INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in "sectioning" the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University Microfilms International

300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4EJ, ENGLAND

TROCHIM, WILLIAM MICHAEL K.

THE REGRESSION-DISCONTINUITY DESIGN IN TITLE I EVALUATION: IMPLEMENTATION, ANALYSIS AND VARIATIONS

Northwestern University PH.D. 1980

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs _____
2. Colored illustrations _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Print shows through as there is text on both sides of page _____
6. Indistinct, broken or small print on several pages ✓
7. Tightly bound copy with print lost in spine _____
8. Computer printout pages with indistinct print _____
9. Page(s) _____ lacking when material received, and not available from school or author
10. Page(s) _____ seem to be missing in numbering only as text follows
11. Poor carbon copy _____
12. Not original copy, several pages with blurred type _____
13. Appendix pages are poor copy _____
14. Original copy with light type _____
15. Curling and wrinkled pages _____
16. Other _____

University Microfilms International
300 N. ZEEB RD. ANN ARBOR, MI 48106 (313) 761-4700

NORTHWESTERN UNIVERSITY

THE REGRESSION-DISCONTINUITY DESIGN IN TITLE I EVALUATION: IMPLEMENTATION, ANALYSIS AND VARIATIONS

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

FIELD OF PSYCHOLOGY

by

WILLIAM M. K. TROCHIM

EVANSTON, ILLINOIS

AUGUST, 1980

ACKNOWLEDGMENTS

This dissertation is the result of the assistance and encouragement of many individuals. Robert Boruch, who chaired my committee, assured me an environment within which to develop my own ideas and the support needed to carry out this project. Donald Campbell patiently introduced me to the regression-discontinuity design and counselled me on technical issues throughout this investigation. Tom Cook exhorted me to strive for academic excellence in this project and my education in general. Dave Cordray diligently reviewed several drafts and offered friendship along with constructive criticism. These few words cannot adequately convey my gratitude to this committee but I hope this volume will offer evidence of their immense contributions to my education. Any errors contained here must be ascribed to me alone and my inability to equal the example which these advisors provide.

I also wish to note the special contributions of two close friends. Chip Reichardt has worked with me for years on the regression-discontinuity design and could always be counted on for advice and encouragement. Ron Visco provided extensive and interesting insights into the use of regression-discontinuity in Title I evaluation. Many others influenced this work through interviews and phone conversations and helped me gain access to the data which is presented here.

This work would not have been possible without the financial support provided under several grants: NSF Grant DAR 7820374 (Robert Boruch, principal investigator), NIE Grant G-79-0128 (Robert Boruch and David Cordray, principal investigators) and NSF Grant BNS 7826810 (Donald Campbell, principal investigator). Kim Cudworth aided greatly in the preparation of several drafts of this manuscript.

Finally, I wish to thank those closest to me for their love and encouragement throughout my education. My parents are without equal -- they stood with me in good and bad times. My sister, Mary Anne, gave me confidence in myself through her confidence in me. This work is dedicated to them and most of all to Mary, my wife. It is her love, encouragement and support which gives meaning to my work and my life.

TABLE OF CONTENTS

I. THE REGRESSION-DISCONTINUITY DESIGN IN TITLE I EVALUATION

Introduction. The regression-discontinuity design first elaborated by Thistlethwaite and Campbell (1960) is a research methodology which can be used to test the causal hypothesis that a particular program or procedure has resulted in a measurable effect. The emphasis in this work is on the relationship between the methodological theory behind the design and its application within a particular context, namely, the evaluation of compensatory education programs.

This study can be viewed as an investigation of three major topics. The first concerns "design implementation," that is, the examination of discrepancies which arise between the theory and application of a research design. There are several reasons for studying theory and application together rather than either one separately. First, it is possible to determine where methodological theory is absent and/or naive. A research design which is methodologically sound but does not fit any real-world application is useless for all practical purposes. More to the point, a study of this type enables a determination of specifically where the design theory is applicable and where it is not. Second, it becomes possible to determine where the design is being incorrectly or incompletely implemented. Within any given research

setting there are many factors which hinder the "correct" use of the design and which are difficult to detect without thorough investigation. Finally, study of how the design is implemented can provide feedback to the methodologist which can be used to derive more appropriate variations of the design or entirely new methods. Especially useful is information concerning how often a particular problem occurs in practice and the parameters of this problem. The topic of design implementation analysis is discussed in Chapter II.

The second major topic involves the statistical analysis of the regression-discontinuity design. A design implementation analysis will detect the most important violations of the requirements of the design, and, because of such violations one will often have to modify the traditionally-applied analytic techniques or develop new ones altogether. One promising analytic strategy is offered here as a useful alternative to the analyses which are traditionally applied. This approach is investigated through illustrative computer simulations and secondary analyses of real data, and its implications for the regression-discontinuity design (especially in the context of compensatory education evaluation) are explored in Chapter III.

The third major topic concerns variations on the regression-discontinuity design. These may serve to reduce the possibilities for design degradation due to implementation

problems and to increase the applicability and feasibility of the design. One useful variation involves "coupling" the regression-discontinuity design with random assignment to group. This reduces the equivocality of estimates of program effect while limiting some of the implementation problems usually associated with entirely randomized designs. Another promising application stems from the fact that in the regression-discontinuity design persons are assigned to treatment and control groups solely on the basis of a quantified measure (or composite of measures). This is similar, in principle, to the way in which the federal government "assigns" or allocates domestic aid on the basis of a formula. If the requirements of the regression-discontinuity design (or a modified version of the design) match the formula allocation procedure, one has a useful and relatively easily implemented manner of evaluating the effect of the domestic aid program. These and other variations are discussed in Chapter IV especially for compensatory education evaluation (although the concepts can easily be extended into other evaluation arenas).

This chapter includes a description of the regression-discontinuity design, background information on the context of compensatory education research, and an explanation of the plan of the study.

Introduction to the Regression-Discontinuity Design

The Basic Design. The regression-discontinuity design is one member of a larger class of quasi-experimental methods which can be termed pretest-posttest group designs. In their most basic form they require a pre-program measure (e.g., the pretest), a post-program measure which can reflect the effect of the program being studied (e.g., the posttest), and a measure which describes the assignment status of the persons in the study (assignment variable), usually whether they received the program (treatment group) or did not (comparison or "control" group).

The regression-discontinuity design can be distinguished from other pretest-posttest group designs by the nature of its assignment strategy. In regression-discontinuity all persons are assigned to treatment or comparison group solely on the basis of a cutoff score on the pre-program measure. Thus, once a cutoff score has been selected, all persons scoring on one side of this score are assigned to one group (e.g., treatment) while those scoring on the other side are assigned to the other (e.g., comparison). The design is useful whenever one wishes to study a program or procedure which is given out on the basis of criteria such as need or merit. This is the case in compensatory education where children in greatest need of additional instruction (as reflected in a pre-instruction measure) are targeted to receive the services; or, in medicine, where those who most

require a new surgical procedure (as reflected in a pre-surgery measure of severity of illness) are selected for surgery; or, in the awarding of scholarships where those who exceed a certain cutoff criteria (as measured in a test of skill or intelligence) are given the award. In all cases, the post-program measure is presumed to reflect the effect of the program or procedure.

Perhaps the best way to understand the idea of the regression-discontinuity design is through graphs. In Figure 1, hypothetical data is presented for the case of regression-discontinuity applied to the study of a compensatory education program. The figure shows the relationship between the pre and post-program measures for each person in the study. The horizontal scale indicates the range of pretest values while the vertical depicts the posttest. Each point on the graph reflects a pretest and posttest score for a given individual. In this hypothetical example, the cutoff score is zero and therefore, all students achieving a pre-program score which is less than or equal to zero receive the special instruction (treatment group) while those whose pretest score exceeds zero do not (comparison group). In the graph, treatment group scores are indicated by a "X" while comparison scores are signified by a "O."

We can understand how the regression-discontinuity design works by considering first just the comparison group scores. The portion of the graph to the right of zero on the

pre-program measure shows the pretest and posttest scores for those students who did not receive the special instruction being studied. A regression line could be fit to these scores to describe the pre-post relationship. If this regression line were extended into the region of the treatment group scores (i.e., to the left of zero on the pre-program measure) the extension would represent the pre-post relationship which would be expected in the treatment group if they, like the comparison group, had not received the instruction. This situation is shown in Figure 2 where the comparison group regression line is indicated by the solid line to the right of zero on the pretest and the expected treatment group line is indicated by the dashed line.

The essential idea in regression-discontinuity analysis is to determine whether the observed pre-post relationship in the treatment group differs from the expected relationship as derived from the comparison group scores. For reasons which will be discussed in Chapter III this is most appropriately reflected in the difference between the treatment and comparison group regression lines at the cutoff point. In Figure 2, the obtained treatment group regression line is indicated by the solid line to the left of zero on the pretest. For any pretest value the obtained line has a higher posttest score than does the line projected from the comparison group scores (in this example, the difference is a

constant amount). This indicates that the treatment group students scored higher on the posttest than was "expected" and one can infer that the instruction had a positive effect on posttest scores. Similarly, if the observed treatment group regression line had been displaced below the expected line one could assume that the instruction affected students negatively.

It should be recognized that the above example illustrates only a very simple version of the design. For instance the pre-program measure can be a pretest (i.e., the same test as the posttest or an alternate form of the same test) or it can be a different measure or composite of several measures. The treatment group need not be the persons scoring below the cutoff score. If the program is given out on the basis of merit, those scoring above the cutoff might receive the treatment. It is not necessary that the pre-post relationship be linear -- any shape which can be modelled with a regression line would be appropriate. In its most general sense, the regression-discontinuity design estimates the effect of the program as the difference between the observed treatment and comparison group regression lines at the cutoff point. In fact, the design takes its name from the fact that when the program has an effect there is a discontinuity in the overall regression line at the cutoff score.

Assumptions of the Regression-Discontinuity Design. All

research designs are based, to varying degrees, on assumptions. Some are common to almost every design -- sufficient quality of measurement, statistical power, implementation of the program construct, and so on. Others are peculiar to a given design and, in fact, serve to set that design apart from all others. The regression-discontinuity design can be characterized by three central assumptions -- perfect assignment relative to the cutoff, correct specification of the statistical model and the absence of "coincidental" functional discontinuities. All three are related to the most distinguishing feature of the design, the assignment to group on the basis of a cutoff on a pre-program measure.

The first assumption of the regression-discontinuity design has to do with adherence to the requirements of the design, and most importantly, adherence to the cutoff criteria. The regression-discontinuity design assumes that there is no misassignment in terms of this cutoff, so that persons who by their pre-program score should be in one group are not placed in the other. Misassignment can occur for a variety of reasons. If potential participants have political "pull," they may be able to prevail upon administrators to misassign them into a desirable program or out of an undesirable one. Sometimes misassignment can occur because a well-intentioned administrator does not wish to deny a potentially beneficial program to certain persons,

especially those narrowly missing being included by the cutoff criteria. Another cause of misassignment is administrative error, that is, insufficient vigilance in assuming that the cutoff criterion is adhered to. Some misassignment will be easily detected by noting cases which did not fall into the appropriate groups, while at other times it will be impossible to detect as when persons are given the program but their participation is not recorded.

Misassignment relative to the cutoff score has been termed "fuzzy" regression-discontinuity by Campbell (1969). When the cutoff is strictly adhered to we can term the design "sharp" regression-discontinuity. With certain types of "fuzzy" assignment (e.g., random misassignment around the cutoff) the traditional analytic strategy will yield unbiased estimates of treatment effect. Other types of misassignment are known to yield biased estimates (Goldberger, 1972). The basic regression-discontinuity design requires sharp assignment. However, the design would be far more flexible if this assumption could be relaxed and fuzzy assignment allowed, thus making it possible for administrators to use more discretion in the assignment of special cases. There has been a good deal of theoretical work on analytic strategies appropriate for "fuzzy" regression-discontinuity (Campbell, Reichardt, and Trochim, 1979; Trochim, 1980; Barnow, Cain, and Goldberger, 1978; Spiegelman, 1976) but a detailed discussion is outside the scope of this work.

The second major assumption is that the analytical model used to estimate program effect accurately describes the true pretest-posttest functional relationship. For example, if the true relationship is in fact linear but the analysis uses only quadratic, cubic, logarithmic, or other functions, biased estimates of effect are likely. Since the true model is seldom known one usually chooses a model either a priori, based upon experience with similar data, or through testing likely alternative models. An analyst might, for example, state a priori that the pre-post relationship in the absence of a treatment is linear, infer from previous experience with the same test and subjects that it is so, or test this hypothesis against likely alternatives such as quadratic, logarithmic, and so on. The problem with models selected in this way is that one is never sure that the chosen model is correct. An alternative approach is outlined in Chapter III. It is appropriate only for a certain class of functions (i.e., when the true pre-post relationship can be expressed as a finite and low order polynomial) and is designed to reduce the possibility of selecting estimates of effect which are biased due to misspecification.

The third assumption is that there are no other factors which, even in the absence of the program, would result in a discontinuity in the pre-post relationship at the cutoff point. This can be conceptualized in two ways. First, if an effect is observed, is it due to something other than the

treatment? Second, could the effect be partially due to the treatment and partially to some other factor? The latter implies that the size of the treatment effect may be incorrectly estimated and that, even if no effect is observed, this may be due to either an ineffective program or to counteracting factors which disguise the effect.

Any factor which affects the scores in one group and not the other can lead to a discontinuity which is mistaken for a treatment effect. For example, if the program participants are put in one setting while the controls are in another, any factor associated with the setting (and not the program) which affects the posttest will manifest itself as a treatment effect. Therefore, the regression-discontinuity design assumes that all group factors which differentially affect the posttest, other than the treatment itself, are accounted for in either the design or analysis.

This brief look at the assumptions which underlie the regression-discontinuity design illustrates the importance of investigating the application of the design in various contexts. While methodologists can outline potential violations of assumptions, a study of design implementation can determine which assumptions are violated most often and why. Much of this work will involve an investigation of the appropriateness of the assumptions within the context of compensatory education.

The Research Context: Title I Evaluation

Background. In 1965, the U.S. Congress passed Public Law 89-10, the Elementary and Secondary Education Act (ESEA), Title I of which is most relevant here. It was the first major piece of legislation in the Great Society program of Lyndon Johnson and the largest single federal education grant program to date. Passage of an education program of its scope was a singular domestic achievement which required a balancing of several divergent interest groups and not a little political straight-arming (Eidenberg and Morey, 1969). The Statement of Purpose (Section 201) holds that Title I is

> ...to provide financial assistance...to local educational agencies serving areas of concentrations of children from educational programs by various means...which contribute particularly to meeting the special educational needs of educationally deprived children.

Each local school district or LEA develops a proposal describing the programs for which it requests funds and submits this to the state department of education, or SEA, for approval. The programs are usually (but not necessarily) confined to basic skill areas such as mathematics, reading and language arts. In 1979, over $5.5 billion was authorized by Congress for Title I and nearly $3.4 billion was finally appropriated. Approximately nine million low income children were served and between 5 and 6 million of these were of elementary school age. Nearly 87% of all school districts receive some Title I funds (NCES, 1979) which account for

between 3 and 4% of all national elementary and secondary education expenses (Office of Education, 1979). Only those aspects of Title I which are relevant to the application of the regression-discontinuity design are discussed here. The reader is referred to documents of the Office of Education and the paper by Wick (1978) for a more general description of Title I.

Title I Evaluation. Title I was the first major legislation which specifically required routine evaluation of its programs. Section 205(a)(5) reads

> ...that effective procedures, including provision for appropriate objective measurements of education achievement will be adopted for evaluating at least annually, the effectiveness of the programs in meeting the special educational needs of educationally deprived children.

Title I has been amended six times since its inception. In the Amendments of 1974 (Public Law 93-380) Congress attempted to improve and standardize the evaluation procedures at the federal, state, and local levels with the development of research models which

> ...shall specify objective criteria which shall be utilized in the evaluation of all programs and shall outline techniques (such as longitudinal studies of children involved in such programs) and methodology (such as the use of tests which yield comparable results) for providing data which are comparable on a statewide and nationwide basis (Section 151(f)).

In doing so, Congress moved away from an approach emphasizing demonstration-type studies by instead attempting to combine the evaluation and reporting functions. The research methods

which were to be used would be clearly specified and results could be aggregated by district, state, and nation. In 1974 a contract was awarded by the Office of Education to the RMC Research Corporation for the development of such models. The system which they generated is characterized by a common measurement metric and by three alternative research designs which are presumed to yield comparable results (Talmadge and Horst, 1976; Talmadge and Wood, 1978). The metric, termed the Normal Curve Equivalent score (NCE), is a standard score with a mean of 50 and a standard deviation of 21.06. Estimates of treatment effect are to be reported on this scale to enable aggregation of gains at higher levels. The three research designs, which differ in the way they generate the expectation for the treatment group, can be described briefly as follows:

> Model A: The Norm Referenced Model. With this model only the treatment group students are pre and posttested. In an attempt to avoid regression artifacts it is required that treatment students be selected by some measure (or measures) other than the pretest. The average test scale score is calculated for the pretest and posttest and these can then be converted to percentiles. Because the average test scale scores are used and these are based on test scores obtained from norming samples the treatment effect consists of any gain over and above that which would be expected in the norming sample group.
>
> Model B: The Control Group Model. Both treatment and control (i.e., comparison) group students are tested with this model. Ideally, students are randomly assigned to group, thus yielding a true experimental design. In practice, this is not often possible, and "comparable"

groups are used as treatment and control, yielding a non-equivalent group design. The treatment effect consists of any gain in the average of the treatment group over and above the gain in the average of the controls.

Model C: The Special Regression Design. This is the regression-discontinuity design as described here (throughout this paper the term "Model C" will often be applied to refer to the Title I implementation of the design). Often, two separate estimates of program effect are computed. The "regression-discontinuity estimate" is the difference between the treatment and comparison group regression lines at the cutoff point. The "regression-projection estimate" is the difference between the lines at the treatment group pretest mean.

All three models can be used with either standardized or locally-developed tests and are distinguished in the model names with a "1" for the former and a "2" for the latter. For example, Model A1 is the norm referenced model with a standardized test, Model A2 is the model with a local exam.

A system of technical assistance was set up to aid LEAs and SEAs in the design and execution of the evaluations. In each of the ten national Department of Health, Education and Welfare regions a Technical Assistance Center (TAC), upon request, is allowed to provide advice, training, and sometimes, assistance with statistical analysis. A good summary of the Title I system in general and the models in particular can be found in U.S. Office of Education (1976) and Talmadge and Wood (1978). A useful bibliography of Title I evaluation issues was compiled by Strand (1979).

Although the basic evaluation system was defined as early as 1976 and used in some LEAs as early as 1977, the

system is only now being formally implemented. The final regulations (45 CFR Parts 116 and 116a) were published on October 12, 1979, and were expected to take effect forty-five days after transmission to Congress. In addition, the RMC Research Corporation is in the process of preparing a Policy Manual of which Subpart F will deal with Title I evaluation policies.

To summarize, we can see in the history of Title I the development of an evaluation system of national scope, specifying three presumably "equivalent" (in terms of treatment effect estimates) research designs, one of which is the regression-discontinuity design of interest here. These conditions make the context of Title I evaluation the richest source which is currently available for information on the application of the regression-discontinuity design.

Plan of the Study

Several distinct tasks were undertaken in order to understand more clearly the characteristics of regression-discontinuity analysis within the Title I evaluation context.

Interviews and Site Visits. Nearly fifty people were contacted in telephone interviews, face-to-face meetings or through site visits. The persons contacted and the forms of contact are listed in Appendix A. The telephone interviews and face-to-face meetings ranged from five minutes to over an hour each and were comprised of open-ended unstructured

questions designed to ascertain the impressions which people have of the regression-discontinuity design in general, the frequency and location of its use, the reasons for its use or lack of use, the major problems which occur in implementation and analysis, the general pattern of results, and so on. Included in these interviews and meetings were several persons from the Office of Education, at least one representative of each Technical Assistance Center (TAC), personnel representing state Title I evaluation units, persons responsible for Title I evaluation at the local level in a number of school districts, and scholars and researchers familiar with the use of the design. The general outline followed in conducting these interviews is presented in Appendix B.

In addition to the interviews, two site visits were made for the purposes of gathering more detailed information and obtaining data for secondary and meta-analysis. The first site visit, conducted December 26 and 27, 1979, was to the Region I (New England) TAC, the RMC Research Corporation in Portsmouth, New Hampshire, and to the Providence, Rhode Island Department of Education. The second, conducted February 4 to 8, 1980, included a visit to the Florida State Department of Education and the school districts of Marion and Osceola County Florida. The sites were chosen on the basis of recommendations from TAC personnel and after discussion with potential respondents.

The major product of the interviews and site visits consists of a listing of school districts which have attempted to use the regression-discontinuity design, a considerable store of qualitative, impressionistic information, and data suitable for secondary and meta-analysis as described below.

Meta-analysis of Regression-Discontinuity Results. The State of Florida has by far the greatest frequency of use of regression-discontinuity within the Title I system. The design has been used for two years (1977-1978 and 1978-1979) by seventeen school districts (in this case, counties). During this time 273 separate regression-discontinuity designs were implemented on data for nearly 185,000 students. A site visit made to the Florida State Department of Education yielded data for each analysis on program characteristics such as subject area, setting, program goals, and the evaluation plan, as well as aggregate statistics including the cutoff score percentile, group sample sizes, pretest and posttest slopes and means, and estimates of gain or program effect. This information is useful in describing current practice and may be analyzed to determine differential effectiveness of the design under varying conditions.

Secondary Analysis of Regression-Discontinuity Data. Annual testing data for two successive years (1978 and 1979) was obtained from three school districts: Providence, Rhode

Island; Marion County, Florida; and Osceola County, Florida. From each district, data was obtained for several grade levels and for both reading and mathematics instruction. Altogether, data was gathered for twenty-four separate regression-discontinuity analyses. This data was analyzed to help determine the frequency and extent of various biasing factors and to apply a number of potentially useful analytic procedures for purposes of contrasting treatment effect estimates. Thus, questions related to the quality of measurement, the presence of floor and ceiling effects, the program effect and so on, were explored.

Illustrative Computer Simulations. The analysis of simulated data with known characteristics is useful in the exploration of data analytic problems. For example, competing analyses can be tested on the same data and the extent of any biases which are obtained can be documented. In this study, simulations are not meant to be definitive. Instead, they are included primarily to illustrate or confirm an argument or to suggest potential solutions for particular problems. For example, the analysis of simulated data is used to explore the effects on regression lines of restricting pretest range as well as to confirm the potential of alternative analytic procedures applied to regression-discontinuity data.

The results obtained from each of these four tasks are reported, where appropriate, throughout this paper, although

the interviews and meta-analysis are described for the most part in Chapter II while the secondary analysis and illustrative simulations are predominantly in Chapter III. The amount of information related to the use of regression-discontinuity in Title I evaluation exceeds what can be described in detail in a single research report. The intent of this project is to examine the area in a general way, define the issues of primary importance, explore these issues in some detail, suggest some potential solutions, and finally, suggest ways in which the topic could be profitably explored in future research.

II. DESIGN IMPLEMENTATION

The study of design implementation involves the investigation of why, where, and how a particular research design is used. In its broadest sense, it can be viewed as applied research methodology or "research on research," that is, the study of research practice for the purpose of improving research design. While most research reports deal in part with the way in which the design was implemented, this topic has seldom been the major focus of attention. An exception has been the work of Conner (1977, 1978) who has examined the implementation of randomization procedures in various settings.

One might also find information on design implementation in studies of secondary analysis or meta-evaluation. The former involves the reanalysis of data in order to explore alternative analytic techniques and/or determine the appropriateness of the original analysis (Boruch and Wortman, 1977). The latter makes use of multiple estimates of program effect in order to describe the effect under a wide variety of conditions or to examine the biases of a research design under different circumstances (Gilbert, Light, and Mosteller, 1975). In both types of studies, the emphasis has largely been either on substantive or analytic questions rather than on questions of implementation.

Some effort has been expended on implementation problems

in Title I of ESEA (NIE, 1977; Johnson and Thomas, 1979) although this has primarily emphasized program implementation rather than design implementation. An informal attempt to estimate the relationship between quality of implementation and estimates of gain yielded a correlation of -.25 for a small sample of project sites (Johnson and Thomas, 1979).

There are three major dimensions to design implementation analysis. The first is concerned with the theoretical aspects of the design, irrespective of context. The second concerns the general features of the context irrespective of the research design. Finally, the third concerns the match between a particular design and research context. The first two dimensions were considered in Chapter I while the last one comprises the primary focus of this Chapter.

This chapter describes the use of the regression-discontinuity design in the context of Title I evaluation. Five major topics are considered:

1. Applicability. Issues related to where and how frequently the design has been used, and why it is or is not chosen by school districts are considered.
2. Assignment. The measures used to assign students to the programs, the degree to which the cutoff criterion is adhered to, and the politics involved in assignment are discussed.

3. Measurement. The context of testing, the quality of the data, and problems in data maintenance are considered.
4. Treatment. The problems in determining who received the treatment, how much treatment they received, whether competing programs exist, and the difficulties involved in measuring treatment implementation are considered.
5. Exclusions. The effects of excluding cases in the process of preparing the data are discussed.

It is important to recognize that many of the implementation problems discussed here will also affect other research designs. Therefore, judgments concerning the quality of the regression-discontinuity design must take into account its advantages relative to other appropriate methods. The major issue of concern is how individual implementation problems can act to distort the estimate of the effect of the program when different designs are used.

The three sources of information for this investigation were described briefly in Chapter I and consist of interviews, meta-analysis data and the results of several secondary analyses. It is useful at this point to describe the method in which the meta-analysis was conducted. The description of the secondary analyses is contained in Chapter III.

The data used in the meta-analysis was obtained through the Florida Department of Education. All information was gathered from two forms which are filled out annually by local school districts. The first form is an application or proposal which describes the Title I programs which the school district intends to carry out, the justification for these programs including a needs assessment, the research plan, and so on. The major variables obtained from this application are presented in Appendix C. The second form from which data was collected is the annual evaluation report which consists of summary data describing the results. Relevant variables gathered from this form are presented in Appendix D.

The State of Florida was chosen primarily because it is the largest user of the regression-discontinuity design within the Title I evaluation system. Data were collected for all occurrences of the regression-discontinuity design within the state for two successive years (1977-1978 and 1978-1979). In all, a total of 273 analyses were carried out, 133 of which were undertaken in the first year, 140 of which were carried out in the second. This represents analyses of several grade levels and types of programs within seventeen separate school districts. The data of primary importance for this study includes the subject matter of the program, the setting, grade level, the number of students in the treatment and comparison groups, the estimate of program effect at the

treatment group mean, the estimate of program effect at the cutoff point, the pretest percentile at which the cutoff point is located and an estimate of the amount of data which was missing from the final analysis. Results obtained from the meta-analysis of this data will be discussed along with the issue to which they are most relevant.

Applicability Issues

"Applicability" refers to the feasibility of using the regression-discontinuity design for Title I evaluation. To some extent, one can demonstrate that the design is feasible simply by showing that it can and has been used. This is the first topic taken up below. However, it is possible that Model C is feasible only for certain types of Title I programs or school districts. Because personnel from each district must decide which evaluation model is most feasible for them, it is important to investigate the reasons which they give for selecting a model. For convenience, the discussion of the applicability of the regression-discontinuity design is further divided into two sections -- the major reasons cited for not using the design and the factors mentioned which favor its selection.

Frequency and Location of Use. In order to determine where the regression-discontinuity design was used, interviews were conducted primarily with personnel of the TAC Centers in each of the national Office of Education regions.

From these interviews a list of school districts which were mentioned in connection with the design was compiled. This list is presented in Appendix E. One must view this list with caution for several reasons. First, not all of the information was corroborated at the school district level. Rather, if during the course of an interview a respondent indicated that a district may have tried the design, this district was then listed as a "user." Whenever possible, attempts were made to verify the information directly. In several cases, a district mentioned as a user, when contacted, claimed to have only expressed interest in the design but never used it. Second, the information obtained from the interviews is likely to be out of date in some cases. Each year a school district reconsiders whether it wishes to continue using the research design which they had previously implemented. Since the Title I evaluation system has only recently been implemented in its present form, districts are likely to switch designs from year to year as they search for one which best suits their needs. Some attempt has been made in listing likely users of the design to indicate the form of that use. For example, an attempt was made to confirm whether the design is currently being used by the district, had been used in the past but abandoned, had been considered in the past but rejected, and so on. In several cases interviewees mentioned that they knew one or more districts in a given state were using the design but were not certain or were

unwilling to specify which districts these were. All things considered, the list illustrates the fact that the design is being used and gives a general idea of where this use is most concentrated.

In all, about sixty school districts (out of a total of about 15,000) were mentioned as possible users of the design. Of these, about 40 are likely to still be using it. The majority of current users are located in East Coast regions, specifically, in the states of New Jersey, Virgina, and Florida. In the Midwest and Western states use is primarily restricted to a few large school districts and to "special case" districts such as the Trust Territories.

Another source of information on the use of the regression-discontinuity design is a survey conducted by the National Center for Education Statistics (NCES, 1979). Of the districts which responded, 87% had Title I programs in the 1978-1979 school year and 63% had used one of the three Title I evaluation models. Of those which had used such a model, 86% used Model A, about 2% used Model B and about 2% used Model C. The remaining 10% of the districts using models made use of an alternative or local one. In general then, while the regression-discontinuity design is used, it is not used frequently, especially relative to Model A.

Reasons Cited for Not Using Model C. It is important, when examining why the regression-discontinuity design is or is not used, to bear in mind the alternatives which a school

district has. While two other models are available, the choice is most often between Model A and Model C (Echternacht, 1980). Consequently, many of the reasons listed as factors in the decision pertain to the perceived relative advantages of one model over the other.

In the process of conducting interviews and site visits, ten reasons were frequently mentioned as important in the decision not to use the regression-discontinuity design. These reasons include the major factors cited in other papers such as those by Echternacht (1980) and McNeil and Findlay (1979). It is possible that for any given district some of the reasons which were cited were ill-considered or inappropriate. Nevertheless, these reasons are included because they serve to define, at least in part, how the design is perceived. The following discussion delineates the ten reasons and discusses their reasonableness.

1. Model C is not chosen because it is likely to yield "negative" program effect estimates.

By far, the most frequently given reason for not using the regression-discontinuity design is the commonly-held perception that it tends to yield "negative" gains for estimates of treatment effect. This reason was cited by representatives from almost every TAC, and by many persons at the state and local level. Because this issue is of central importance in assessing the appropriateness and feasibility of the design it is given considerable attention here.

A number of questions are related to the perception that the design yields "negative" estimates of effect. First, is it true that the average gain generated by the regression-discontinuity design is in fact negative when used to evaluate Title I programs? Second, if so, are similar negative gains found with the other two models, most notably, with Model A? If all of the models yield similarly negative results, this factor would probably not affect the design choice. On the other hand, if the average gain with Model C is negative, while for Model A it is positive, one would expect districts to choose the latter because on the average it would reflect more favorably on their Title I programs. Finally, if this is the state of affairs, why would results differ across designs? Three reasons are possible -- gains from Model A tend to be biased upwards, gains from Model C tend to be biased down, or a combination of both of these.

We can begin to determine whether the average gain yielded by the regression-discontinuity design is negative by looking at the distribution of the 273 effect estimates obtained from the Florida Department of Education. In the Title I evaluation system these gains are estimated at two points, the treatment group mean and the pretest cutoff score. Although in Chapter III it is recommended that the treatment effect not be estimated at the treatment group mean, the distribution of gains will be shown for both estimates. Figure 3 depicts this distribution for the

estimate at the treatment group mean while Figure 4 shows it for the one at the cutoff score. Both distributions appear to be roughly normal in shape and, in general, the means and medians tend to be similar. For the estimate at the mean, the average is -1.132 NCEs with a standard error of .398 (n=273) while the median is -.983. For the estimate at the cutoff, the average gain is -2.689 NCEs with a standard error of .377 while the median is -2.714. One of the estimates at the cutoff point looks suspiciously low (gain = -48.0) but even with this value excluded the average gain is estimated to be -2.518 with a standard error of .337 where the median is -2.69. The commonly-held notion that the average gain with the regression-discontinuity design is negative is clearly supported by these distributions.

Even if the average gain is negative one can question whether this may be an accurate result. On the basis of what is known about the effects of compensatory education programs in general, it is difficult to say what the "expected" gain should be. Most of the studies of Title I programs which had previously been conducted only indirectly addressed the question of program effect and often, these results were suspect due to methodological, measurement, and analytic flaws (Wick, 1978; Campbell and Erlebacher, 1970; Campbell and Boruch, 1975). Even granting that the biases in analysis have been against finding effects, programs of this nature have not been found to be conspicuously effective when more

appropriate modes of analysis have been used (Magdison, 1977; Bentler and Woodward, 1978; Director, 1974). However, significantly harmful effects have so far primarily been explainable as mistaken methodology.

Within the Title I milieu the results from the regression-discontinuity design are not considered alone. Instead they are contrasted with the generally positive gains reported by those districts using Model A. In the state of Florida in 1978-1979 the average gain using Model A was in the vicinity of 6.0 NCE units (Florida Department of Education, 1979). This is considerably higher than the slightly negative estimate for the regression-discontinuity design and suggests the possibility that Model A may be biased upward, the regression-discontinuity design may be biased down, or both. Each of these will be considered in turn.

First, estimates from Model A do appear to be biased. Several persons have commented on this fact (Echternacht, 1979, 1980; Murray, 1978; Linn, 1979; Glass, 1978) and it can be easily demonstrated through the following argument which is based upon the discussion presented by Glass (1978).

It is well known that when a group is selected from one end of a distribution of scores their mean on any other measure will tend to "regress" toward the overall mean of this other distribution even in the absence of a treatment effect (Roberts, 1980). This has traditionally been

labelled the "regression artifact" and is discussed in Campbell and Erlebacher (1970) and Campbell and Boruch (1975).

We can specify the amount of regression to the mean which occurs between any two measures, A and B. Assume that we select a group from the extreme lower end of the distribution of variable A. When the standard deviations of the two distributions are equal (e.g., they are in standard score form), the correlation between the two measures for the selected group is a direct reflection of the amount of regression to the mean. In fact, the symbol for the correlation, "r," was originally used to signify "regression" in this sense. Specifically, (1 - r) x 100 gives the percentage of regression to the mean (assuming equal variances). For example, if there is a perfect correlation between measure A and B for the selected group (i.e., r = 1.0) there is no regression to the mean (i.e., (1 - 1) x 100 = 0% regression). Conversely, if there is no correlation (i.e., r = 0), there is maximum regression to the mean (i.e., (1 - 0) x 100 = 100% regression). Or, if r = .6, the mean of the selected group on measure B will be 40% closer to the overall mean of B than the mean of the same group on measure A is to the overall mean of A (i.e., (1 - .6) x 100 = 40% regression). Each of these cases is shown in Figure 5 (i.e., r = 1.0, 0, or .6). If measure A is a pretest and B a posttest, except when they are perfectly correlated, the

selected group will appear to "improve" from A to B even in the absence of any treatment simply because of the regression artifact resulting from a less than perfect correlation and selection from the extreme of the distribution.

This fact was well known when Model A was devised for use in Title I evaluation. In attempt to avoid the regression artifact with Model A, some measure other than the pretest is required for the assignment of students to the program. One might use an achievement test given in the previous year, a separate one (i.e., separate from the pretest) given in the current year, grade point averages, teacher ratings, and so on. By using a separate measure it was thought that assignment would be "independent" of the potential regression artifact. The fact is that this separate selection measure is likely to remove only part of the regression artifact.

To see this, consider an assignment variable, z, a pretest, x, and a posttest, y, all in standard score form and hence having equal standard deviation units. Assume that a treatment group is selected from the lower end of the distribution of measure z as is commonly done in Title I (if the measure z is related to achievement we wish to select those "most needy"). Further assume that the correlation between the assignment measure and the pretest is $r = .8$. There would be regression to the mean from z to x such that the mean of the selected group on the pretest, x, would be

20% closer to the overall pretest mean than this group's mean on z is to the overall mean of z. Assume also that the correlation between the assignment measure and the posttest, y, is r = .5. There would then be regression to the mean from z to y. Here, the mean of y for the selected group would be 50% closer to the overall mean of y than the mean for this group on z is to the overall mean of z. By subtracting these two regression artifacts we find a "residual regression artifact" of 30%. That is, the treatment group's mean on the posttest, y, would be 30% closer to the overall mean of y than their mean on the pretest, x, is to the overall mean of x. The separate selection measure only removes part of the regression artifact. One would find that students improve from pretest to posttest even though no program is ever given.

We can conclude that if the treatment group is selected from the extreme of the distribution of z and if $r_{zx} > r_{zy}$ there will be some regression to the mean unaccounted for by the separate selection measure. If $r_{zx} = r_{zy}$ there will be no residual regression to the mean. If $r_{zx} < r_{zy}$ there will actually be regression away from the posttest mean and the selected group will appear to have lost ground from pretest to posttest. These last two cases are not likely to occur in practice. Given the usual finding that correlations between measures tend to be lower the longer the time between testings (e.g., the correlations may follow a simplex model

as described in Humphreys, 1960, and in Campbell and Boruch, 1975) it is more reasonable to argue that $r_{zx} > r_{zy}$ and that Model A is biased due to a residual regression artifact (labelled "B" in Figure 6).

To examine relative bias in the two models some attempts have been made to compare the estimates of gain from Model A and Model C when implemented simultaneously on the same program. In several such cases uncovered during interviews (Louisiana State Department of Education, 1980; Hill, 1980; House, 1979) the comparison can be considered questionable because one or both of the designs was incorrectly implemented. For example, several districts compared a "Model A analysis" with a "Model C analysis" but did not follow the requirements of the designs. Specifically, these comparisons are often based on version of Model A where a separate selection measure was not used or of Model C where assignment was not "sharp" relative to the cutoff. Thus the comparisons are made using Model A estimates which are biased due to regression artifacts or Model C estimates which may be biased due to "fuzzy" assignment. In one study which was carried out in Escambia County, Florida (Lloyd, 1979), the evaluator claimed that both designs were correctly implemented. Even then, different results were obtained with the two designs. For the reading programs the average gain using Model A was equal to 8.0 NCE units while for Model C it was 3.0 at the treatment group mean and 0.0 at the cutoff. For

programs in math the difference was less clear cut. Here the estimate of gain for Model A was 4.0 NCE units while for Model C it was 2.0 NCE units at the mean and 3.0 NCE units at the cutoff point. Another study (Horwitz, Long and Pellegrini, 1979) compared Model A and C analyses when applied to Model C data. For all four grades studied, Model A yielded estimates of gain which exceeded the Model C estimates by anywhere from 8.5 to 13 NCE units.

In general then, it is reasonable to argue that the slightly higher average estimates of effect obtained with Model A may be due to the lingering presence of the "residual regression artifact" which results from selection from the distribution extreme and the higher correlation between the assignment measure and the pretest than between the assignment measure and the posttest.

A second possible cause for the observed negative gains may be unrelated to bias in Model A. While it is known that under perfect conditions the estimates of treatment effect yielded by Model C will be unbiased, it is quite possible that these necessary conditions are not prevalent. Several factors which could have systematically degraded the results obtained from Model C were identified. The first three comprise the three assumptions of the regression-discontinuity design as outlined in Chapter I, the fourth involves an implementation problem and the fifth consists of a commonly-held misconception. They hold that negative gains result

from degradation of the Model C design due to:

- Misassignment relative to the cutoff.
- Model misspecification.
- Differential factors across groups.
- Computer programming errors.
- Restriction of pretest range.

Each of these is discussed in turn.

Misassignment Relative to the Cutoff. The potential for misassignment is primarily discussed later along with assignment issues. Errors in assignment are usually handled in one of two ways in the statistical analysis -- either the misassigned cases are excluded or they are included with the group to which they were assigned. Both of these procedures are likely to bias the estimates of effect and it is argued later that the bias will be negative given typical Title I implementations.

Model Misspecification. The Title I analytic model assumes that the pretest-posttest distribution is linear. If it is not linear, the Title I model is misspecified and the estimate of gain is likely to be biased. A nonlinear relationship could be reflective of the true distribution or the result of problems like floor and ceiling effects, chance level effects, misassignment (in relation to the cutoff) or data exclusions. The analytic approach outlined in Chapter III is designed primarily to reduce the chances of calculating estimates biased due to model misspecification.

Differential Factors Across Groups. As mentioned earlier, any factor which affects the posttest and which is differentially distributed between groups can masquerade as a treatment effect. For example, treatment group students may be more likely to make errors in coding identifying information on answer sheets. This would result in fewer pretest-posttest score matches and consequently, in under-representation of treatment group scores in the analysis. Problems of coding and score matching are discussed more fully with measurement issues. Similarly, the treatment group may have a greater attrition rate or may be more mobile than comparison students. This leads to disproportionately more treatment group students being "excluded" from the analysis (because both tests were not obtained) and, as a result, can bias the estimate of effect. These problems are discussed along with other exclusion issues below.

Computer Programming Errors. Negative gains for Model C (and, for that matter, positive ones for Model A) could in part be due to data processing errors such as faulty computer programs. While the following error actually acted to bias Model C results in a positive direction, it is illustrative of such computer difficulties and demonstrates that they can affect the results. The 1978 annual report from Florida (Florida Department of Education, 1978) states:

Due to a computer programming error it is

> impossible to differentiate between Title I reading participants and Title I mathematics participants. Therefore, when the Panhandle Area Cooperative (PAEC) analyzed each component, test scores for students not participating in that particular component were inadvertently included. The results were invalid. Because of a second computer programming error, some of the Model C results submitted as 0 may actually be negatives. The computer program as implemented by PAEC was not set up to accomodate negative findings. Although the 0 Model C results are not 100% accurate, they are included in the statewide aggregation. The computer errors are being corrected.

It is interesting to speculate that the optimistic expectation was that no negative gains would be found for Title I programs and consequently, the computer program was not originally able to generate such findings. Errors of this type are not likely to be a factor in the pattern of negative gains unless they occur consistently in a variety of school districts. While this is not likely, it is possible if a number of districts use the same computer program (e.g., a program distributed nationally) and the program yields erroneous negative gains for Model C or, for that matter, positive ones for Model A.

Restriction of Pretest Range. One reason mentioned by several interviewees (Murray, 1978; Blair, 1980) as a possible cause for the pattern of negative gains concerns the effect which the restricting of the range of pretest scores might have on the estimate of within-group slopes and consequently, on the treatment effect. This notion is related to the well-known fact that restricting the range of

scores on one variable will lead to lower estimates of the correlation coefficient between that variable and any other. In the context of Title I evaluation the treatment group usually encompasses a smaller pretest range than does the comparison group. Later, we will see that the median cutoff value for 273 analyses in Florida was the 28th pretest percentile. If it is true that restricting the pretest range alone will result in lower slopes for the treatment group, this would in part explain the negative gains. This, however, is not the case.

While it is true that restricting the pretest range will lead to lower within-group correlations, it will not bias the within-group estimates of slope. To demonstrate this argument several illustrative simulations were conducted. First, a normally distributed "true score" was generated using a pseudo-random number generator to have a mean of zero and a standard deviation of 3.0. Second, a pretest and a posttest were constructed by adding to the true score separate random "error" variables each having a zero mean and various standard deviations. Third, for each pretest-posttest combination constructed in this manner correlations and slopes were calculated after restricting the pretest range by various amounts.

The results of these simulations are presented in Table 1. By gradually narrowing the pretest range (i.e., by discarding all pretest-posttest pairs greater than a given

pretest value) one can see whether or not the slope changes as a result of the restriction. It is clear from Table 1 that all of the estimates of slope for any given pretest-posttest combination are very close in value and tend to agree with the true slope. Table 2 shows more detailed results for one such simulation. This includes estimates of the standard deviations of the pretest, s_x, the posttest, s_y, the covariance between the pretest and the posttest, s_{xy}, the sample size, n, the correlation, r, the ratio of the variances, s_y/s_x, and the estimate of slope. It is clear that while the estimates of slope tend to be distributed around the true value of .9, the correlation coefficient declines as expected as the variance of the pretest is restricted more. The reason for the stability of the slope can be seen in the ratio of the standard deviation of the posttest to the standard deviation of the pretest. Since one definition of the slope is $r(s_y/s_x)$, we can see that as the correlation coefficient declines this ratio needs to increase so as to keep the estimate of slope at a constant. This is demonstrated in Table 2.

The reason for the increase in the ratio of standard deviations can be seen more clearly by looking at a graph of the pretest-posttest relationship as shown in Figure 7. With the bivariate normal distribution the assumption is made that the posttest variance is the same for any pretest value (i.e., we have homogeneity of variance). In the figure, the

the variances of the pretest and posttest are equal. Therefore, if the pretest range is not restricted at all, the ratio of standard deviations will be equal to 1.0. Consider now what happens if we discard all cases having a pretest value greater than "a" in the figure. It is clear that the posttest variance greatly exceeds the pretest variance when this great a restriction is made on the range of the pretest. Thus we can conclude that with greater restriction on the range of the pretest, the ratio of the standard deviations increases exactly enough (assuming a bivariate normal distribution) to offset the change in the correlation, thereby keeping the estimate of slope at a constant value. Because of this it is unlikely that the pattern of negative gains which have been observed are due to the restriction of the treatment group pretest range as a result of assignment by a cutoff.

We can summarize the negative gain issue as follows. First, it does appear that the average gain for Title I programs using the regression-discontinuity design is in fact negative. Along with this, it also appears that the average gain using Model A is positive. Second, even if perfectly implemented, Model A yields biased estimates of effect due to what is termed here a "residual" regression artifact. The difference in average gain can be explained in part by this bias. Third, it is likely that the negative gains are in part also due to design degradation in Model C. Although

ideally Model C will yield unbiased estimates of effect when "perfectly" implemented and there are no violations of assumptions, there is good reason to believe that in Title I settings this is seldom the case. If one could estimate the degree to which implementation problems distort the estimate of gain for Model C, it would be possible to subtract this value from the observed average gain to obtain an unbiased estimate of effect. It is difficult to gauge the relative seriousness of these implementation problems primarily because a number of them are likely to be confounded in any given setting. Nevertheless, future design implementation analyses would be well-directed to attempting to determine how greatly any implementation problem can be expected to affect the estimate of gain.

2. Model C is not chosen because Model A is close to what most school districts had already been using for Title I evaluation.

Prior to development of the three models many districts evaluated Title I programs simply by testing treatment students before and after the program and determining whether their gain exceeded some goal or norm. In many cases, the only change that was needed in order to use Model A correctly was the addition of the separate selection measure.

3. Model C is not chosen because it requires adherance to the cutoff point criterion for assignment.

Several administrators at the school district level claimed they want greater discretionary power in assignment both to allow for favoritism and for the possibility that a given test score may be inaccurate. Two administrators stated that they would like to have a "range of cutoffs" rather than a single cutoff point. In the simplest case, one could have two cutoffs. All students scoring below the lowest cutoff would automatically be assigned to the treatment group, all those scoring above the higher cutoff would automatically be assigned to the comparison group, while all those scoring between the cutoffs would be assigned at the discretion of administrators, teachers, and so on. Although this strategy has not been used in Title I evaluation, it is a possibility as will be discussed in Chapter IV.

4. Model C is not used because of premature release of research results which erroneously showed it is a biased model.

One frequently cited reason for not using the regression-discontinuity design provides an illustration of the interaction between social contexts and methodological decisions. Subsequent to the original contract for the development of research models awarded to the RMC Research Corporation, the Office of Education awarded a follow-up grant to the same corporation for purposes of investigating the models in more detail. To examine whether Model C might be biased, the RMC Research Corporation acquired two sets of data used to norm standardized tests. The first was obtained

from the Sustaining Effects Study and consisted largely of scores from the Comprehensive Test of Basic Skills (CTBS) while the second was comprised of the 1977 norming data for the California Achievement Test (CAT) issued by the McGraw-Hill Company. For both of these databases it was assumed that the tested students did not receive compensatory education services. For Model C, cases were assigned to either treatment or comparison groups using an arbitrary cutoff point and an estimate of treatment effect was calculated. Presumably, since no treatment was administered, the analyses should have yielded an average treatment effect in the vicinity of zero.

These tests were carried out in two phases, first on the CTBS data base, and second on the CAT data. The initial results for the CTBS data indicated that Model C tended to show a biased effect in the direction of a positive result. Subsequent to this portion of the study, and before looking at the CAT data, persons from the RMC Research Corporation, in seminars given throughout the country, suggested that there may have been problems with Model C which had not been previously anticipated. When the data from the CAT data base was examined using the same procedures, all three models yielded estimates of effect in the vicinity of zero NCE units. At this point several researchers from RMC began to search for reasons for this apparent discrepancy (Wood, 1979). The re-examination of the CTBS database turned up the

existence of non-normal or skewed distributions. After transforming the data to minimize this skew, the estimates for Model C were recomputed and while slightly positive, approached the expected zero value. While this study is still in progress, the general conclusion seems to be that Model C is not biased, although it may be affected adversely by such things as floor and ceiling effects and abnormalities in the marginal distributions (Stewart, 1980a and 1980b).

A distinction needs to be made between the actual results from this study and the effect of premature release of the results. A number of respondents, especially at the TAC level, stated that to their knowledge RMC had shown that Model C did not work correctly and should be avoided. None of those who mentioned this seemed aware of the subsequent investigation and the change in the overall conclusion. While it is reasonable to expect that eventually this information will be passed along it is hard to gauge the effect which premature partial information had on the selection of research models by school districts. It is reasonable to argue that some districts used this information in deciding against Model C, while others who wished to use Model A cited it to enhance the credibility of their decision.

Even granting that a consistent positive treatment effect may be detected in this study, one must be careful to distinguish between the conclusion that Model C is biased or

that implementation problems (e.g., testing problems, matching problems, etc.) degrade estimates of effect with this design. Each conclusion has different implications. If one concludes that Model C is biased it would be reasonable to recommend suspension of its use. If however, the conclusion is that with typical Title I implementations there are problems which act to degrade estimates, one can address these problems individually and attempt to improve the quality of the results.

5. Model C is not chosen because it is more technically difficult than Model A.

The fact that Model C is more technically difficult than Model A seems to be related, at least in part, to the fact that it is based upon notions of regression lines and projections from regression lines, ideas which are not easily understandable to those who are unfamiliar with research and statistical analysis. In addition, several respondents mentioned that in presenting the three research models, the materials used and the explanations provided for Model C were the most difficult to understand. It is not at all clear whether this is a deficiency in the manner of presentation or a problem which stems from inherent differences in the technical difficulty of the models. Nevertheless, it is a commonly held perception that Model C is more difficult than Model A, that it requires a more technically trained staff and that it is most appropriately used by districts which have ready access to computer facilities capable of computing

the necessary regressions (Bridgeman, 1979). Although all of these considerations have some validity, the investigation of where Model C has been used turned up several districts which used the design despite the lack of sophisticated personnel and facilities. In most cases these districts either pooled their resources with other small districts or hired outside consultants who were able to conduct the analyses.

6. Model C is not used because it requires testing of comparison students.

Model C requires the testing of comparison students. While it is not absolutely necessary that all non-Title I students be tested (e.g., a randomly selected subsample of non-Title I students could be used instead), the model does require more than simply the testing of treatment students. For those districts which use a district-wide annual standardized achievement test this requirement poses no problems. However, for other usually smaller districts, the increased cost of additional testing is a considerable burden.

7. Model C is not used because it requires at least 30 students in the treatment group.

The Users Guide suggests that Model C should not be used unless there are at least 30 students in each group. The basis of this is that with larger samples one can achieve greater statistical precision and more accurate estimates of regression lines (McNeil, 1977). While this is certainly an important consideration, the requirement of at least 30

students seems arbitrary and is technically unnecessary.

8. Model C is not used because it requires that within-group correlations be at least .4.

The Users Guide suggests that Model C should not be used unless the pretest and posttest correlation is at least .4. While it may be desirable, from a conceptual point of view, to have a strong pretest posttest correlation (McNeil, 1977), this is not technically necessary (see Echternacht, 1978). In fact, any measure can be used for the pre-program measure whether it is correlated with the posttest or not. Figure 8 presents an example of the regression-discontinuity design where the pretest and posttest are uncorrelated. In this case the regression line in each group is flat, that is the slope is equal to 0. Here, one has effectively reduced the design to a test between the treatment and control group posttest means. In this case because there is no relationship between the pretest and posttest, one "approximates" random assignment to treatment. It is difficult to conceive of such a case occurring in Title I evaluation because of the desire to assign students to a treatment group on the basis of some measure related to achievement. Nevertheless this sample is included primarily to illustrate the fact that a low pretest-posttest correlation, should it occur, will not by itself result in biased estimates of treatment effect.

9. Model C is not chosen because it can cause individual schools to lose Title I teacher positions.

Another problem which occurs in relation to the sharp cutoff point concerns the allocation of Title I teachers to schools. From one year to the next, the number of children within any given school within the district who are eligible for Title I services might change considerably. If Title I teachers are assigned to schools on the basis of the number of children who will be served, there is a possibility that a given school will lose Title I teachers from one year to the next. Few schools are likely to accept unchallenged a reduction in the number of teachers they are allocated. Other types of allocation procedures could be developed as in one case, where a very creative approach which capitalized on an already existing busing program equalized teacher allocation by means of moving students from one school to another.

10. Model C is not chosen because the SEA discourages its use.

Another reason cited for not using the regression-discontinuity design concerns the fact that in many states the SEA strongly recommends the use of another design, most often Model A. While the decision to use a particular research design technically is the option of the LEA, the SEA does have the right to review and approve program proposals. While no respondent stated that the SEA rejected their use of Model C, many mentioned that the existence of a stated or implied state policy favoring Model A did have an effect on their decision.

Reasons Cited in Favor of Using Model C. Fewer reasons were listed as favoring the use of Model C. They are:

1. Model C is used because it is conceptually closest to the idea of Title I training.

Title I programs are supposed to be given to those students who need them most. More than the other two models, Model C provides an explicit quantified measure of such need and a clear decision rule for the allocation of service. There are other ways to exploit the correspondence between the regression-discontinuity design and allocation procedures used for Title I. For example, each school district must designate which schools are eligible for Title I programs on the basis of a quantified measure of poverty. If a cutoff on this measure is used to assign the schools, one can conceive of a regression-discontinuity analysis of the effects of Title I training at the school level. Variations of this type are discussed in Chapter IV.

2. Model C is used because it fits in well with annual district-wide testing programs.

Those districts which have annual testing programs find that no additional testing is required for Model C. The test given in each year acts as the posttest for the current year and as the pretest and selection measure for the years to follow. In the typical implementation of Model A, all students in the school district are tested (for example in the Spring), those students selected for Title I service are

pretested (for example in the Fall), and all students are posttested the following spring. Thus, in the annual cycle, the Spring testing provides both the selection measure for the subsequent year and the posttest for the current year. However, it is necessary to include a separate pretest. Essentially, this issue involves a trade-off between the two models. Model A requires less testing in that only treatment students are given the pretest and posttest (although all are given the selection test). Model C involves less testing in that a separate selection measure is not needed. Depending on the type of district testing program which exists, a given district might find it less costly to use either Model A or Model C.

3. Model C is used because it is perceived as methodologically stronger than Model A.

Another set of reasons cited in favor of using Model C is related to perceptions about the quality of design. Those districts which employ researchers who have been specifically trained in research methodology are more likely to be aware of the academic history of Model C and of its advantages from a methodological viewpoint. In addition, because of this history and the common perception within Title I evaluation circles that Model C is the most technically difficult of the three models, a degree of higher status might be accorded those districts which use the model. One Title I evaluator from a large metropolitan school district said that he felt his district should "lead the way" in

attempting to implement more technically difficult research designs. This is more likely to be the case for districts which have well-trained staffs and sufficient computer facilities.

Summary of Applicability Issues. The regression-discontinuity design is used for Title I evaluation, although infrequently. While it is certainly appropriately used in this context, it tends to be judged relative to Model A. From the typical school district's perspective, Model A is easier to understand, easier to use, less costly, and likely to yield "favorable" results. If both models were of similar methodological quality Model A would clearly be the better. The issue is clouded by two facts. First, even if ideally implemented, Model A yields biased program effect estimates. Second, although unbiased in the ideal, Model C estimates can be seriously distorted due to poor implementation.

Therefore, the issue of which method is generally preferable hinges on the seriousness and prevalence of implementation problems with Model C weighed against the inherent bias and implementation difficulties of Model A.

Implementation problems can be reduced in three ways. First, one can simply be more vigilant about implementation. Design implementation analysis helps to indicate the major problems. For example, if attrition rate is a serious degrading factor, more time can be spent on following up on "missing" cases. Second, implementation problems can, to

some extent, be explored as part of the statistical analysis. If floor effects are suspected, one might estimate gains using all the data and after discarding selected portions at the lower end of the pretest distribution. If the effect is similar for both analyses one can conclude that gains are not seriously distorted by the potential floor effect. Finally, implementation problems can be anticipated and minimized by improving the research design. Especially useful would be the inclusion of random assignment to groups, at least for a portion of the pretest range. If one can minimize Model C implementation problems in these ways, it is reasonable to argue that it is methodologically stronger than Model A.

Assignment Issues

One of the distinguishing features of the regression-discontinuity design and, in fact, the characteristic which sets it apart from other pretest-posttest designs, is the use of a pretest cutoff value for the assignment of students to treatment or control groups. The need for adherance to a strict cutoff poses serious administrative and political problems for a school district. The requirement of a sharp cutoff has been clearly stated within the Title I evaluation literature. The reader is referred to Campbell, Reichardt and Trochim (1969), Trochim (1980) and Goldberger (1972) for more detailed discussion of the rationale for this requirement.

The implementation questions of primary importance in the section are:

1. What is the typical cutoff value?
2. How do districts select the cutoff?
3. What measures are used for assignment?
4. How well do districts adhere to the cutoff criterion?
5. In what way do assignment problems affect the estimate of gain?

Each issue is discussed below.

Placement of the Cutoff Value. In their annual evaluation reports to the State of Florida, all but two school districts expressed the assignment criterion in terms of a pretest percentile or stanine. Thus, a school district might report that all students scoring below the 20th percentile on the pretest would be eligible for Title I training while all those above that percentile would not. The other two school districts in Florida reported the cutoff criterion as a pretest scale score value. Here, the rule might hold that all students scoring below a CTBS scale score value of 340 are eligible for Title I services while those scoring above that value are not. The way in which the cutoff value is reported to the state, however, is not necessarily indicative of the way in which the assignment procedure was handled within the school district. For example, a school district which reported the cutoff as its pretest percentile might actually have used a pretest scale

score and simply reported in terms of percentiles for convenience. Within the State of Florida the median pretest percentile of the cutoff was 28.6. The lowest percentile over the two year period studied is 7.0 while the highest is 50.0. Thus, there never was reported a case where over half the eligible students were assigned to Title I services.

Selection of the Cutoff Value. Regardless of the manner in which the cutoff score is reported, it is useful to look at how a district chooses it. The majority of respondents at the local level stated that they selected the cutoff value so that they will be able to have the maximum number of students served given the available resources. However, a large number of school districts reported the exact same cutoff percentile in 1978 as in 1979 suggesting that there is either a remarkable consistency in the available resources and the number of eligible students, or that school districts have some latitude from year to year in the way in which they handle allocation of funds. One school district reportedly chooses its cutoff point in terms of standard deviation units on the pretest. Thus all students scoring below, say, one standard deviation unit below the pretest mean are eligible for Title I services while all those who score above are not.

It is worthwhile to investigate in more detail how such allocation procedures are managed at the district level. The remarkable consistency in the cutoff percentile from year to year suggests a fair amount of discretionary power in the

allocation of funds on the part of the district or the possibility of the existence of a "reported" cutoff score which is different from the "real" cutoff used for assignment. Part of the problem may be that in order for a school district to truly allocate on the basis of the pretest distribution, that is, to choose the cutoff on the basis of the available funds, they need to know the shape of the pretest distribution in advance. If the proposals for Title I funds which are submitted to the state are required before a district has the time to examine this distribution, it is possible that they would give a reasonable guess for the cutoff percentile and change this if it proves unmanageable.

Measures Used for Assignment. There are essentially three ways in which a school district can assign students to treatment group while adhering to a sharp cutoff. First, they can use the same test or different levels of the same test for both the pretest and posttest. Second, they can use different tests for pretest and posttest. Third, they can use a combination or composite of several tests or measures as the pretest and some other measure as the posttest.

Most school districts use the same test for the pretest and posttest, sometimes relying on different levels of this test for each. There are special cases where this is not feasible. For example, for students entering the first grade there is not likely to be a standardized achievement test

which would be appropriate. In this case the district often relies on a "readiness" test of some sort for the pretest and a standardized achievement test for the posttest. Another case where using the same test for both testings is not feasible is when there is no readily available standardized achievement test which is appropriate. Thus, in Puerto Rico, because there is a dearth of standardized achievement tests written in Spanish, the pretest is often a measure of grade point average while the posttest will be a locally developed achievement test. One reason for reliance on the same test or different levels of the same test for pre and post-testing is the advertised requirement of a pre-post correlation of at least .4. Although this is not technically required as mentioned earlier, some districts report reticence to using different tests for pre and post-testing because of the fear that the necessary correlation would not be obtained and the analysis would be invalid.

One of the more interesting potential assignment procedures involves the use of a composite of several measures for the pretest. Since a major problem in adhering to the strict cutoff seems to be the impressions of teachers, parents, or administrators that certain children were misassigned, it would be useful to attempt to quantify their implicit assignment strategies and incorporate them directly along with a pretest. One can conceive of an assignment measure which is a weighted average of a pretest value, a

grade point average, a quantified estimate of skill from the classroom teacher, and so on.

The major problems with such a procedure are that it requires a good deal extra effort, it can be arbitrary, and it makes explicit an assignment decision which for political reasons might in specific cases be difficult to justify. A respondent from one school district which investigated the possibility of a composite assignment variable claimed that the major problem occurred in trying to determine the relative weighting of the different variables. Other respondents stated that it is likely, on the basis of their experience, that the assignment to treatment group would not differ greatly regardless of whether standardized achievement test scores or teacher's subjective impressions are used. Some work which has been done on the degree to which teachers can correctly estimate the performance of students on standardized achievement tests shows that there is a strong correspondence between the two measures (RMC Research Corporation, 1979). However, much of this work has been informal and has gone undocumented (Hill, 1979).

The possibility that subjective impressions might be quantified and subsequently incorporated into the assignment strategy offers great potential and should be investigated further. Such a strategy is not without its hazards however. One school district which has tried it cited dissatisfaction with the results primarily because of conflicting advice

given to them by their regional TAC. In this case, the district attempted to combine an achievement test score with the classroom teacher's impression of each student's skill. According to local officials, the regional TAC advised them to transform the teacher rating from a continuous scale to one where all students scoring below a given value received a score of 0. When this transformed variable was added to the pretest score it obviously resulted in a nonlinear assignment measure and led to difficulties in the analysis. In the subsequent year the TAC corrected this advice and suggested the use of a continuous scale. This example seems indicative of the lack of experience with composite scales in the Title I system. Nevertheless, with more experience this approach to assignment under Model C appears to hold promise.

Adherence to the Cutoff. Virtually every school district which was contacted initially claimed adherence to the strict cutoff value in assignment. On further examination most indicated that this "adherence" was strictly nominal. Sometimes district evaluators are aware of the extent of misassignment while at other times this is disguised by the procedures which are used to analyze the data. For example, in some cases, what was advertised as strict assignment turned out to be so only because the analyst directly excluded or disqualified cases which were misassigned relative to the cutoff point. In some cases, the extent of the misassignment

is more subtly disguised. Several districts, at the time of data analysis, simply took the pretest and posttest scores at hand, and considered all those with pretest scores below the cutoff as participants in the treatment group, while those above the cutoff were considered comparison students. Unless this formal decision rule was actually adhered to within schools and/or classrooms, there is likely to be come misassignment present. Classroom records of which students actually receive Title I services need to be compared with the assignment by means of the cutoff or it is impossible to detect directly the degree to which misassignment occurs. Some districts have formal procedures for comparing the students serviced with assignment by the cutoff but even in these cases the degree to which the information agrees is seldom calculated.

Another problem related to adherence to the cutoff in assignment concerns the timing of feedback from the district level to individual schools and classrooms. For example, if test scores are processed at the district level and this information is not forwarded to the schools or classroom teachers prior to the beginning of Title I service, the teachers may begin service to those who they believe will be eligible and this will often be in disagreement with eligibility by the test scores (Hill, 1979). It is difficult to assess, given the data at hand, the degree to which such problems occur within Title I evaluation.

By far, the most frequent form of misassignment is through formal procedures by which the assignment can be challenged. Almost every district has some mechanism which makes it possible for a teacher, principal, or parent to challenge the assignment by the cutoff criterion. By far the most common method for handling challenges involves retesting of the student in question. Several problems occur at this level. First, as one respondent stated, because of the number of challenges which are submitted annually in his district, as many as "600 or more among the grades" (Visco, 1980), retesting of all these students is prohibitive. Second, there is often no limit put on the number of times a student may be tested. Thus, as one respondent reported, a teacher who challenges a particular child's assignment could have that child retested again and again until the score falls below the cutoff value. Third, the time of retesting is critical. As mentioned later, it is well-known that average achievement test scores in the Fall tend to be lower than those in the Spring, all things being equal. As a result, if a teacher wishes to challenge the assignment of a given student by the Spring pretest and administers a Fall makeup test the chances are greater that the student would have a lower score and be more likely to qualify for Title I services. Along with this problem one must consider the difficulties in equating scores given at a Fall testing with those given in Spring. It is not clear whether one should

use the Fall test scores as they are or to attempt to transform these scores to a scale which is similar to the Spring scale. Most likely, students who are challenged into Title I programs tend to have pretest scores just slightly above the cutoff value, (although this is not necessarily the case). One might attempt to verify the existence of such challenges by looking for distortions in the pretest frequency distribution in the vicinity of the cutoff score.

It is not always clear that the existence of a challenge is reported to the district level. For example, a challenge could theoretically be handled entirely within a school or a classroom. Thus, a teacher who challenges the original test score could administer a makeup test, grade that test, assign the student to Title I service, and never inform the district evaluator of the procedure. Part of the problem then is related to the lack of documentation of such cases while the other part concerns how it should be handled in the analysis. In any case, most districts report using such procedures and believe that the frequency of challenges poses significant problems for the data analysis.

The existence of formal challenge procedures tends to disguise political factors related to assignment. It is not surprising that this is the case when one considers that the regression-discontinuity design tends to eliminate the discretionary power of teachers and school administrators. Only in-depth interviewing and observation is likely to turn

up some of the subtle ways in which challenge procedures can be used to political advantage. This is illustrated in one example cited by Visco (1980):

> One principal apparently insisted on having (with few exceptions) only certain challenges approved, namely those submitted from a particular group of classroom teachers (the principal's old friends, I'm told). When the Coordinator spoke to the principal regarding this, the principal emphasized that it was her school and she had final say. Apparently, she was right.

Misassignment relative to the cutoff point has long been recognized as a major difficulty for the analysis of the regression-discontinuity and the "fuzzy" regression-discontinuity design. It is important to note that there are a variety of possible causes for misassignment, all of which might operate in the same setting. Teachers or administrators might, for example, use the challenge procedure as a vehicle for showing favoritism to certain students. Another source of misassignment is likely to be related to administrative error. Errors in test correction, score reporting, and the like, might lead to incorrect test score information upon which the assignment is based. Finally well-intentioned teachers and administrators might be unwilling to deny service to students who they feel are deserving of that service. An unusual example of this type of well-intentioned misassignment is documented by Visco (1980):

> At least one teacher (maybe it was two) told me that she would not challenge certain students "out" of the program even though they did not

> require service, because if they officially left the program they were no longer eligible to receive clothing through the Title I clothing component-grant. Instead she just stopped servicing those students and never reported it to Title I administration....Obviously, it affects the evaluation, but I have no way of knowing the extent to which it occurs.

Assignment Problems and the Estimate of Gain. An important question is whether program estimates are distorted due to assignment problems. Two possibilities are discussed here. First, if the pretest-posttest distribution is non-linear, the placement of the cutoff can affect the gain. Second, the manner in which misassigned cases are treated in the analysis can distort the program effect. We can indirectly examine the effect of cutoff placement by looking at the average gains for those districts which used a low cutoff percentile versus those using a high one. When the 273 gains obtained from the State of Florida are dichotomized at the median cutoff percentile, those school districts having a cutoff below the 28th percentile had an average gain of -2.31, while those who had a cutoff above the median of 28.0 had an average gain of -2.42. It is not clear then that there is any difference in the estimate of treatment effect for districts which differ in their placement of the cutoff.

More important is the impact which including or excluding challenge cases from the analysis might have on estimates of program effect. Consider the following hypothetical example. Let us assume that all of the challenges in a district are in the direction of into the

program. Students scoring just above the cutoff would be the most likely candidates. If the challenges are reasonable, these students would be the ones who score immediately above the cutoff and would be expected to do poorly on the posttest. This group is indicated by the "darkened" portion of the graph in Figure 9. Further assume that these students are retested as part of the challenge procedure and that their retest scores fall below the cutoff point.

First, consider the consequences of including them in the analysis using their retest (below the cutoff point) scores. This will have the effect of increasing the number of cases immediately below the cutoff point which also tend to fall below the regression lines as shown in the graph of Figure 9. The addition of these cases in the treatment group will attenuate the slope of the regression lines for the group. Similarly, the removal of these cases from the comparison group will also lower the slope of their regression line. One would expect to find a negative pseudo-effect in the hypothetical case if retest scores were used.

Second, if the challenged cases are simply excluded from the analysis altogether, one expects that only the comparison group regression line will be affected. As above, the slope of the line will be attenuated somewhat. Again, a negative pseudo-effect will result, although it will probably not be as great as that caused by using retest scores. The problem is not avoided by using the original test scores in

the analysis because this is a "fuzzy" design which is likely to yield biased estimates for that reason.

One can construct a similar scenario for challenges out of the program. Again, if we assume that these cases are likely to be students scoring just below the cutoff who would be expected to do better than average on the posttest, we would find negative pseudo-effects (because their removal from the treated group and their addition to the controls acts to attenuate the slope of the line in each group). Therefore, if it is reasonable to believe that students challenged into the program are those who are likely to do worse than expected on the posttest, while ones challenged out are likely to do better than expected, in both cases we would expect to find pseudo-effects, and these will tend to be negative. This is the reason that misassignment was mentioned earlier as a likely factor, at least in part, for the observed pattern of negative gains in Title I evaluation. There is no clear way to avoid biasing program estimates when challenge procedures are used although the argument above seems to indicate that excluding such cases from the analysis may yield less bias than using their retest scores would.

The issue of misassignment under the regression-discontinuity model is important then for a number of reasons. First, the results of the study can be distorted. Procedures need to be developed for determining the extent of

this distortion. This might involve looking at the distribution of challenges relative to the pretest scores to determine whether it is homogeneous or tends to show a coincident jump in the vicinity of the cutoff score, or the analysis of challenge cases separate from those which are correctly assigned, or both. Second, an investigation of the reasons for challenges and the relationship of such challenges to test distributions is essential for developing reasonable models for "fuzzy" regression-discontinuity designs. These models would be useful for the generation of simulated data for purposes of testing new and alternative analytic procedures for the "fuzzy" case. Finally, it is important that the occurrence of such problems be routinely documented and that their implications be communicated to those responsible for evaluation at the school district level.

Measurement Issues

The assumption implicit in most research is that the data which is collected for analysis is accurate and has been collected correctly. Put in other terms, we assume that the tests or measures have reliability, validity, and are free of bias. Because most Title I evaluation relies upon the use of standardized achievement tests, issues related to who administers the test, when it is administered, which test is given, the scales on which scores are reported, and the possibility of falsified or "revised" scores are important in considering design

degradation.

For convenience, these issues are divided into four topics:

1. Test Administration
2. Test Characteristics
3. Test Problems
4. Data Maintenance and Access

Each of these is considered especially in terms of their potential for distorting estimates of program effect.

Test Administration. First, it is important to consider how the test is administered and the problems which might arise in the administration. School districts differ considerably in the manner of administration. In some cases the school district requires that all schools be tested during the same week under specific conditions. This might include testing of all students within the school in the same setting, such as an auditorium or a school gym, or might allow each class to test within its classroom. Sometimes the test is administered by district personnel while at other times individual classroom teachers are in charge. Similarly, the test can be corrected by the teachers themselves, by school personnel, or at the district level. An enormous number of factors can affect the testing context, but for our purposes we are primarily interested in those which differentially affect treatment or comparison group students.

Any conditions other than pretest scores which serve to

target students as either potential treatment or comparison group participants should be considered suspect. If, within a particular classroom Title I students tend to sit in a pre-assigned section of that classroom, it is possible that this "segregation" can lead to differential attitudes or behavior at the time of testing. Under these conditions, for example, Title I students from the previous year, when taking the next year exam may have more anxiety because of fear of assignment in Title I training. In other ways, the teacher or test administrator might subtly communicate differently with Title I and comparison students. This can manifest itself in the instructions, in offhand comments or in any behavior which tends to confirm the segregation. Similarly, control students might display more anxiety concerning the test because of their fear of being assigned to Title I programs. Generally, any factor which creates different expectations or anxieties in the different treatment groups can have an adverse affect on the research which is conducted. These factors tend to change the scores for one group and not the other and at the point of data analysis might be confused or mistaken for an effect of Title I service.

Another troubling factor concerns differences in testing conditions between classrooms or schools. Again, such factors will only affect the results if they are manifested differentially within subgroups. Thus, if those classrooms with higher proportions of Title I students also happen,

through circumstances or by plan, to have systematically worse or "better" facilities, classrooms, and so on, the results might be affected.

Other, seemingly more mundane problems related to the processing of the test might affect results seriously. This is especially true when a large number of people have responsibility for handling the test administration. Thus, when the tests are processed by classroom teachers, problems related to miscoding of information often arise. Several respondents mentioned problems of this nature such as miscoding of identification numbers, names, sex, age, and so on. Some problems, such as the occurence of 80 year old first graders (Visco, 1980) can be flagged at the district level, although it is not clear that most districts have procedures capable of doing this. Especially important in the case of Model C is the coding of student names or ID numbers, because it is necessary to match up pretest and posttest scores for each student. Some of these problems will be discussed in more detail later.

Another problem related to administration concerns when in the school year the test is given. It is commonly found, for example, that average achievement test scores tend to be lower for Fall testings than for the previous Spring testings. This is sometimes attributed to a degradation of skills due to summer vacation, lack of practice, and the like. While this is a far more serious concern under Model

A because using a Fall pretest instead of a Spring pretest will tend to inflate gains, this problem can affect Model C results if, for any reason, the "loss of skill" is differential between the treatment and control groups.

Test Characteristics. Another problem related to testing concerns which test is given. For this study, it will be assumed that the test is an appropriate reflection of the skills which are likely to be affected by Title I training. This is an important and difficult issue, and one that lies outside the scope of this work (see for example, Linn, 1979). More germane are questions related to the level of the test, procedures for makeup testing, and so on. One issue which is important in determining the use of Model C concerns the need to administer non-English exams to some students. Because there is a lack of achievement tests in languages other than English which have been appropriately normed, such places as Puerto Rico, the Northern Mariana Islands, Guam, and so on, are more likely to use Model C which does not explicitly require the translation of scores to norms and thus encourages use of locally developed tests. This might also be a problem in large metropolitan areas where a considerable proportion of the student population is composed of foreign-speaking students.

An issue of more widespread importance concerns the level of test which is administered. Most achievement tests which are currently marketed have several different levels or

forms of the test designed to measure different levels of achievement or skill. Usually the test maker will recommend a particular test level for a given age group or grade level. While the assigned test level is likely to be appropriate for the majority of students in any grade, it will tend to fail precisely where the most accurate information is desired for purposes of Title I evaluation, that is, at the lower end of the testing distribution.

The issue of choosing the appropriate test level is related to the potential for floor and ceiling effects, and the effect of the chance level of the test. Since most Title I students come from the lower quarter of the test distribution, if that test is too difficult one is likely to find a floor effect differentially affecting the treatment group. Several suggestions have been offered to avoid test level problems, most noteably the ideas of "out-of-level" or functional level testing. This is commonly implemented in one of two ways. First, if the district is concerned about the possibility of floor effects they could simply require that each grade be tested at the next lower level of test, although this obviously increased the potential for ceiling effects. A second approach is to allow individualized testing. Here, the classroom teacher or some other appropriate and knowledgeable administrator, assigns whatever level of test is most reasonable for any given student. Within the same grade several different levels of the same test might be

used for different students who are believed to have different ability. The major problem with individualized testing stems from the fact that at some point these test scores must be converted to a common scale for purposes of analysis. Thus, one must rely on "extended" standardized scales provided by the test makers and obtained from norming samples of students who received multiple levels of the test. While the use of out-of-level or functional level testing holds promise for Title I evaluation, especially when Model C is the method of choice, one must be careful to anticipate the special problems which occur when moving from on-level testing to one of these procedures.

A related question concerns the scales in which the scores are reported and on which the analysis is based. Most test makers will report test results on a variety of scales such as raw score, extended standardized score, local standardized score, and so on. The issues involved in converting scores from one scale to another for the most part are outside of the scope of this work. As a general recommendation it is probably best in conducting an analysis to rely on the raw scores for developing estimates of gain and to convert these estimates into NCE units for reporting purposes. By using the raw scores one can for the most part avoid concerns about the quality of various standard scales provided by the test maker.

Another issue related to test characteristics stems from

the finding that students first entering grade school are likely to differ considerably in their ability to read or to handle simple mathematics problems. However, after several years of education, it is more likely that the ability levels will become more similar. One might expect, as a result, that estimates of treatment effect will be higher at higher grades where tests are more likely to reflect ability better. There is some corroboration for this notion in the analyses of Model C results from the State of Florida. Across all Title I programs (n=273) there was a tendency for test results to be more positive for the higher grades, but whether this is due to developmental phenomena or other factors such as better measurement, better training, loss of less intelligent students, and so on, is not certain.

Another issue related to the quality of test information concerns the fact that standardized achievement tests are designed to discriminate best near the middle of the test distribution. For the most part these tests are constructed to yield good estimates of average performance for any given group. Thus, the tests are most fallible at the tails of the distributions. This is exactly where the most precise information is desired in Title I evaluation because treatment group cases tend to be confined to the lower end of test distribution.

Test Problems. Some of the most serious measurement problems for Model C result from factors which distort the

pretest-posttest functional relationship and consequently, estimates of program effect. Two such factors, floor and ceiling effects and chance level performance, are described here.

Floor and ceiling effects result, respectively, from a test which is either too hard or too easy for the group in question. For example, if the test is too difficult, a "floor" effect results because most students will receive a low test score and consequently the distribution will be positively skewed. The test will not sufficiently discriminate between students of different ability who all receive low scores. The issue is especially complicated here because it is possible to have either a floor or a ceiling effect or both on either the pretest or the posttest or both. In Figure 10, several expected pre-post patterns for various combinations of floor and ceiling effects are depicted. The figure shows that if there is a floor effect on the posttest the treatment group slope would be likely to be lower than the comparison group slope. Similarly the same pattern would result if there is a ceiling effect on the pretest. Either of these could occur in practice. To consider the likelihood of such occurrences one must look at the test or tests which are used. Assuming that the same test is used for both the pretest and posttest it seems more likely that one will have a floor effect on the pretest and a ceiling effect on the posttest (assuming that subjects grow or gain in skill between

the testings). Thus, in the case where the same test is used both times it seems unlikely, or less likely, that floor and ceiling effects would yield the pattern of negative gains commonly attained.

The situation becomes more complicated if different tests or different levels of the same test are used for the pre and posttest, as is commonly the case in Title I. Here, depending upon the difficulty of either task, one could get the floor effect on the posttest or the ceiling effect on the pretest which could yield a lower treatment group slope and result in a negative gain. Thus, for any given analysis, the likelihood that a ceiling or floor effect results in a lower treatment group slope must be assessed within the context of the tests which are used. As a potential source for the observed pattern of negative gains mentioned earlier these effects can not be ruled out.

Another measurement related problem which could result in a lower treatment group slope than control group slope and consequently in negative gains, is related to the chance level of the test. The concept of "chance level" can be best understood through a simple example. Consider a multiple choice test of 100 items with each item having four possible answers. If a respondent were to guess on all 100 items one would expect by random chance alone that the score would be 25. Thus any student scoring in the vicinity of or lower than a score of 25 could have been guessing throughout the

exam. If a student guesses on the pretest and either does or does not guess on the posttest there should be no statistical relationship, or correlation, between the two tests. As a result, those pretest scores in the vicinity of the chance level of the test on the pretest are likely to be less related to the posttest than are those having other pretest scores, assuming that a portion of those persons were guessing. This is illustrated in Figure 11.

It is important to recognize that there is a difference between a chance level effect and a floor effect. A floor effect occurs when there is value on a test below which students cannot score even though their true ability would indicate that they should. A chance level effect occurs when students respond to questions by chance or guessing. A floor effect should be evidenced by an absolute flattening of the distribution at the floor value while a chance level effect flattens the distribution but allows for test scores on both sides of the chance level. Thus it is theoretically possible to have either a floor or chance level effect or both in any set of data.

The relationship of the chance level of the test to the cutoff score in the regression-discontinuity design has been investigated by Visco (1980). Generally, he finds that in most cases the cutoff score is in the vicinity of or lower than the chance level of the test. While it is impossible to say that for these cases all those in the treatment group

were guessing, it is not unlikely that a large number of students were. In this regard, it is important to determine whether the instructions of the test explicitly encourage guessing. In addition, in the 273 Florida analyses the median value for the cutoff was about the 28th pretest percentile and one can infer that it is quite likely that the cutoff is often very close to or even below the chance level for the test. Thus in the case of Title I evaluation it can be assumed that this phenomenon can result in the lowered treatment group slope and, at least in part, the resulting pattern of negative gain estimates described earlier.

Data Maintenance and Access. One area which has received little attention from methodologists concerns processing of test information. Many school districts had little experience prior to the initiation of the current Title I evaluation system with processing vast amounts of achievement test data for purposes of analysis. There are several types of problems which can occur in this context. First, are problems of miscoded information which can result from the students themselves or from the test correctors. Second, are errors in processing the test made by the test producer or the company which markets the test. Third are errors made at the school district level such as calculation mistakes or errors made when using tables to convert scores. Fourth, are problems related to accessing the data and correctly processing the data for purposes of analysis.

Problems of coding are illustrated well by the difficulties involved with trying to match pretest and posttest scores by computer. For example, it was necessary to match the data obtained from Osceola County, Florida (see Chapter III for a fuller description) by name of student before conducting the analysis. The test that was used allowed 12 columns for the last name, 6 columns for the first and 1 column for the middle initial. Ideally, one would like to match students on the basis of an exact correspondence between all 19 columns of the name field on the pretest and posttest. However, in practice, there are many ways in which the name for the same student can be miscoded on different tests. Some of the problems detected when matching the Osceola County data include:

- Code "long" name one time, "short" name the other:

  SMITH, CATHER
  SMITH, CATHY

- Code middle initial one time, not the other:

  JONES, JOHN M
  JONES, JOHN

- Errors (keypunch or entry) leading to "unlikely" names:

  ADKINR instead of ADKINS
  ALBRECNT instead of ALBRECHT
  MAAK instead of Mark

- Different coding of name due to number of columns allowed:

  SAMANT OR SAMATH for Samantha
  JACQUL OR JACQUE for Jacqueline

- "Joking" or suspicious names:

  MOUSE M
  DUCK D

- Placing middle initial in first name field:

  TERREN A vs TERRYA

- Extension of long name into next field (e.g., if first name and middle initial are Vincent E.):

  VINCEN E VINCEN T

- Coding of abbreviation of first name:

  ROBT instead of ROBERT

- Legitimate name change from one testing to another:

  CLAY to ALI

At the point of actually matching the cases, problems like these can act to prevent a successful match.

One might attempt to eliminate some of these problems by reducing the amount of information used to match, but this will lead to more frequent mismatches. For example, for the first 100 alphabetically sorted cases (including both pretest and posttest) for Osceola County, use of a 19-column match using the complete name (last, first, and middle initial) yielded 21 matches (i.e., 42 cases were paired). When the number of columns was reduced to the first 15 (last name and first three letters of first name) 28 matches were found (i.e., 56 cases paired). Other variables might be included to improve the match such as sex, age, race, school, and so on, but each of these will have its own errors and their addition may considerably complicate the match procedure.

In general, the more information included in the match field, the greater the probability of excluding a legitimate match due to coding errors. Conversely, the less information included, the greater the chances of obtaining illegitimate matches. Perhaps the best way to deal with matching problems is to match as conservatively as possible (i.e., include many matching variables) and, for cases which do not have a match, attempt to determine the reason (e.g., did not take both tests, miscoding of answer sheet by student, error in keypunching of data, etc.) making corrections where possible.

The matching issue is important for two reasons. First, the need to match cases with Model C is often cited as a factor in favor of choosing Model A which does not require matching. Second, there is good reason to believe that matching problems related to miscoding of answer sheets will be more prominent in the treated group which is comprised of lower-achieving students. This leads to differential "exclusions" between the groups which, in turn, have the potential of biasing the estimates of program effect.

A second issue related to data maintenance concerns errors which are made by the test producer or corrector. Most districts which use standardized achievement tests have them corrected by either the test producer or using automated procedures within the district itself. A number of problems can occur at this stage which have serious effects on the analysis. Thus, one respondent reported receiving the

achievement data for the school district from the test company and subsequently discovering that the wrong correction key had been applied in processing the data. Obviously, had this not been detected the results for that district would have been totally erroneous. In another district, pre and posttest scores were matched up by the test producer using ID numbers which were assigned by the district. In this case however, while ID numbers were assigned to all public school students, those who attended private schools were not given numbers. For some reason the test producer separated these data into two groups, those with IDs and those without. Unfortunately, only the data with ID numbers were returned to the school district. The analyses of the Title I programs for that year were conducted and it was not until the subsequent year that the evaluator realized that none of the private school data had been included. It is clear that situations of this nature must be avoided and that procedures for doing so need to be developed.

In an excellent paper, Crane and Maye (1980) examine the frequency and effect of three types of "correctable" errors made at the school district level: calculation errors, misread norms tables, and use of incorrect norms tables. The data consisted of aggregate values reported to the State of Illinois for two successive years (n = 1,788 for 1978 and n = 198 for 1979). They found that the total percentage of aggregates having at least one correctable error was 51.6%

for 1978 and 57.6% for 1979. In 1979, these errors alone result in an overall positive bias of .74 NCE units (with a range of from .43 to 3.20 NCE units depending on the error). In addition, "the effect of undetected errors on a single aggregate (e.g., for a grade within a project in a single school district) ranged from -24 NCEs to 41NCEs!" They suggest that statistical quality control procedures are needed to detect such errors and reduce their influence on estimates of gain.

Another set of problems concerns the school district's ability to access the data. These problems become prevalent when large amounts of data are processed and/or when the data is computerized. Sometimes the problem is exacerbated by the test producer. For example, the CTBS standardized achievement test reports scores in three different computer formats, depending on the level of the test. This is espcially a problem when the format statements for the pretest and posttest differ. The application of the wrong format to a given set of scores will obviously result in erroneous results and sometimes these can be difficult to detect. Other problems in accessing the data result from the processing of magnetic tapes, errors in reading such tapes, and the like.

It should be clear that irrespective of research design the quality of measurement is crucial. Sometimes measurement problems will occur differentially between treatment and comparison group students. Procedures need to be developed

for assessing the extent of these problems and their effect on the statistical analysis. These should include methods for flagging "outliers," miscodings, floor and ceiling effect and chance level problems, as well as procedures for determining the proportion of such errors in each subgroup of interest.

Treatment Issues

It is not the purpose of this work to comment on what types of treatment or program produce what types of results, or to make a statement on the substantive issues involved in Title I service. Nevertheless, there are important issues related to the treatment which can have a detrimental effect on the quality of the results. Three problems of this nature are discussed here:

1. Identifying recipients of treatment.
2. Determining the amount of service received.
3. Determining the type of treatment received.

If unaccounted for, each of these can distort the estimates of effect.

Identifying Recipients of Treatment. Implicit throughout the Title I evaluation system is the notion that Title I service is somehow "standardized," that is, that all students receive service which is "of a kind." This is evidenced by the existence of a system for national aggregation of gains; it would be unreasonable to aggregate gains from different

types of programs. Within a particular program students are usually divided into treated or comparison group, implying that on the average they are homogeneous with respect to experiences in their condition. In practice, however, the assumption of standardized treatment is unreasonable and the analysis will be improved if the actual service can be more accurately described.

Many of the problems related to measuring the treatment are involved with the seemingly mundane task of determining who actually received service. This was mentioned above, for example, in citing the difficulties which arise at the district level when one attempts to match up who actually received treatment with who should have received treatment according to the cutoff point criterion. The amount of error in documenting who actually received treatment is difficult to determine given the data at hand. However, several respondents cited examples of cases where the assignment to treatment was incorrectly reported. For example, a student who was assigned to treatment by the cutoff and who is listed as having received treatment, may not have received it because the classroom teacher felt that it was not warranted and never informed the evaluator. Similarly, some Title I teachers may simply have too many students assigned to them and may, as a result, have to resort to putting some on a "waiting list" to be serviced later in the year, if at all. Often these cases go undocumented and it is assumed that all

students designated as receiving Title I service were given the same amount and type of treatment. Equally difficult to detect and identify are those students who did not receive service. For example, those students who began the school year in a designated Title I school but transferred to a non-Title I school may have been eligible for service on the basis of their pretest score, and may be listed as receiving such service, even though they did not. As of many other problems in this context, the extent of the problem is unknown and its effect on the analysis is undocumented.

Amount of Service Received. Even if all students could be correctly identified as having either received Title I service or not, the amount of service which is received is difficult to determine. Most districts report that students receive service for a given number of minutes or hours per day or week. This figure is usually illusory and in fact the amount of service received tends to differ from student to student. For example, a Title I teacher is likely to show favoritism or give special instruction to those students who most need or appreciate it. Little, if any, attempt is made to document the degree to which students do not receive service due to such factors as absenteeism, holidays, canceled sessions due to school plays, field trips, weather, and so on (Visco, 1980). In addition, there are systematic factors which tend to reduce the time available for instruction such as transportation time, setup time, clean-up

activities, and so on.

In general, it would be desirable to document the degree of treatment received and to incorporate this information directly into the analysis. Some guidance on how treatment implementation information can be incorporated into the analysis is provided by Boruch and Gomez (1977).

Type of Treatment Received. A major problem in documenting treatment-implementation involves determining what type of specific instruction students received. While it is fairly reasonable to assume, for instance, that students who are supposed to be receiving reading instructions are in fact getting some type of instruction related to reading, the type of instruction might differ considerably from one class to another. Any attempt to document the difference in instructions or to standardize them within a school district is likely to improve the quality of the evaluation, although it may pose political and managerial difficulties.

Another problem related to type of instruction concerns the setting in which instruction is given. The State of Florida, for example, requires each district to report the setting under one of five classifications: self-contained classroom, regular classroom, pull-out small group, pull-out individual instruction, and other. This is certainly an improvement over no information at all, but falls short of a good description of the setting for instruction.

The most common setting is the in-class approach which involves the Title I teacher providing service to Title I students in the back or corner of a room while regular classroom activities continue. Several respondents indicated that although this might be the accepted practice it may or may not be popular with any given Title I teacher. The essential problem is whether one should provide instruction within the classroom (in which case the instruction is likely to be intrusive or disruptive of classroom activities), or whether one should pull out students for instruction in another room (in which case students are more likely to be labeled or stereotyped in a negative way). In any event, issues related to the treatment setting have important implications for the analysis and should be documented where possible.

Another problem related to the type of treatment received concerns the instruction given to comparison students while Title I students are receiving service. Obviously, the students do not remain "fallow" while Title I students receive service and this time is often dedicated to additional instruction in social sciences, the arts, and so on. The level of instruction may, in fact, be raised due to the removal of the least advanced pupils. The problem which arises here concerns whether any skills related to those being communicated to Title I students are also being communicated to the comparison students. This would tend to

diminish their comparative quality and reduce the differences which one might expect between the two groups. For example, if comparison students are receiving instruction in social sciences and this instruction also serves to improve their skills in reading, one would expect that the Title I students as a group would gain less relative to the amount the comparison students would gain had they not received such instruction.

Because of the nature of the subject matter, one might expect that this would be a greater problem for reading programs than for math programs. This is because other skills tend to generalize to reading ability more readily than to math skills. On the basis of this alone we might expect that the gains found for Title I math programs might be less negative than those found for the reading programs. In fact there is evidence for this proposition in the results from the State of Florida. There, the average gain at the cutoff for all reading programs was -3.32 while for math programs it was -.56. One must interpret such differences cautiously as they could possibly be attributed to other factors such as better test reliability for math than for reading. Nevertheless, it is important to recognize how the existence of the competing forms of the treatment or instruction might tend to degrade the comparison.

Another problem in this same vein concerns the existence in many school districts of state or locally supported

compensatory education programs (Visco, 1980; Parkes, 1980). For the most part districts attempt to keep such programs separate from Title I programs by incorporating them into different grade levels. Nevertheless, there are occasions where these treatments overlap in the same grade level, school, and classroom. In one such case, students were assigned Title I service if their pretest score was below a given cutoff, but those students within this eligible group who scored at the lowest end of the pretest distribution were often put into more intensive locally funded compensatory education programs. Unfortunately, no second lower cutoff was used to assign these eligible students to the local programs. As a result, while assignment to group may have been sharp relative to a cutoff value, assignment to different levels of treatment was not, and the analysis was consequently complicated.

The determination of the type of treatment which is received needs to be studied in more detail. In fact, this was a sentiment which was seconded by many of the respondents, although viewed in slightly different terms. Many respondents felt that the emphasis within the Title I evaluation system on outcome evaluation or the estimate of gain was misplaced and that more work should be done on studying the process of Title I training. This would involve a more careful look at the type and amount of treatment which is administered and is to be encouraged especially if this

information will be coupled or incorporated into the analysis of gains.

Exclusion Issues

In conducting a data analysis it is often necessary to exclude or disqualify certain cases for a variety of reasons. Sometimes the cases "exclude" themselves through attrition from the program, that is, for one reason or another it is impossible to obtain both a pretest and a posttest score. In other cases the data analyst decides to exclude certain subgroups from the final analysis because they do not represent the target group of interest or for some reason their data is suspect. Thus, what is included here under the title "Exclusion Issues" includes those groups which are disqualified for one reason or another as well as a discussion of the process of defining the final data which will be used in the analysis.

Three types are considered here:

1. Background and Prevalence of Exclusions.
2. Common Exclusions.
3. Data Processing Issues.

Very little formal work has been done on the systematic effects of such exclusions, and in fact the wide variety of decisions which must be faced by a school district evaluator suggests that the task at times can become overwhelming. Many respondents, especially at the local level, reported

that problems of this nature pose tremendous difficulty and that there was little formal guidance on how they should proceed.

Background and Prevalence of Exclusions. The Office of Education along with the RMC Research Corporation has indirectly commented on the issue of data exclusions, most often within the context of attrition. A recent draft of the Policy Manual (USOE, 1979) states that:

> ...some units (students, schools, etc.) will no longer be available when the second measurement (posttest) is made. This loss is referred to as attrition. If a large percentage of units has been lost, the effects of this attrition should be checked to see if the data are still representative.
> As a general rule, if the posttest includes fewer than two-thirds of the students that were pre-tested, the data should be considered probably not representative. The evaluator can check the representativeness of the data by determining whether:
>
> -the mean score of students having both measurements is significantly different from the mean score of students having only one measurement.
> -certain subgroups in the project (e.g., the lowest or highest scoring students, the lowest or highest socio-economic status) tend to have only one score.
>
> In its evaluation plan (the LEA may wish to consider whether attrition has been a problem in the past. If certain types of students have been lost to previous evaluations, the evaluator may be able to develop a plan to locate those types of students for posttesting.

It is clear that attrition is a primary concern of the Title I evaluation system.

Attrition technically only pertains to those cases where it was impossible to obtain a pretest and posttest score. As

will be shown below, there are many other cases which are routinely excluded from the final analysis even though both pretest and posttest scores are obtained. While a district may have more than two-thirds of its cases with a pre and a posttest, this does not necessarily mean that two-thirds of the eligible cases will finally be used in the data analysis. Little guidance is provided to aid the local evaluator in deciding which data must be used in the final analysis. Some guidance is implicit in the definition of who is eligible for Title I service. Thus, the Office of Education Annual Report for 1979 (USOE, 1979) states that:

> Exceptions to serving the most needy students are: continuation of services to educationally deprived students no longer in greatest need; continuation of services to educationally deprived children transferred to an eligible area in the same school year; skipping children in greatest need who are receiving services of the same nature and scope from non-federal sources; and a school with 75% or more of its students from low-income families may have a project for all of its students (called a school-wide project).

This statement provides some idea of students who may or may not be eligible for Title I services but does not indicate whether or not these cases should be included in the data analysis if treated.

We can begin to determine the extent of the problem by looking at estimates of attrition for the data obtained from the State of Florida. The annual report for the State of Florida (Florida Department of Education, 1979) states that the "unduplicated participant count" (that is the number of

students estimated statewide for the 1978 school year) was 173,477, while the total number unaccounted for in reporting Title I analysis was 40,145. This means that approximately 76.8% of the data was present and that there was a loss of about one quarter of all these cases. This was significantly less than in the previous year and tended to occur, according to the report, primarily in kindergarten and first grade.

Some attempt was made to verify this information by individual school district. The annual evaluation report for each school district requires a listing of the number of participants in Title I evaluation programs broken down by subject area. Although this is listed as a "duplicated count" it might still serve to allow estimates of the percentage of missing data and at worst is likely to yield conservative estimates. The procedure followed here was to total up the number of participants listed and to compare this to the number of students listed as included in each analysis. Thus, a ratio can be formed with the number of students in the analysis as the numerator and the number of participants as the denominator. While one must be cautious about interpreting these results because of the possibility of faulty or missing information in the annual report, it is nevertheless instructive to compare these estimates of missing data with those reported by the state. This tabulation indicates that the median amount of missing data is 32% with a range from 1% to 91%. While this range

suggests the possibility of inaccuracy in this calculation, the median value of 32% does not differ considerably of the value of 25% as reported by the state. It is obvious from this information that a good deal of the data falls into the category of attrition. What is not clear is the degree to which additional exclusions are made on the data which are not reflected in these figures.

One reason for calculating within district estimates of missing data is so that it might be possible to compare the effects of various amounts of missing data on the estimates of treatment gain. Thus, when the estimates of missing data are dichotomized at the median, that is, all districts having less than 32% missing data are considered in one group while those having more are considered in the other, it is possible to see whether there are differences in gains under these conditions. In fact, the differences do not appear to be great. Those estimates of gain based on relatively little missing data yielded an average gain at the cutoff point of -2.02 while those districts with evidence of higher percentages of missing data had an average gain at the cutoff point of -2.61. While no difference appears to be present when the data is aggregated for the state, it is not clear how the exclusion of data might affect a given analysis. One study (Kaskowitz and Friendly, 1980) shows that while national aggregates of gains are not likely to be seriously biased due to attrition, individual project estimates can

sometimes be distorted by as much as ±15 NCE units. What is clear is that a significant percentage of the data does not reach the final analysis.

Common Exclusions. The list of subgroups which might be excluded from any given data analysis reads like a list of "special problems" which tend to arise in the process of conducting research. The most common exclusion, as recognized above, is due to what is traditionally considered attrition, that is, the lack of both a pre and posttest score for a given student. Several factors may be responsible for this. First, it may in fact be the case that students do not take both tests. This could be due to factors such as absenteeism or to the movement of students either into or out of the school district. It is often the case that these standardized achievement tests take several days to administer and as a result a student missing class on one particular day may not have a total test score. Transfer students may or may not have both tests depending upon whether the students were retested or whether test data was available from the district which they previously attended. Another source of attrition cases has to do with the inability to match up successfully pretest and posttest for given students. Thus, the problems mentioned above concerning the miscoding of such items as identification numbers, name and so on, will result in an inability to successfully match pre and posttest score.

A second exclusion which tends to occur is the result of

misassignment relative to the cutoff score. One respondent noted that it was discovered that schools within the district were not sending in cases which were misassigned relative to the cutoff score. When school administrators were asked why these scores were not sent their reasons usually had to do with the desire to comply with the requirements for strict adherence to a cutoff score. Sometimes this exclusion is made by the data analyst in order to insure that the requirement of a sharp cutoff has been adhered to.

Data is often excluded because of the mobility of the students. Generally, there are two types of mobility of concern here. First, students may transfer into and/or out of schools within the same district. Second, students may transfer into or out of the district itself. Within a given school district a distinction is often made between Title I and non-Title I schools. The Title I allocation procedures require at the district level that schools be ranked according to some measure which is an indicator of poverty. Most often schools within a district are ranked according to a measure like the number of free lunches which are served which is a readily available measure and is thought to be a reasonable proxy for the degree of poverty. Depending upon the size of allocation to the district it is often the case that some schools will receive no Title I funds at all. The schools which receive funds are usually determined by a cutoff point on the measure of poverty. Problems occur for

the data analysis when a student who is eligible to receive Title I service begins the school year in a Title I school and transfers to a non-Title I school or vice-versa. It is possible in this situation that eligible students will receive total service, partial service, or no service. Often the documentation is poor and it is difficult to know whether or not such cases should be included in the final analyses. Along with this is the difficulty in tracking any kind of transfers within the same school district. Illustrative of such difficulties is an example cited by Visco (1980):

> A child started out at School A transfered to School B and then transfered to School C all during the fall. All three schools are Title I and the child qualified for service in each. Thus, she was dropped at A added at B, dropped at B, and added at C, (so far, that is!).

In this case, even though the student might have received service throughout the year, she was excluded from the analysis. This may or may not be justified depending upon how one wishes to interpret the effect of the individual circumstances.

The second type of mobility which occurs involves transfers into and out of the school system. Depending upon the time of year the student transferred it may or may not be possible to obtain the necessary data. Thus, students transferring into the school district for the fall semester who did not receive the spring pretest may be able to be pretested and included in the analysis. Obviously in cases such as this one runs into the administrative difficulties

posed by such additional pretesting as well as the problems associated with equating scores from tests given at different points in time.

Another problem related to mobility concerns the situations where students simply "drop out of sight" for a period of time only to reappear later somewhere within the school district. This occurs with alarming frequency especially in school districts with large migrant populations (Winbush, 1980), such as districts in Florida in orange growing regions. Especially unnerving is the case of the prime harvest season being coincidental with the administration of district-wide tests. Given that it is not unreasonable for many school districts to have mobility rates as high as 25 or 30%, it is clear that the problems posed in trying to track such students can rapidly become formidable. Even if a certain percentage of this "mobile" population can be identified for both the pretest and the posttest, it is not clear whether their score should be included in the analysis.

Another exclusion which is commonly made is the repeaters or those students who are held back a grade from year to year. This exclusion offers a particularly clear illustration of the deterimental effects which exclusions in general can have on the data analysis. It is certainly reasonable to expect that most of the students who are required to repeat a grade come from the group of students

who are eligible for Title I services. If there are a fair number of repeaters and if these cases are routinely excluded from the data analysis it is likely that the treatment group regression line and subsequently the estimate of gain will be distorted. In addition, it is not unlikely that these students come from even the lower portion of the Title I treatment group distribution and that they may have more than the average share of disciplinary problems, learning disabilities, and so on. It is possible to conceive of a situation where excluding such students will actually make the Title I program look worse. A polar case is illustrated in Figure 12. The most likely repeaters are those at the lower end of the treatment group distribution who also do, on the average, worse than expected on the posttest for a given pretest value. These cases are indicated by the blackened portion of Figure 12. If such students are excluded the treatment group regression line would be lower in slope and an apparent "negative gain" would result.

Another set of exclusions is made for students who graduate from or drop out of the Title I program. Some respondents reported that when the Title I teacher perceives a student to have "caught up" the teacher will sometimes remove the student from the program without recording this information. Excessive absenteeism may also cause a child to be dropped or not depending upon whether or not this information was duly noted. Title I students are also

sometimes "picked up" by other programs such as English As A Second Language, Special Education, local compensatory education, and so on. Again, depending upon how accurately this information was recorded and the view of the individual data analyst such cases may or may not be excluded.

Cases are also excluded from the final analysis if it appears that they are "outliers." These exclusions are often made without any clear rationale to guide them. Thus, an analyst may in viewing graphs of the univariate and bivariate distributions decide that certain cases lie well outside the normal range for the distribution and on this basis exclude these cases. One respondent for example cited the case of a student who scored near the top of the pretest distribution and near the bottom of the posttest distribution. The question in such cases often revolves around whether the observed score is legitimate or whether it is the result of some processing error or miscoding. Sometimes it is possible to investigate such cases in-depth but often this is difficult or too costly. More rational guidelines for excluding outliers have been suggested and analytic procedures less sensitive to their presence have been recommended (Fleming and White, 1980).

It is legitimate within the Title I system to allow students who received services in one year to continue to receive services in subsequent years even if they no longer qualify in terms of their pretest score. This provision is

largely left to the discretion of individual district administrators. In cases where school districts decide to take advantage of this option, two problems tend to occur. First records of which students got serviced in previous years tend to be spotty and/or nonexistent and thus it is difficult to determine which cases fall into this category. Second, even for cases which do fall into this category whether they should be included in the analysis or not is not clear.

It is obvious from the above discussion that a large number of exclusions or disqualifications are considered or are incorporated prior to the statistical analysis. While any one exclusion may have only a minimal effect on the analysis the combination of several can pose formidable difficulties. Further work needs to be conducted to determine the extent to which such exclusions are made and the effect which they have on the analysis.

Data Processing Issues. Another class of problems related to exclusions or disqualifications concerns the order in which these are made and the variables which are used to select the excluded cases.

In the large amounts of data analyzed by computer it is often difficult to know which variables should be used to select subgroups for the analysis. For example, if one wishes to perform a task as simple as dividing students into those who were in Title I or non-Title I schools, it is not

clear whether it is better to use the school at the time of the pretest, the posttest, or a match between the two. Similarly students are often selected for the analysis of a particular grade level by selecting on the posttest grade. This seemingly simple rule would include in the final data setup all students who repeated the grade. This may or may not be desirable. Generally, one of two strategies is followed by a school district in the processing of its data. The first strategy involves creating successively smaller data sets by excluding certain subgroups from the analysis. The second strategy involves creation of a comprehensive data file including all cases and the application of a multiple If-statement for purposes of selecting the appropriate subgroups. When all exclusions are considered together, the process of selecting the appropriate subgroup is seen to be a formidable task. Visco (1980) provides an illustration of the complications which can occur in an example of the multiple If-statements which might be used to select the data for an analysis of a fourth grade reading program:

```
If Title I school on pretest
If Title I school on posttest
If school has reading service
If grade 4
If not repeater
If Title I service in reading
If below cutoff in reading
If not challenged in reading
If no makeup in reading
If stationary (no add-drop)
```

These are the statements which might be used to select the treatment group cases assuming that these cases have already

been successfully matched on the pretest and posttest and assuming that other exclusions (such as of Title I program repeaters) are not desired. Other sequences might be designed to achieve the same subgroup of interest and, in fact, the order in which the selections are made might significantly affect the cases which are finally included.

Summary of Design Implementation Analysis

The litany of design implementation problems listed in this chapter serves to illustrate the wide variety of these problems and their potential impact upon the data analysis. One would be naive to believe that these problems could be considered minor or that they are not likely to affect the data analysis seriously. On the basis of the investigation documented above it is possible to list the major goals of a complete design implementation analysis. These would include:

1) An identification of the major problem areas in implementing the design. Five general areas were identified in this Chapter, with each area containing subcategories of potential problems.

2) A determination of the prevalence or frequency of each problem for purposes of prioritizing the expected seriousness of their effects upon the data analysis. Although this task was addressed to some degree, one can conceive of carrying it out in a more formal manner. This could take the form of frequency counts of the number of times the problem occurs in various settings; or with percentage estimates, as with the estimates of the amount of missing

data; or by charting the distribution of the problem as when one looks at the percentage of grade repeaters for different pretest values. Because it is unreasonable to believe that all problems can be identified and accounted for, the purposes of this step would be to determine which are the major or most frequent problems which need to be addressed.

3) An assessment of the effect which the major implementation problems have on the estimates of treatment effect. There are several procedures which might be used to accomplish this. First one might simply look at pretest means for various subgroups of interest. This is especially appropriate if posttest data is missing and one wishes to determine whether for those cases there is any likely bias. Thus, if one finds that the pretest mean for students who do not have a posttest is not in the vicinity of the overall pretest mean one might suspect that this subgroup differs from the overall population of interest. In other cases one may be able to look at the bivariate relationship between an estimate of a particular problem and either the pretest or the posttest. For example, if one is interested in attrition, it would be useful to plot the percentage of attrition cases for various pretest scores. One would expect that if attrition has no serious effect upon the results, this percentage would be evenly distributed across all pretest values. It is especially useful in the case of the regression-discontinuity design to determine whether there is a coincidental jump in this distribution near the cutoff, as well as whether there are a proportionately greater number of cases in the treatment group range than in the control group. Probably the best way to determine the effects which various implementation problems have on treatment effect estimates is to conduct multiple analyses. For example, one could conduct the analysis with or without grade repeaters, transfers, and so on. If the results do not differ significantly, one would feel more confident in including such data in the final analysis. Similar strategies might be applied to determine the effects of measurement problems such as floor and ceiling effects. If, for example, a ceiling effect is thought to

exist, one might do a series of analyses where various amounts of data at the upper end of the distribution are systematically excluded. If a ceiling effect is present, the estimates of gain under such a series of analyses should change systematically. One must be careful, of course, to avoid misinterpreting such analyses. In this case for example, a true quadratic relationship, if interpreted and analyzed as a ceiling effect, could lead to erroneous conclusions.

The analysis presented in this Chapter illustrates the importance of design implementation issues. Such a study tends to degrade the distinction which is often made between process and outcome evaluations. While the conduct of a design implementation investigation is similar to the types of activities often involved in process evaluation, the goal is primarily to improve the quality of outcome studies. Nevertheless, much of the information which is gathered would be useful to administrators and researchers for understanding better how their programs work and improving procedures for future applications.

What specifically can be concluded about the implementation of the regression-discontinuity design within the context of Title I evaluation? The results of this investigation can be summarized briefly in terms of the five major topics considered in this chapter:

1. Applicability. Although Model C is used it is not used frequently. While there are many factors which affect use, the major one is the fact that on the average the model has yielded a negative estimate of program effect. This pattern seems to result primarily from curvilinearity in the data. While the curvilinearity may be reflective of the true

relationship it seems more likely due to chance level problems, floor and ceiling effects or data exclusions. Chapter III shows that when the typical Model C analysis is applied to curvilinear data one should expect biased estimates of gain.

2. Assignment. Misassignment occurs routinely in many districts as part of formalized "challenge" procedures. In typical Title I programs misassignment is likely to produce biased estimates of effect in spite of how these cases are handled in the analysis. Including challenges in their original group (disregarding their retest score), including them in the group indicated by their retest score, or excluding them from the analysis altogether will not in many cases remove the bias.

3. Measurement. Floor and ceiling effects and chance level problems can induce curvilinearity in the data and result in biased estimates of gain when using the Title I Model C analysis. The differences between chance level problems and floor effects have been underinvestigated and deserve further study.

4. Treatment. The assumption that all treated (or comparison) students experience standardized conditions is unrealistic. This is acknowledged in the Title I system by use of gross categories for differentiating program types (e.g., reading versus math, in-class versus pull-out, etc.). Nevertheless, more accurate descriptions of treatment and comparison group experiences (based perhaps on a detailed process analysis of the program) will improve the quality of the statistical analysis and the estimates of gain.

5. Exclusions. Significant amounts of data are routinely excluded prior to analysis. Cases are excluded because they are perceived as coming from "special" groups (e.g., Title I repeaters or graduates) or because of attrition (e.g., due to absenteeism, migrancy, transfers, and so on). Any combination of exclusions can act to distort estimates of program effect. Multiple

analyses may be useful for assessing the extent of these problems.

Overall, a major problem is poor documentation which makes it impossible to determine the extent and influence of implementation problems. Even where documentation exists there can be doubts as to its accuracy. Improvements in documentation are needed so that implementation problems can be ranked accurately by order of importance and procedures or analyses can be devised to deal with them.

III. THE STATISTICAL ANALYSIS OF THE REGRESSION-DISCONTINUITY DESIGN

This chapter is concerned with the statistical analysis of "sharp" regression-discontinuity designs. Several topics are considered. First, a polynomial regression model is offered as a useful general model for regression-discontinuity analysis. This model allows for any number of polynomial and interaction terms. Second, the selection of an appropriate subset of variables from this general model is discussed. The goal of procedures used to select such a subset is to exactly-specify the true model in the data. Problems of model misspecification arise because of the fallibility of available subset selection procedures. Under-specification of the true model will generally result in biased estimates of effect. Over-specification will yield unbiased estimates although these will be less efficient than estimates from an exactly-specified model. Third, because of the difficulties involved in selecting an appropriate subset of variables, it is argued that a conscientious analysis should be based on multiple estimates of effect using a variety of subset selection procedures. Fourth, over-specified models are shown as useful for obtaining unbiased (although less efficient) estimates of effect. Because of this it is recommended that such models should always be included as one part of a multiple analysis scheme. The use of over-specified models

is illustrated on simulated data as well as on data for twenty-four Title I regression-discontinuity designs obtained from three school districts. Fifth, the advantages of auxiliary analyses to explore problems associated with curvilinearity in the data (e.g., floor and ceiling effects, chance level effects, exclusions, etc.) are discussed and illustrated on real data.

A Statistical Model for Regression-Discontinuity

A good deal of work has been addressed to the issue of how one should analyze the regression-discontinuity design when the assignment to groups has been correctly implemented in terms of the cutoff criterion. Sween (1971), in a dissertation on the topic, advocated the calculations of separate regression lines for the treatment and comparison groups and the estimation of the treatment effect by means of a t-test of the difference between the regression lines at the cutoff score.

In several papers, Boruch (1973, 1974) and Boruch and DeGracie (1975) summarize a wide variety of potentially useful analytic models. They distinguish cases where the pretest variable is random or fixed, assignment is sharp or fuzzy, and the groups come from the same normal distribution or comprise separate ones. The recommended model for a commonly expected situation, that is, when assignment is sharp and the groups are thought to represent a single

population, is suggested in the work of Chow (1960) and consolidated in a more general model by Gujarti (1969). Boruch also discusses analyses which are appropriate under special circumstances or when certain assumptions are violated such as when heterogeneity of variance is present.

The Title I evaluation context has resulted in several papers related to the statistical analysis of Model C. Most notable among these are the papers by Echternacht (1978) and Echternacht and Swinton (1979) which discuss a number of approaches which might be used when there is evidence for a curvilinear relationship such as might be caused by floor and ceiling effects.

Econometricians and statisticians have also shown some interest in the regression-discontinuity and similar designs. Related discussions can be found in Spiegelman (1977, 1979), Rubin (1977) and Goldberger (1972). Some work on a general model appropriate for the nonlinear case or when the distributions are not normal has been carried out by Sacks and Ylvisaker (1976).

Description of the Model. The analytic model presented here is similar to many of the above procedures and has been described in Trochim (1979) and Reichardt (1979). The model can be stated formally as follows:

Given:

$$x_i \sim \text{NID}(\mu_x, \sigma_x^2)$$

$$y_i \sim \text{NID}(\mu_y, \sigma_y^2)$$

$$z_i = 1 \text{ iff } x_i \leq x_0$$

$$= 0 \text{ iff } x_i > x_0$$

Then:

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 z_i + \beta_3 x_i^* z_i + \ldots + \beta_{n-1} x_i^{*s} + \beta_n x_i^{*s} z_i + e_y$$

Where:

x_i^* = pre-program measure for individual i minus the value of the cutoff, x_0 (i.e., $x_i^* = x_i - x_0$)

y_i = post-program measure for individual i

z_i = assignment variable

β_0 = parameter for control group at cutoff (i.e., intercept)

β_1 = linear slope parameter

β_2 = parameter for treatment effect estimate

β_n = parameter for the s^{th} polynomial in x^*, or for interaction terms if paired with z

e_i = random error such that $e_i \sim NID(0,\sigma_e^2)$

The major hypothesis of interest is

$$H_0: \beta_2 = 0$$

tested against the alternative

$$H_1: \beta_2 \neq 0$$

Some explanation of the analysis is in order. First, it is only appropriate if the assignment strategy is correctly implemented, that is, there is no misassignment relative to the cutoff score. The analysis of "fuzzy" regression-discontinuity designs has been discussed elsewhere (Barnow, Cain and Goldberger, 1978; Campbell, Reichardt and Trochim, 1979; Goldberger, 1972; Spiegelman, 1976, 1977 and 1979; Trochim, 1980) and is outside the scope of this work.

Second, it is assumed that the true pretest-posttest relationship can be adequately described as a polynomial in x. If the true model is instead logarithmic, exponential or some other function, this model is misspecified and the estimates of treatment effect will be inaccurate. Even in these cases it may sometimes be possible to transform the data so that a polynomial model is appropriate. For example, if the true relationship is logarithmic, one can use the logarithm of y instead of y in the analysis specified above.

Third, this approach allows for changes (or interaction terms) in slope or function between the treatment and comparison groups. Thus, one fits the same linear, quadratic and cubic functions <u>across</u> both groups as well as separate ones <u>within</u> groups.

Fourth, the estimate of the treatment effect is made only at the cutoff point. This differs from the traditional Title I analysis where estimates are calculated at the treatment group pretest mean (Talmadge and Horst, 1976; Talmadge,

1976). The reason for this restriction is based on the inherent instability of extrapolations of regression lines. Estimates of effect at the cutoff point require no extrapolation whereas those at the treatment group mean involve extrapolation of the comparison group model into the region of the treatment group scores. Unless the true relationship is exactly-specified, the further one extrapolates from the cutoff the more likely it is that pseudo-effects will be generated.

Model Specification. From the general model described above one wishes to select that subset of variables (i.e., polynomial and interaction terms) which describes the true model in the data. Hocking (1976) discusses five computational procedures which are useful for subset selection: all possible regressions, stepwise methods (forwards and backwards), optimal subset procedures, sub-optimal methods and ridge regression techniques. For most of these procedures one usually defines selection criteria which are used to determine whether a particular variable will be included in the final model.

Among the criteria which are commonly used are the change in residual mean square, the squared multiple correlation coefficient and the total squared error. In addition, if a stepwise procedure is used, one usually needs a stopping criterion to help determine when an acceptable subset of variables has been selected. The reader is referred to

Hocking (1976) for an excellent discussion of subset selection.

The choice of computational procedure and criteria for subset selection should be guided by the goals of the analysis. The primary goal of regression-discontinuity analysis is to obtain an unbiased and statistically efficient estimate of program effect. This is achieved if the subset of variables which is selected from the general model exactly describes the true model in the data. In practice, exact specification is difficult to achieve. Hocking (1976) states: "The problem of interpreting the subset information is complex. The question of whether any of the proposed criteria are effective in identifying "true" variables is difficult to answer" (p. 44).

If the selected model is not exactly-specified then it must be either over or under-specified. A model is over-specified if it includes all variables in the true model along with additional terms which are not in the true model. It is under-specified if it does not include all variables in the true model. Sween (1971, 1977) and Hocking (1976) argue that parameter estimates will generally be biased when the true model has been under-specified. In addition, Sween (1971, 1977) presents evidence that an over-specified model will yield an unbiased estimate of program effect, although this estimate will be less efficient (i.e., have greater variance) than one obtained from an exactly-specified model.

Obviously, it is desirable to exactly-specify the true model. Given that this is difficult to achieve, one wishes at least to avoid under-specification and the potential for a biased estimate. Over-specification is preferable to under-specification in that it avoids this bias.

A reasonable approach to model specification would be to utilize several of the computational procedures and compare the estimates of program effect. For example, one might first attempt to exactly-specify the true model using a stepwise procedure and recognizing that the selected model might be under-specified. One might then deliberately over-specify the likely true model. Here, the estimate will probably be unbiased but will be less efficient than the previous one. A more restrictive model might also be tried. For example, the Title I analysis for Model C, which is under-specified if curvilinearity is present in the data, can be applied. One would then compare the estimates of effect generated under each model. If they are similar in direction and magnitude one can accept the exactly-specified model (and have greater confidence in the Title I estimate which is reported). If they are not similar one must weigh the potential for bias in the more restrictive models against the likely lower precision of the less restrictive ones. While a multiple analysis approach of this type seems conceptually unsatisfying it accurately reflects the inherent frailties involved in any type of model selection.

Computational procedures for attempting to exactly-specify the true model have received considerable attention (Hocking, 1976). However, except for the dissertation by Sween (1971) little has been said about the advantages of over-specification. Within a multiple analysis framework estimates from over-specified models can act as a benchmark against which one can judge the potential for bias in more restricted models. The remainder of this chapter illustrates the problems which arise when under or over-fitting the true model as well as the application of the overfitting approach to simulated and real data.

The Effect of Under and Over-fitting the True Model

The advocacy of an approach based on overspecification rests on the dual premises that underfitting the true relationship generally results in pseudo-effects while overfitting does not. To see this, consider first a true model which is linear throughout and exhibits no discontinuity or treatment effect. If higher-order terms are added to the model these new terms should be non-significant and the treatment effect should still be zero. Because there are more parameters in the model one has fewer degrees of freedom available for estimating the effect as well as multicolinearity between variables and the standard error of the effect will be larger. Nevertheless, no pseudo-effect should result from adding in the higher-order terms.

Consider next the null case for a true quadratic model as shown in Figure 13. While this function might be unlikely in practice, it provides a useful illustration of pseudo-effects. Here, one expects that linear models will yield pseudo-effects while quadratic ones will not. When cubic terms are added to the model these parameters should be non-significant and the estimate of effect should be similar to the one generated with the quadratic model.

The same logic can be extended to models of any order. If the true function is quintic (i.e., fifth-order), analyses which underfit this function (i.e., first through fourth-order models) are likely to yield pseudo-effects while models of fifth-order and higher should not.

One exception to the pseudo-effect assumption should be noted. An example is shown in Figure 14. Here, the true relationship is quadratic and the cutoff point is in the middle of this function (as measured on the pretest). This results in symmetrical functions on either side of the cutoff point. In this case, fitting a linear model will not lead to a pseudo-effect even though the true model is quadratic. The figure shows two linear models, one which allows for changes in slope (i.e., an interaction) and one which does not. The lack of a pseudo-effect in these cases is simply the result of symmetry in the function. Despite this anomaly, use of the analysis outlined here should still yield unbiased estimates of effect as higher-order terms are

added (although statistical power will be lower).

An Outline of the Analysis. The model set forth above would be estimated as one of the final steps in the analytic scheme. There are a variety of additional tasks which should be undertaken as part of a complete analysis. Several major components which should be included are presented here.

1. Summary Statistics. There are three uses for summary statistics. First, they give an indication of the quality of the data. Many of the problems cited in Chapter II can be assessed through computations of means or simple crosstabulations. For example, one might wish to compare the pretest means of matched cases and of those cases having only the pretest. This would give some indication of the potential for selection bias as a result of attrition. The crosstabulation of pretest grade and posttest grade will turn up miscodings such as children who "skip" from first to eighth grade or who "drop back" one or more grades from pretest to posttest. A simple crosstabulation of pretest and posttest school will give an estimate of within-district transferring. In the same vein, frequency distributions of the data will indicate likely "outliers" such as 80 year old first graders, or other miscodings. In addition, one can look at frequency distributions of any pretest characteristic (e.g., age, income, GPA, etc.) for pretest intervals to examine the possibility that the groups are discrete populations or to examine the possibility of coincident discontinuities at the cutoff. The latter is especially helpful in ruling out alternative explanations for a discontinuity on the posttest (i.e., alternatives to a treatment effect).

   A major concern with the regression-discontinuity design is the potential for floor and ceiling effects and for non-normal distributions in general. To check such possibilities it is useful to compare the pretest mean and median and to calculate distributional characteristics such as the skew or kurtosis.

   A second purpose for the summary statistics is to obtain data which are necessary for conducting

subsequent analyses. For example, it is often necessary to know the total sample size for the most efficient use of computer time as well as providing a value against which other analyses can be checked for accuracy. It is also helpful to obtain the minimum and maximum values and the range for the pretest and posttest. These are often needed for setting up graphs of the distributions. In general, it is important to build into the analysis multiple estimates of the same statistics. One might, for example, estimate the treatment group mean, standard deviation and pretest-posttest correlation as part of the summary statistics as well as in the process of computing the regression line. Multiple estimates of this type are inexpensive and increase confidence in the results by assuring that the correct subgroups (e.g., the treatment group) are being used in later analyses.

The third purpose of the summary statistics is purely descriptive. Means, standard deviations, correlations and the like should be reported in the write-up of the analysis.

2. Graphs. The importance of graphs in regression-discontinuity analysis cannot be overestimated. Graphs aid in the detection of problems with the data, in determining how well the design and analysis are implemented and in verifying the existence of a treatment effect. To begin with, it is useful to graph the univariate distributions of the pretest and posttest and to search for evidence of non-normality (e.g., skew), floor and ceiling effects, and outliers or coding errors.

There are two types of bivariate graphs which are helpful. First, one can look at the relationship between the pretest and a variety of measures which might act to degrade the analysis or in some way influence the estimates of effect. For example, one might graph an estimate of mobility within the school district against the pretest values in order to determine whether there is differential mobility between groups and whether this function shows a discontinuity in the vicinity of the cutoff score.

The most important graph is the pretest-posttest bivariate distribution. As above, this

can be used to detect problems with the data. Failing to inspect such graphs can lead to costly or time-consuming errors. One example of this is seen in several analyses of Model A data on which a Model C analysis was tested (Louisiana State Department of Education, 1980). Although the data was fuzzy relative to the cutoff, the intent was to discard the misassigned cases and try a sharp regression-discontinuity analysis. Unfortunately, the data as presented to the analyst included these cases. It would have been immediately apparent, had the data been graphed, that these cases had not been discarded. This was not detected until the analysis had been completed and, as a result, the data had to be reanalyzed. Without discussing the merit of analyzing data from one design with the procedures of another, it is obvious that a graph of the bivariate distribution would have saved considerable time and money.

A second use of bivariate graphs involves plotting posttest means for various pretest column intervals. A major problem in regression-discontinuity analysis concerns the order of polynomials which is needed to overfit the likely true model. Including more polynomials than necessary serves to reduce statistical power and increase the probability of committing a Type II error. A plot of posttest column means will usually present a smoother portrayal of the pre-post relationship than a simple bivariate graph.

A third use of bivariate pre-post graphs is as a verification or confirmation of the statistical analysis. One important feature of the regression-discontinuity design is that a treatment effect should be evidenced by a "discontinuity" in the function at the cutoff score. One would hope that this would be apparent in the graph. In fact, Riecken and Boruch (1974) suggest that one should even "distrust the statistical results if visual inspection makes plausible a continuous function with no discontinuity at the cutting point."

There are several considerations involved in the construction of appropriate graphs including the choice of scales for the variables, the selection of graphing symbols, and the determination of plot type (e.g., hand

drawn, line printer, digital plotter, and so on). Bad choices in these and other factors can lead to confusion and misinterpretation. One commercial evaluation company, Educational Consulting Services (1980), plots pretest values on the vertical axis and posttest ones on the horizontal. This is probably due to the fact that scores are sorted by the pretest values and it is easier to construct a line printer plot with the sorted score determining the line and the other value (posttest) determining the number of spaces to move left or right on that line. Whatever the case, such a format violates accepted graphing practices and further complicates interpretation. A major problem which occurs in Title I evaluation concerns the fact that the recommended analysis involves fitting separate regression lines in the treatment and control groups. Because of this it is usually more convenient, at least in terms of programming, to construct separate bivariate plots for each group. This is a practice which should be avoided because it makes it impossible to see the entire distribution at a glance and because, as a result of the computer plotting algorithms, the scales for the two plots are often different.

3. Regression Analysis. After computing summary statistics and inspecting the graphs, one can estimate the treatment effect using the approach presented earlier. This can be done with any standard multiple regression computer package (e.g., SPSS, BMDP, Datatext, SAS). If the package does not allow for stepwise fitting a separate regression analysis can be run for each of the steps. A simple regression-discontinuity analysis program written in SPSS is documented in Trochim (1979).

4. Additional Exploratory Analyses. It is important to go beyond the simple estimation of the treatment effect. Additional analyses are especially needed, if it is difficult to determine what order of polynomial is necessary in order to overfit the true model. For example if one suspects that the true function is not higher than a second-order polynomial, the data would be analyzed at least to a cubic power. If the estimates from the quadratic and

cubic steps differ greatly the true function and may be of higher order. Two strategies for exploring this problem are illustrated in the secondary analyses below. The first involves fitting higher-order polynomials than used in the original analysis and determining whether estimates at these steps begin to converge. The second involves weighting cases closer to the cutoff point more than those at the extremes of the distribution so as to reduce the effect which curvilinearity may have on the estimates. Two weighting strategies are illustrated in the secondary analyses. The first involves discarding all cases (i.e., truncating the distribution) which fall above or below certain pretest values. This is in fact an extreme form of weighting where all cases included are assigned a weight of one and those excluded a weight of zero. The second weighting procedure makes use of a continuous weighting variable such that those cases at the cutoff are fully weighted, those at the greatest distance from the cutoff receive zero weights and those in between are assigned weights proportional to their absolute distance from the cutoff.

Other analyses might also be used. Whatever the case, when one considers the enormous effort involved in testing and data processing and the complexity of the problems illustrated in Chapter II, it becomes apparent that the need to conduct a rigorous detailed analysis is warranted and should be encouraged.

Regression-Discontinuity Analysis in Title I. Before presenting results of the simulations and secondary analyses it is useful to summarize the analytic strategy recommended by the Office of Education and the RMC Research Corporation and compare it to the one offered here. The two major distinguishing features of the Title I approach are the calculations of separate regression lines for the treatment and comparison groups and the estimate of the treatment effect at both the treatment group pretest mean and the

cutoff point.

Although two separate regression lines are calculated, it is possible to get the same estimates through a single regression model. In fact, the procedure used for Title I evaluation is simply one submodel of the general model outlined above. For the estimate of gain at the cutoff their model can be expressed as:

$$y = \beta_0 + \beta_1 x^* + \beta_2 z + \beta_3 x^* z + e_y$$

where all parameters are as stated in the model outlined earlier. For the estimate of gain at the mean, the same model can be used with the exception that one subtracts the treatment group pretest mean, x_m from each pretest score, x_i which is used in the model. Thus the only difference between the two estimates lies in the construction of x_i^*, in one case by subtracting the cutoff value, x_0, and in the other by subtracting the treatment group pretest mean, x_m.

There are two major problems with the Title I analysis. First, if the true relationship is higher than first-order, this analysis is likely to yield pseudo-effects. In Figure 15 the underlying relationship is non-linear but is analyzed with separate linear functions in each group. The true curvilinear relationship shows no discontinuity, that is, there is no treatment effect. Nevertheless, the Title I analysis would show a discontinuity at both the cutoff and the treatment group mean. In this case a positive pseudo-effect would result. One could just as easily construct

scenarios which would yield negative pseudo-effects.

The second problem with the Title I analysis involves the practice of estimating the effect at the treatment group pretest mean (Talmadge and Horst, 1976; Talmadge, 1976). Figure 15 shows that the pseudo-effect for the estimate at the mean (designated by the letter "B") is considerably larger than the one at the cutoff (designated "A"). In fact, the figure demonstrates that the farther one extrapolates into the treatment group range, the greater this pseudo-effect becomes. It is primarily for this reason that the estimate at the mean should not be used.

To summarize, the Title I procedures comprise one subset of the general model offered here. If the true relationship is not linear it is likely that pseudo-effects will result. In addition, these pseudo-effects will often be greater for the estimate at the mean than for the one at the cutoff.

Illustrative Simulations of Under and Overfitting

To illustrate the effect of under and overfitting the true model computer simulations were conducted. Two independent assumptions are being verified in these simulations.

1. Does underfitting the true model result in pseudo-effects?
2. Does overfitting the true model yield unbiased estimates of effect?

Regarding the latter, it is also important to examine

how the statistical precision of the estimate changes with overfitting. The general strategy involves creating pretest-posttest distributions with a known relationship and then overfitting this relationship to estimate effects.

A total of sixty computer runs were carried out to examine these issues. Twenty bivariate distributions were constructed for linear, quadratic, and cubic true models. Within each of these three sets of twenty runs, ten were constructed for the null case and ten having a gain of five units. Each file had 1000 cases with between one-fourth and one-third assigned to the treatment group. The data were constructed in the following manner:

1. A pseudo-random number generator was used to construct a common true score, t, and independent random error for the pretest and posttest such that:

$$t \sim N(0, \sigma_t^2)$$

$$e_x \sim N(0, \sigma_{e_x}^2)$$

$$e_y \sim N(0, \sigma_{e_y})$$

2. A pretest, x, was constructed using the true score and error component such that:

$$x = t + e_x$$

3. An assignment variable, z, was constructed by selecting a cutoff point, x_0, such that:

$$z = 1 \text{ iff } x \leq x_0$$

$$= 0 \text{ iff } x > x_0$$

4. A posttest, y, was constructed using the true score, error component and gain, g, such that:

For linear model:

$$y = t + gz + e_y$$

For quadratic model:

$$y = t^2 + gz + e_y$$

For cubic model:

$$y = t^3 + gz + e_y$$

The size of the gain was set at either 0 or 5 units. Since z is a dichotomous variable (i.e., 0,1) this constant gain was added to the posttest only for the treatment group cases (i.e., when z = 1).

It is desirable in these simulations to avoid the exception to the pseudo-effect assumption which results from symmetrical distributions in both groups. Because the pretest has an expected value of zero, use of a cutoff point equal to zero would have resulted in symmetrical distributions on either side of the cutoff and, consequently, pseudo-effects would not have been detected for underfitted models. In addition, it is shown in Chapter II that in practice the cutoff point is seldom as high as the 50th percentile. Consequently, these simulations are based on cutoff points less than zero on the pretest.

Another problem occurs in setting the values for σ_t, σ_{e_x}, and σ_{e_y}. If these are set to be the same for the linear, quadratic and cubic functions the variance of the

posttest will be considerably larger for the cubic than for the quadratic and for the quadratic than for the linear. This is because for the cubic function y depends on t^3 whereas for the linear it depends on t alone. Because of this it is desirable to alter the ratio of σ_t/σ_e for the three functions. In all runs σ_{e_x} and σ_{e_y} are equal to one unit but σ_t differs for the three functions. Because σ_t differs the variance of the pretest will also differ. Therefore, use of the same cutoff value would have resulted in considerably different average treatment group sizes for the three functions. As a result, the cutoff point is set differently for each function so as to obtain approximately the same size treatment group in all runs.

The parameter values used in the simulations are shown in Table 3 for the standard deviation of the true score, σ_t, the standard deviations of the error terms, σ_{e_x} and σ_{e_y}, the cutoff value, x_0, and one of two sizes of gain, g = 0 or g = 5. For each of the sixty runs entirely new data files were constructed, that is, each run is independent of all others.

Although each analysis could be conducted by fitting a model to a polynomial only one order higher than the true function, it is instructive to deliberately overfit by several orders of polynomials to see what happens to the estimates of gain and their standard errors. The analysis of all runs uses up to seventh-order polynomials and their interaction terms fitted in a series of eight steps. The

model used at each step is shown in Table 4. The first step fits a linear term, the second adds in a parameter allowing for a change in slope (i.e., an interaction term), the third adds both the quadratic and its interaction, the fourth adds the cubic and its interaction, and so on to the eight step which adds the seventh-order term and its interaction. It is important to note that Step 2 of the analysis is identical to the model used in Title I evaluation to estimate the gain at the cutoff.

For the true linear model the estimates of gain and standard error are shown for the twenty runs in Table 5. These data are summarized in Table 6 which shows the mean gain (n = 10 for gain = 0 and for gain = 5) and the standard error of the mean gain for each step in the analysis. For the true quadratic model the estimates of gain and standard error are shown for the twenty runs in Table 7. The data are summarized in Table 8. For the true cubic model the results of the twenty runs are shown in Table 9 and summarized in Table 10.

The first question of importance is whether underestimation of the true model results in pseudo-effects. For the linear model the true model is not underestimated and no pseudo-effects are expected. For the true quadratic model one underestimates the function in Step 1 (linear) and 2 (linear interaction added) of the model. For the null case, the average gain for Step 1 is 6.896 with a standard

error of .338. For Step 2 the gain is 2.137 with a standard error of .271. When the gain is 5 units pseudo-effects also appear to be generated for Step 1 ($\hat{\beta}_2$ = 11.723, S.E.($\hat{\beta}_2$) = .147) and Step 2 ($\hat{\beta}_2$ = 6.758, S.E. ($\hat{\beta}_2$) = .158). Clearly, the notion that pseudo-effects can result from underestimation of the true model is supported here for the quadratic model. When the true model is cubic one expects that estimates from Steps 1, 2, and 3 (quadratic added) will be biased. Table 10 shows for the null case that the average gain differs from zero by more than two standard error units for all three steps. Similarly, when the gain is equal to 5 units the average gain for the first three steps differs from 5 by more than two standard error units. Therefore, the first assumption, that underestimation of the true model may lead to pseudo-effects, is supported by these simulations.

The second major question is whether overestimation of the true model results in unbiased estimates of effect. For the linear case we are interested in examining Steps 2 through 8, for the quadratic, Steps 4 through 8, and for the cubic, Steps 5 through 8. The three summary tables show that only two estimates of average gain differ by more than two standard error units from the true gain. Coincidentally, these both occur at Step 6 in the analysis (quintic and quintic interaction term) for the quadratic case when gain = 5 and for the null cubic case. Nevertheless, the notion that overfitting the true model yields unbiased treatment effect

estimates is supported here.

The results of these simulations should be viewed cautiously for several reasons. First, they are based on only ten runs for each combination of conditions. More definitive simulations would require a larger sample size. Second, the pattern of results may differ for simulations using different parameter values.

As mentioned earlier, selection of a cutoff value which yields symmetrical distributions in the two groups may not result in pseudo-effects for underestimated models. It is not clear how different values for the variances, gain or cutoff would affect the results. Finally, these simulations are based on polynomial true models. These may or may not be reasonable approximations of the distributions found in practice. If they are not, pseudo-effects may result for any order of polynomial which is used. In spite of these qualifications, the simulations illustrate that under certain conditions the assumptions involved in overfitting are reasonable.

The results of these simulations have an important implication for the Title I analysis used to evaluate regression-discontinuity data -- if the true pretest-posttest distribution is non-linear the Title I analysis is likely to generate pseudo-effects. This problem of nonlinearity has generally been recognized in the Title I evaluation literature (Echternacht, 1978). The additional contention

here is that as long as the distribution can be reasonably approximated by a polynomial function, the practice of overfitting the likely true relationship will result in unbiased estimates of effect.

Secondary Analyses of Title I Data

Data for eight Title I programs in the 1978-1979 school year were obtained from the school districts of Providence, Rhode Island, Marion County, Florida and Osceola County, Florida. Although one could use a number of criteria to determine the degree of overfitting, these analyses are based on the assumption that a true relationship of higher than third-order is unlikely for these data. As a result, the analyses are comprised of only the first four stps (linear through cubic) used to analyze the simulation data above as shown in Table 11.

Secondary Analyses of Providence Data. Providence is the largest of the three school systems in this study and the most urban. The test used for both pre and posttest is the Comprehensive Test of Basic Skills (CTBS). It was administered in March, 1978 and March, 1979. Title I reading programs were implemented in grades two through six while mathematics programs were carried out in grades two through four. The grades, test level and forms are shown in Table 12.

Nearly seventy separate data files were acquired from

Providence. These include the files of raw scores for the pretest and posttest separately, a large number of "intermediate" files (e.g., unmatched and separate reading and math files for each grade), and the final matched files comprising a separate file for each subject (e.g., reading or math), group designation (e.g., treatment, comparison, or non-Title I) and grade level combination. The data used in these reanalyses were constructed by combining for each subject and grade level combination a merged file of all three group designations. In this manner, eight files were obtained, five for the reading programs in grades two through six and three for the math programs in grades two through four. The variables in these files are listed in Table 13. Each file contains the appropriate pretest and posttest score and a group designation. In addition, different subscales are available depending on the level of test which was administered.

Descriptive Statistics. For each treatment group and grade level the sample size, mean, variance, skewness and kurtosis is shown for the reading pretest and posttest in Table 14. The same information is given in Table 15 for the mathematics program. The transfer of data was successful as indicated by the exact correspondence between the estimates calculated here and those published in the Providence Title I Evaluation Final Report (The Providence School Department, 1979). As expected, given the typical cutoff percentiles

which were detected in the meta-evaluation of the Florida data in Chapter II, the treatment group sample sizes are considerably smaller than the comparison group ones. In general, the pretest variances for the treatment and comparison groups are considerably smaller than posttest variances (or, for that matter, the pretest variances of the non-Title I group) confirming that the distribution was truncated at the cutoff point.

Results. The bivariate distributions for the five reading programs and three math programs are shown in Figures 15a to 15h. The estimate of gain, $\hat{\beta}_2$ and standard error, SE ($\hat{\beta}_2$), for each step in the analysis is shown in Table 16 for all eight programs. There are several ways to view these estimates. First, one can look at the magnitude of the estimate generated in the fourth step of the analysis. If the true function is third-order or less this should be an unbiased estimate of effect. Viewed in this way, the highest positive gains were for the reading programs in grade 2 ($\hat{\beta}_2$ = 21.69) and grade 5 ($\hat{\beta}_2$ = 28.47). Second, one can look at the statistical significance of estimates generated at step 4. At the .05 level only the estimate for the math program in grade 4 ($\hat{\beta}_2$ = 28.47, SE ($\hat{\beta}_2$ = 8.60) is statistically significant. Third, for any given analysis one can look at the stability of estimates across steps. If these fluctuate considerably, as with the reading programs in grades 3 and 4, one might suspect that the true model is

overfitted by one or more orders of polynomial. For example, for the second grade reading program the estimates are consistently positive and, up to step 3, are statistically significant at the .05 level. Given that statistical precision declines the more one overfits the true model it might be reasonable to conclude in this case that the estimate at step 4 is not significant primarily because of lower statistical power due to the higher-order terms. Finally, it is important to compare these results with the estimates obtained by the Title I analytic strategy (which is estimated in Step 2). The Title I analysis generates statistically significant effects for the reading programs in grades 2 ($\hat{\beta}_2$ = 35.84, SE($\hat{\beta}_2$) = 10.07), grade 3 ($\hat{\beta}_2$ = 16.96, SE($\hat{\beta}_2$) = 6.94), grade 4 ($\hat{\beta}_2$ = 15.78, SE($\hat{\beta}_2$) = 7.53) and grade 6 ($\hat{\beta}_2$ = -22.13, SE($\hat{\beta}_2$) = 9.94) and the math program in grade 4 ($\hat{\beta}_2$ = -19.03, SE($\hat{\beta}_2$) = 5.31). The overfitted models find a significant effect only for the math program in grade 4, although it may also be appropriate to consider the grade 2 reading program as having a marginal effect.

Secondary Analyses of Marion County Data. Marion County is a moderate size county of approximately 100,000 people situated in central Florida. The CTBS was used as the pre and posttest in all grades except the first grade where the Metropolitan Reading Readiness Test (MRRT) was administered for the pretest. All students were pretested in March, 1978 and posttested in March 1979 with the exception of the first

grade students who were pretested in September of 1978.

The data were acquired in the form of one large file which included all scores. Contained in this file is a variable which designates treatment condition (i.e., comparison, reading only, math only, both reading and math), a school type (i.e., Title I or non-Title I), and a merge code (i.e., both tests, pretest only, posttest only). The variables in the file are listed in Table 17.

A distinguishing feature of the Marion County data concerns the use of functional-level testing. Within any grade level a variety of test levels are used depending on an assessment of the level of difficulty most appropriate for a particular student. Because of this it is not possible to list the levels and forms of the test which are used at any grade level. In addition, test scores are reported on the CTBS Extended Standard Score Scale which is a single scale spanning all levels and forms. A total of five reading programs (grades one through five) and three mathematics programs (grades three through five) were implemented.

Descriptive Statistics. For the treatment and comparison students at each grade level the sample size, mean, variance, skewness and kurtosis is shown for the reading pretest and posttest in Table 18. The same information is given in Table 19 for the mathematics programs. The transfer of data was successful as indicated by the correspondence between the estimates generated here and those listed on computer

printouts acquired from the district. Except for the first grade reading program, the treatment group is always larger than the comparison group. The mean scores increase with grade level demonstrating the continuous extended score scale. The low pretest scores in the first grade are due to the use of the MRRT (which uses a different scale) as the pretest.

Results. The bivariate distributions for the five reading programs and three math programs are shown in Figures 16a to 16h. The estimate of gain, $\hat{\beta}_2$, and standard error, $SE(\hat{\beta}_2)$, is shown for each step in the analysis in Table 20. The highest positive gain at step 4 is for the reading program in grade 2 ($\hat{\beta}_2 = 26.10$, $SE(\hat{\beta}_2) = 12.50$). These are also the only Marion County programs for which the estimates at step 4 are statistically significant at the .05 levels. The estimates are fairly stable across steps for the grade 5 reading program and the math programs in grades 3 and 4. Closer inspection of the figure and estimates for the third grade math program indicates that there is some evidence for a positive effect although the estimates of the third and fourth steps are not statistically significant. The Title I analysis (step 2) would have detected significant effects for the reading programs in grades 3 ($\hat{\beta}_2 = -11.89$, $SE(\hat{\beta}_2) = 4.40$) and grade 5 ($\hat{\beta}_2 = -22.16$, $SE(\hat{\beta}_2) = 6.31$) and for the third grade math program ($\hat{\beta}_2 = 8.07$, $SE(\hat{\beta}_2) = 3.90$).

Using the overfitted model one could conclude that the grade two reading program had a positive effect, the grade five reading program had a negative effect, and possibly, that the third grade math program can be considered marginally positive.

Secondary Analyses of Osceola County Data. Osceola County is a small rural county of approximately 40,000 people situated in central Florida. It is the smallest of the three districts in this study and probably has the largest migrant population. The Stanford Achievement Test (SAT) was used for the pretest (April, 1978) and the posttest (March, 1979). Title I programs in reading and math were implemented in grades two through five.

Osceola County is the only district for which the unmatched test scores were obtained and analyzed. With the Providence, Rhode Island and Marion County data the files which were acquired already contained matched data and some exclusions had been made. For Osceola County, the data consists of national percentile scores from the SAT. The form and level of the test used for each grade are shown in Table 21. The variables which are available for analysis are shown in Table 22 and, as with the other tests, depend to some extent upon the level and form of the test.

Because separate pretest and posttest files were obtained it was necessary to match these scores. Many of the problems associated with matching these data were described

in Chapter II and comprise a serious threat to the quality of the analysis. There are several additional problems which also act to degrade the analysis. First, "misassignment" relative to the cutoff point occurred. There were, in fact, two separate cutoff points. All students scoring below the fifteenth percentile on the test were eligible for service but, in addition, all those scoring below the thirtieth percentile who received service in the previous year were also included. If the latter group is included, the analysis is a "fuzzy" regression-discontinuity one. In this analysis all students in this group were excluded. Unfortunately, it it is difficult and costly to determine which students received service in the previous year and not in the current one. Ideally, if any exclusions of this type are made they should probably include all students receiving service in the previous year (thus limiting the population of interest to those students who received service for the first time in the current year).

Another problem with the data is that it is reported in terms of national percentiles. This yields square univariate distributions and violates the normality assumption of the analyses. Because of this the data were transformed from percentiles to Z-scores using a table of standard scores. The transformation has the effect of increasing the density of scores near the middle of the distribution. To illustrate why this is necessary consider the bivariate plot of the

second grade reading scores in percentiles in Figure 17 versus their standard scores as in Figure 18a. It is easy to see that the former is a more "squared" distribution and does not resemble a bivariate normal distribution as closely as the latter does.

It is difficult to assess the accuracy of the data transfer because the district analysis was done by hand (e.g., the data was not computerized). As a result it is not easy to determine what exclusions were made (and at what point in the analysis). To get some idea of the comparability of the analyses one can compare the reading program sample sizes as reported by the county to the state versus the sample sizes obtained here. These are shown in Table 23. In all cases but one the sample sizes for the secondary analysis were smaller. This is probably due to the conservative matching procedure which was used. For example, while the district analyst may have known that John Jones and John P. Jones are really the same child, this could not be assumed here and these cases were excluded for lack of a match. When one considers the complexity of the data preparation procedures it is remarkable that the sample sizes are as similar as they are.

Descriptive Statistics. For each treatment group and grade level the sample size, mean, variance, skewness and kurtosis is shown for the reading pretest and posttest in Table 24. The same information is given in Table 25 for the

mathematics programs. Note that all analyses are performed on data which is transformed to standard scores. Thus, for any distribution of scores we expect the mean to be zero and the variance to be one. A major problem with the data is evidenced in the treatment group sample sizes. No treatment group had more than 39 students.

Results. The bivariate distributions for the four reading and four math programs are shown in Figures 18a to 18h. The estimate of gain, $\hat{\beta}_2$, and the standard error, $SE(\hat{\beta}_2)$, is shown for each step in the analysis in Table 26. The largest positive gain at Step 4 is for the math program in grade 5 ($\hat{\beta}_2 = .95$, $SE(\hat{\beta}_2) = .54$) while the largest negative gain is for the reading program in grade 4 ($\hat{\beta}_2 = -1.16$, $SE(\hat{\beta}_2) = .36$). The latter program is the only one having an estimate which is significant at the .05 level for step 4. Grades 2 through 4 in math and grade 2 in reading were programs which yielded similar estimates across all four steps. The Title I analysis (step 2) generated statistically significant estimates for the grade two reading program ($\hat{\beta}_2 = .52$, $SE(\hat{\beta}_2) = .18$) and the grade three math program ($\hat{\beta}_2 = .75$, $SE(\hat{\beta}_2) = .27$). Using the overfitted model, one would conclude only that the grade 4 reading program had a negative effect on achievement test scores.

Additional Analyses. A number of questions can be raised about any analytic procedure and supplemental analyses can be useful in answering them. To illustrate this, several

additional analyses are presented which are primarily related to two questions:

1. Is the highest order of polynomial used in the analysis (in this case, cubic) sufficient to overfit the likely true function?

2. To what extent might problems associated with curvilinearity in the data (e.g., floor and ceiling effects, chance level effects, outliers, etc.) be distorting the estimate of effect?

The data from the Providence, Rhode Island fourth grade math program are used to investigate these questions. The bivariate distribution for the data was presented earlier in Figure 15h. This data set was chosen because there is clear evidence of curvilinearity and, as a result, the data may not have been overfitted by a cubic model (question 1) and because there appear to be some "outlier" cases, especially in the upper ranges of the pretest distribution (question 2).

A simple way to examine whether the true model had been overfitted is to extend the original analysis by adding even higher-order polynomials. If the estimates of gain found for these additional steps are similar to the estimates in the original analysis, one can fairly conclude that a sufficient order polynomial has been used. For the grade 4 math program additional steps were added up to a tenth-order polynomial and its interaction term. The same models as used for the simulations described earlier were used here for steps 1 through 8. Steps 9, 10 and 11 added in eighth through tenth-order polynomials and their interactions. The estimate of gain, $\hat{\beta}_2$, and standard error, $SE(\hat{\beta}_2)$, is shown for

each of the eleven steps in Table 27. For all steps added to the original analysis (Steps 5 through 11) the estimate of gain is negative and statistically significant at the .05 levels. The estimates become slightly more negative with increased overfitting but appear to level off near a value of -41.0 between Steps 7-11. One can conclude that it may be appropriate to overfit a model to at least a sixth-order polynomial.

The second question concerns the effects of curvilinearity in the data. There are several possible outlier values in the grade 4 math data but only those in the upper ranges of the comparison group will be considered here. There are two distinguishable patterns in this area of the distribution. First, between the pretest scores of 325 and 450 the distribution appears to "tail upwards" relative to lower pretest value cases. Second, there appear to be two points having a pretest score greater than 450 which tend to stand apart from the rest of the scores. The question of interest here is how the estimates of effect might differ if the assumption is made that these scores do not reflect the true relationship. This can be investigated by using a variety of schemes which weight cases nearest the cutoff point more heavily than ones at the extremes of the distribution. If the estimates with weighting differ from those generated in the original analysis one might suspect that these cases are having an effect.

The first strategy used involves discarding cases (i.e., truncating the pretest distribution) if their pretest score falls above certain values and reestimating the effect using the original four-step model. For the grade 4 math data, the distribution was truncated at eight different pretest values at twenty-five point scale score intervals. Thus, all cases having a pretest score greater than 475 are excluded in the first analysis, those greater than 450 in the second, those greater than 425 in the third, and so on to the eighth analysis which excludes all cases with a pretest value exceeding 300. Discarding cases in this manner can be viewed as a form of weighting where all cases below a given pretest value receive full weight (i.e., w = 1) while those above receive no weight in the analysis (i.e., w = 0).

The estimate of effect, $\hat{\beta}_2$, and standard error, $SE(\hat{\beta}_2)$ is shown for each step in the original analysis and the eight truncated analyses in Table 28. At the fourth step in the analysis the estimate of gain across all truncations is similar to the one in the original analysis. In addition these estimates are significant until the data is truncated at pretest values of 350 or lower. As greater amounts of data are excluded, estimates of gain at lower steps in the analysis tend to approximate more closely the estimates in the fourth step. The conclusion which can be reached is that even if the cases in the upper ranges of the comparison group are "outliers" or are affected by floor and ceiling problems,

the exclusion of these cases does not significantly alter the estimate of program effect.

One can also weight the data continuously (rather than dichotomously) in terms of distance from the cutoff point. This was done for the grade 4 math data by assigning cases at the cutoff full weight (i.e., w = 1), those one-tenth of the way from the cutoff to the most extreme value nine-tenths weight (i.e., w = .9), those eight-tenths of the distance to the extreme value a weight of two-tenths (i.e., w = .2) and so on to the case farthest from the cutoff which receives no weight (i.e., w = 0). To determine the point most distant from the cutoff one can use the point most distant in both groups or separate ones in each. The former is used in these analyses. The same eight-step model used in the simulations is used here after cases have been appropriately weighted.

The estimate of gain, $\hat{\beta}_2$, and standard error $SE(\hat{\beta}_2)$, is shown for all steps in Table 29. These estimates correspond closely to the ones generated above from the original and truncated analyses. One can assign even greater weight to cases nearest the cutoff point by using higher-order polynomials of the weighting variable described above. For example, if the square of the weighting variable is used (i.e., w^2), a case which had a weight of 1 would still have that weight, one which had a weight of .5 would now be weighted .25, and so on. The estimate of gain, $\hat{\beta}_2$ and standard error $SE(\hat{\beta}_2)$, is shown for each step in Table 29

using the square of the weighting variable. Again, the pattern of gains agrees with those found above.

On the basis of these analyses one can be more confident in concluding that the grade 4 math program had a negative effect on achievement test scores. This effect cannot be considered a pseudo-effect due to underfitting the true model because when polynomials as high as tenth-order are added the estimates are similarly negative. In addition, the effect cannot reasonably be considered an artifact due to problems with the data in the comparison group because of the results which were generated with a variety of weighting schemes. Other explanations for the effect might also be explored. For example, the potential for chance level effects in the treatment group scores could be investigated by discarding verying numbers of cases from the lower end of the pretest distribution. This was not done here because the treatment group has so few cases to begin with. Whatever the case, it is useful to conduct additional analyses to rule out alternative explanations for the estimate of effect and increase one's confidence in the conclusions which are reached.

Conclusions

This chapter illustrates that misspecification of the statistical model in regression-discontinuity analysis can lead to pseudo-effects. If the true pretest-posttest

relationship (in the absence of a treatment effect) is known, one simply uses this model to obtain unbiased estimates of effect. The major problem is that one is seldom certain about the nature of the true model. Even if certain, other factors related to design implementation (e.g., floor and ceiling effects, chance levels, etc.) can act to degrade or distort this relationship.

With overfitting one will obtain unbiased estimates of effect. These estimates become less efficient as more higher-order terms are added. Because of model specification problems multiple analyses should be conducted which include attempts to exactly-specify the true model and deliberately overfit it. Thus, if the goal in regression-discontinuity design is to obtain a program effect estimate which is both unbiased and efficient, one can use overfitting to help determine the former and exact-specification to help achieve the latter.

It is important to keep in mind that the general model presented here is based on the assumption that the data are reasonably expressed in terms of a polynomial model. If the true model is not a polynomial in x, including any number of polynomial terms is likely to still result in biased estimates. In general, it may be that the more one is able to approximate the true model with a polynomial model the smaller the bias which would result under this analysis. It is important to note that the model presented here shares

this difficulty with most others. The Title I procedure, for example, is even more restrictive in that it assumes that the true model is linear.

One interesting implication of overfitted models is especially appropriate for Title I evaluation. A purpose of the Title I evaluation system is to aggregate estimates of gains across school districts, grades, type of program, and so on. With the current Title I analysis it is likely that some, if not many, of these gains are biased due to underfitting the true model. If all school districts using the regression-discontinuity design routinely overfit the likely true model (for example, as high as a fifth-order polynomial), these estimates, while probably inefficient for a particular program, will be less biased in the aggregate than with the current Title I analysis. Thus, regardless of the estimate of gain which a given school district wishes to use, from a statistical point of view there is some advantage in using estimates from an overfitted model for purposes of national aggregation.

More research needs to be done to test the limits of the analytic approach outlined here. Especially useful would be an examination of the pattern of gains yielded when the true model is only approximated by a polynomial and an investigation of the reasonableness of the assumption that polynomial models are appropriate. In addition, problems in interpreting estimates of effect generated by competing analyses deserve

greater exploration. While not entirely satisfying, the multiple analysis approach described here does, however, provide a reasonable method for obtaining unbiased and efficient estimates of treatment effect with the regression-discontinuity design.

IV. DESIGN VARIATIONS

The intent of this Chapter is to expand the traditional definition of the regression-discontinuity design by suggesting alternative variations and applications. It is important to keep in mind that the major distinguishing feature of the design is the assignment to conditions on the basis of a cutoff on some quantified measurement. Any analysis which is based at least in part on such a strategy will be considered here a variation of the design.

There are two major topics considered. The first includes a discussion of design variations and analyses with special emphasis on extending the versatility of the design for Title I evaluation. The second involves a more general look at the appropriateness of the design for evaluating other Federal programs. The latter topic offers some interesting possibilities, especially for coupling the design with the formulas which are used for allocating funds on the basis of need, merit, and other similar constructs.

Major Variations

The regression-discontinuity design is far more versatile in principle than its present applications might suggest. This section is devoted to consideration of some of the major

useful variants. Although the discussion is aimed primarily at applications within the Title I evaluation system, the design may be appropriate for the evaluation of other federal social programs. Five major areas are considered. These are, different strategies for handling assignment to condition, treatment-related variations, alternative post-program measures, the use of the design on aggregate units and the use of the design on a post hoc basis.

Assignment Variations. The major distinctive characteristic of the regression-discontinuity design is the assignment to condition on the basis of a sharp cutoff score on a pre-program measure. There are many aspects of this assignment strategy which can be altered to improve the estimate of program effect which are generated.

In theory, the most methodologically sound way to assign units or persons to conditions is through random assignment. In fact, the regression-discontinuity design is recommended only when random assignment to conditions is not feasible. A primary advantage of random assignment (assuming that the treatment groups have a sufficient sample size) is that the groups which are assigned will on the average have the same expected value on any characteristic measured prior to the administration of the program. Thus, groups can be considered to be "equivalent" prior to the program and differences which occur can be attributed to either the program or some other event which occured subsequent to

assignment. For a variety of reasons random assignment to condition is not used as frequently as would be desired (Boruch, 1978).

Little work has been conducted to directly compare the regression-discontinuity design with the true experiment (i.e., a design based upon random assignment to condition). Goldberger (1972) suggests that all things being equal, one needs at least 2¾ times as many program participants for the regression-discontinuity design as for a randomized design in order to attain the same degree of precision in estimating program effect. One of the strongest variations of the regression-discontinuity design therefore would be one which incorporates random assignment to condition, at least in part. Much of the following discussion on coupling the regression-discontinuity design with random assignment follows from the suggestions outlined by Boruch (1973).

Many of the difficulties which occur in the implementation of the regression-discontinuity design stem from the requirement of adhering to a strict cutoff. While lack of adherence to this requirement may be motivated by a factor like political favoritism, it is often the case that it is due to the reasonable belief that persons closest to the cutoff score may be misassigned simply as a result of error in the pre-program measure. If it is assumed that error occurs in the pre-program measure it might be more reasonable to define a cutoff interval rather than a single

cutoff point and to randomly assign persons within this interval to treatment or comparison group. Within this cutoff interval it would be assumed that all persons have approximately the same true score but differ on the measure primarily due to measurement errors. Conceptually it might even be useful to set the width of the interval on the basis of some estimate of measurement error.

One way to view this modified design is as a true experiment "imbedded" within a regression-discontinuity design. Several analytic strategies would be possible. One could analyze just the data within the cutoff interval as a true experiment of its own. Alternatively one could analyze the data on either side of the cutoff level by means of a standard regression-discontinuity analysis where the cutoff is the midpoint of the interval. Such a strategy would only be recommended for cases where the cutoff interval is relatively narrow and regression lines could be fairly estimated for the remaining data. Finally, it would be possible to include all the data in a standard regression-discontinuity analysis as outlined in Chapter III, even though assignment is not sharp relative to the cutoff. This is acceptable as long as random assignment within the cutoff interval has been adhered to because such misassignment around a cutoff will not result in biased program effect estimates.

This design would be especially useful in Title I

evaluation because of the possibility of nonlinearity in the data as a result of factors described in Chapter II. Assuming that there is sufficient sample size for groups within the cutoff interval, one would be most confident in the estimates generated by the true experiment. The analysis of the data for the whole range on the pre-program measure could be conducted to increase the precision of the estimates and as an aid in determining whether there are problems which occur in the tails of the distribution. This design has been discussed as an extension of the "tie-breaking experiment" by Campbell (1969), Riecken et al. (1974), and Boruch (1973).

Random assignment could be incorporated at other points along the pre-program continuum than in the vicinity of the cutoff point. This might be most useful if random assignment is not feasible for all the participants but can be justified for a smaller number of "test" participants. There are three ways in which this could be done. First, one could randomly select points on the pre-program measure at which random assignment to condition would be used. This is superior to randomizing within a small interval because it allows one to generalize across a wider range of scores. Second, one could systematically select values or ranges on the pre-program measure within which participants are randomly assigned (Hansen, 1977). This would be useful in Title I evaluation if it could be done at the extremes of the pre-test

distribution. Here, any floor or ceiling effects would be accounted for by the fact that they would be exhibited equally, on the average, in both the treatment and comparison groups. Finally, one might be able to randomly assign within certain subgroups or subpopulations. While it might not be possible to randomly assign throughout an entire school district it might be feasible to select several schools as test sites and to randomly assign within those schools.

As above, several analyses would be appropriate. One would place greatest confidence in the analysis of the randomly assigned cases and could include other cases for purposes of increasing precision or for completeness in reporting. A major reason cited for not using random assignment in Title I evaluation is because it will result in a situation where students who need and may be eligible for service are denied such service. This criticism makes less sense when one considers that the procedures for allocating Title I services to schools lead to students (who qualify for the program on the basis of their pretest score) being denied service because they come from designated non-Title I schools. It should be noted that random assignment to treatment or comparison group is advocated within the Title I evaluation system as a variation of Model B but this design is almost never used. Nevertheless, some of the advantages which result from using random assignment can be achieved through coupling such strategies with Model C, the regression-

discontinuity design, as mentioned here.

There are several other useful applications of random assignment within the regression-discontinuity design. This includes the random assignment of schools to Title I or non-Title I status or the use of a cutoff interval in assigning schools with all schools within that interval being randomly assigned. One could easily devise analogous designs at the classroom level or, perhaps, even at the school district level. In general, embedding random assignment procedures within the regression-discontinuity design will improve the methodological qualities of the study and should be encouraged.

In addition to random assignment to conditions there are times when random selection from a larger population will improve the design and its analysis. It is, for example, acceptable within the Title I evaluation system to select randomly a comparison group from the population of comparison students in a district. In a district which has a large number of students it may be more efficient to select randomly a small comparison group for purposes of analysis, although in general, it would be preferable to include all such students if that is possible.

Another use for random selection would be possible if Title I programs are administered in two phases with half of the eligible students receiving service in the first half of the school year and the other receiving it in the second

half. Here, one randomly selects from eligible Title I students those who will get treatment first and those who will get it second. The advantage of such a strategy is that one has a suitable comparison group (namely, those students who are not serviced in the first half of the school year), without the need to ultimately deny service to any eligible students. If such a strategy proves unworkable it might be possible to modify this by instead randomly selecting from eligible Title I schools those schools which will implement the program in the first half of the year and those which will do so in the second. This is a very powerful methodological alternative to the three models which are currently used for the Title I evaluation, and should be encouraged. Three separate analyses could be used. The first would involve a comparison of students who received Title I service in the first half of the year and those eligible students who did not. The second would involve the comparison of the same two groups after the students have all been served. In the first case, one would expect that if the programs have an effect such an effect will be detected. If one is found, and if the second phase of the program also has an effect, one would expect that the original difference between the two groups would diminish in the second analysis. Finally, a standard regression-discontinuity analysis could be applied for both treatment groups separately or the two combined. One can readily conceive of interesting, and at times, even more

complicated design variations. For example, students who receive service in the first half of the year could be assigned to either continue service or not for the second half of the year. This assignment could be random or could be by means of a cutoff score. Thus, one could have a regression-discontinuity design imbedded within a true experiment which is in turn imbedded within a larger regression-discontinuity design.

When random assignment is not possible it is often useful to search for other comparable groups of students which can act as comparisons in the analysis. This essentially involves coupling the Pretest-Posttest Non-Equivalent Group Design (Reichardt, 1979) with the regression-discontinuity design. With this design, equivalence of the groups is not assumed, that is, non-equivalence is allowed. Consequently, the potential for biases due to selection is ever present. A strategy of this type is readily available in Title I evaluation because of the existence of non-Title I schools which are likely to have students who score below the district cutoff point and hence, could be used as comparison students. Matching of students from Title I and non-Title I schools is not advisable (Campbell and Stanley, 1963) and no ready strategy for selecting a "comparable" group from the non-Title I schools is available. As a result, an analysis like this should be interpreted cautiously. There may be times when it is possible for a design of this type to be

incorporated at school district levels. This might be so if a school district which is not receiving any Title I funds can be located and is comparable to the district receiving funds. This is not likely, however, due to the fact that most school districts in the country receive at least some Title I funding.

Another variation on assignment strategies would involve the use of multiple cutoff points with the group most in need being served first, the next needy group second, and so on. In this case, the first group could stop receiving service when the second group ends or each group could continue receiving service once it has begun. Separate analyses could be conducted comparing the most recently serviced group with all persons not yet receiving service, or a single analysis could be run by including suitable treatment by order of admission terms into the analysis. It is important to recognize with this design that there will probably be a need for repeated testing of all subjects. Thus, a pretest is administered, a test is administered as a posttest for the first treatment group and all comparisons, another test is administered as the posttest for the second treatment group and all other subjects, and so on.

A variation of this stagewise design is suggested in Riecken et al. (1974) and is termed a "trickle" design. This design is especially useful for an agency which is continuously processing persons through its program such as a hospital,

mental health center, or social agency. It is necessary in this variation that the number of applicants exceed the number of available spots in the program at any given point in time. All applicants are tested when they present themselves to the agency and put into a pool for the next cycle of the program. For this pool of applicants a cutoff score on the pre-program measure is selected on the basis of the number of available spots in the program. The problems which occur here are related to the fact that one is likely to have different cutoff points each time the program is run depending upon the number of applicants who present themselves. Nevertheless, this variation of the design would be a useful method for agencies which are continually processing applicants.

The assignment to conditions under the regression-discontinuity design is dependent upon the measure or measures used to assign. Technically, the assignment measure need not be statistically related to the posttest although this may be desirable for conceptual reasons and for statistical power. In addition, it will often be useful to construct a composite variable made up of several measures suitable for assignment. In Title I evaluation such a composite variable might include an achievement test score, a grade point average, a rating of need by a teacher or admissions committee, and the like. Technically, it is not necessary that the variables which go into this composite be continuous ones. One could use

dichotomous designations, rankings, or ratings, such as on a five-point scale ranging from very acceptable to not very acceptable.

It is desirable that the composite measure be close to a normally-distributed continuous variable with adequate measurement properties. For example, if a five-point Likert rating of the acceptability of a student for the Title I program is used, it would be advantageous to have several teachers, administrators, or others complete this rating and to average the scores for any given student. As a result the five-point scale can be transformed into a scale with finer gradations than the original five points. The major problems in developing composite assignment measures are related to how individual variables should be combined and how missing observations on these variables should be handled. The formula for the composite measure could be additive, multiplicative, a combination of these, could be based on a ratio of several variables, and so on. Each variable could be weighted equally or in accordance with the preferences of the particular schools. Another strategy could be to use multivariate statistical methods such as multiple regression analysis or factor analysis for devising a composite measure. It is also not clear how one should proceed in the event that there are missing observations on one or more of the variables which enter the formula. This would be analogous to the univariate case where the pretest

value is missing. One might attempt a more complicated assignment model for the case of missing data by utilizing other information to estimate the missing values, but this would probably be of dubious quality. Issues related to the development of composite measures have been discussed by the RMC Research Corporation (1976), Campbell (1969), Cordray (1978) and by Boruch (1973).

Treatment Variations. Variations of the basic regression-discontinuity design can be developed by altering the usual dichotomous treatment-control designation. Many of these variations have been mentioned earlier and, as a result, will only be discussed briefly. First, it is possible to have multiple levels of the treatment or to test different versions of the program. In Title I evaluation, for example, one could test different amounts of treatment (e.g., five days versus three days a week versus one day a week), different settings (e.g., in-class versus pull-out), as well as different methods of instruction. The sets of comparisons can be made with or without a comparison group, although they will be stronger if such groups are included. When this is not possible, as in the case of federal grants where all potential recipients receive some degree of funding, comparison of different levels of treatment, especially if the difference between them is quantifiable, will often provide useful information relevant to the program effect.

Post-Program Measure Variations. It has been mentioned throughout this work that it is not necessary for the pretest and posttest to be the same measure. In fact, any measure which is relevant to or is likely to be affected by the program may be used as a posttest. In addition, it is not necessary to limit one posttest for each pretest, that is, several measures can be analyzed for the same pretest.

In Title I evaluation there are several ways in which alternative post-program measures might be incorporated. Most standardized achievement tests, for example, are comprised of several subscales or subtests. Thus, a reading score might be a composite of scores obtained from tests of spelling, grammar, and punctuation. There is no reason why the analysis of the effects of a reading program should be limited to looking only at effects on the total reading score. Separate analyses could be conducted using each of the subscales as post-program measures.

To illustrate use of different measures of post-program performance, data from the third grade reading and fourth grade math programs in Providence, Rhode Island, were reanalyzed using subtests in place of the posttest. In the secondary analyses described in Chapter III, the third grade reading program evidenced no effect, or possibly, a slight positive one, while the fourth grade math program exhibited clear negative gains. Analyses of posttest subscales might make the interpretations of the reading program less

equivocal or the negative gains of the math program more specific.

Two posttest subscales which measure vocabulary and comprehension were available for the reading analysis. Three subscales measuring computation, concepts, and application were available for the math data. The sample size, mean, variance, skewness and kurtosis for all five measures are shown in Table 30. The same four-step polynomial regression model used for the secondary analyses and described in Table 11 is applied here. For the third grade reading program, the bivariate plots are shown for the vocabulary subscale in Figure 19a and for the comprehension measure in Figure 19b. For the fourth grade math program the plots for the computation, concepts, and applications subscales are shown in Figures 20a to 20c respectively.

The estimates of gain and their standard errors are shown for each step in the analysis in Table 31. Both reading subscales show a similar pattern of gains across steps. On the basis of the analysis we would conclude that there is no clear program effect. Although inspection of the graphs provides some evidence for an effect, assuming a true linear relationship, there is some evidence for slight curvilinearity especially at the lower end of the treatment group scores. This could be due to a floor effect or chance level problem.

The analyses of the math subscales are also illuminating.

The only measure which shows the consistent negative gains across steps is the concept subscale. Since similar gains were detected in the secondary analysis using the total posttest math score, one might conclude that they can be attributed primarily to this subscale. The implication is that Title I students in that program lose ground (relative to comparison students) most on conceptual mathematical skills and that increased attention to this aspect of instruction may improve future program results.

One might wish to look at other post-program measures of importance such as the self-esteem of students and estimate the effects of Title I programs on these constructs. Other variables useful in "process" evaluation or to examine potential threats to the major analyses could be considered, such as attendance rates, migrancy, and so on. Just as the assignment variable can be a composite there is no reason why the posttest cannot be comprised of several measures. One might wish to develop indices related to program effect and use these as the post-program measures.

A major question of interest in Title I evaluation concerns monitoring the effects of the program over extended periods of time. In this regard, it would be useful to look at a posttest two or three years removed from the program. Generally, it would be beneficial for the context of Title I evaluation to encourage the use of the regression-discontinuity design for the estimation of program effects on

more than simply a single measure of achievement.

Aggregation Variations. Many variations of the regression-discontinuity design for aggregated data have been discussed throughout this work. Sometimes the aggregation is in terms of the unit of analysis as when the design is applied to a higher level of allocation formula (e.g., when allocating from the federal government to the states). At other times the analysis can be performed on aggregated data, such as means for various units of analysis. Thus, if one is looking at the effects of Title I instruction at the school level one might use an estimate of the mean reading achievement score for each school. The use of the regression-discontinuity design on aggregated data is discussed in more detail in Boruch (1973).

Post Hoc Analysis. The regression-discontinuity design would appear to have special attractions for conducting analyses "after the fact," that is, on a post hoc basis. There are several reasons why post hoc analysis with this design is promising, especially for social programs which are allocated on the basis of a formula as described later. First, it is often the case that the data for these programs, that is, the pre and post-program measures, are routinely collected by government agencies. Even in cases where this data is not collected routinely it is often possible to construct appropriate indices which can be used. Second, this data is often readily accessible in the form of

government documents or computerized magnetic tapes. Often the major problems which will arise concern locating and accessing this data and determining its quality (Trochim, in press). Third, it will often be possible to conduct a regression-discontinuity analysis even if the allocation formula does not require sharp assignment in terms of a cutoff score. As long as there is an interval on either side of which the units are perfectly assigned to condition, the regression-discontinuity analysis may be legitimate.

Coupling the Regression-Discontinuity Design with Federal Allocation Formulas

We turn our attention here to the broader range of applications of the regression-discontinuity design which involve evaluations of federal social programs in general. Because Title I of ESEA is one example of such a program these design variations may be appropriately applied there as well as elsewhere. No one research method could ever be appropriate for the evaluation of all federally-sponsored social programs. In order to get a clearer idea of where the regression-discontinuity design could be most usefully applied, it is helpful to first define the types of federal grants which are available and the different allocation formulas which are used. Much of the material for this discussion was obtained through the many reports of the Advisory Commission on Intergovernmental Relations (A.C.I.R.,

1978). This group is charged by Congress with monitoring the operation of the American federal system and recommending improvements. Over the past few years they have conducted a study of the intergovernmental grant system, including an in-depth assessment and policy recommendations. Other sources of importance are the Catalog of Federal Domestic Assistance (O.M.B., 1979) which describes each federal grant and lists major documentation, and the Report on Statistics for Allocation of Funds (U.S. Department of Commerce, 1978) prepared by the Subcommittee on Statistics for Allocation of Funds, Federal Committee on Statistical Methodology, U.S. Department of Congress.

Types of Federal Grants. The grant system can be divided generally into three types of grants, the typology primarily dependent on the degree of intended specificity for the use of the funds. The first type of aid, General Revenue Sharing (GRS) has the fewest spending strings attached. Funds from GRS are allocated to local governments by formula and can be spent on whatever the local government considers necessary. Therefore, in some communities GRS money is used to defray the costs of necessary municipal services like police and fire protection while in others it is directed to other services which the community would be hard-pressed to provide on its own, like social services and aid to the elderly.

The second general type of grant is known as the block

grant. By one counting there are five major block grants: the Partnership for Health, the Omnibus Crime Control and Safe Streets Act, the Comprehensive Employment and Training Act (CETA), the Housing and Community Development Act, and the Title XX Social Services Act. Funds for these grants can be spent on a wide variety of programs, as long as they can be justified in terms of the intent of the act. Therefore, while they are more specific than GRS, they are not as narrowly defined as the third type of grant -- categorical aid.

By far, the largest number of federal grants (in 1975, 442) are categorical, that is, expenditures are confined to a specific category or type of program. Furthermore, the categorical grants can be divided into two categories: formula or project. Generally, with formula grants (146 in 1975) the funds are allocated from the federal government to smaller governmental units by means of a formula. Project grants (296 in 1975) are open for proposal bids. Funds are given out based on judgment in the allocation process. First, they are often used to allocate money from the federal government to the major recipient, most often a state. In turn, a second level of allocation may be set up to disperse funds within each state to the local government. There can, in some cases, be additional levels of allocation from local governments to sub-units. The procedures used to disperse funds for Title I provide a good illustration of a

multi-level allocation system. Here, funds are first allocated to the states and territories, then to school districts (often coterminous with country or parish boundaries), then to individual schools within the district, and finally to the programs which serve the students. At each level of allocation a variety of formulas incorporating different measures may be used. These can include demographic variables, measures of need such as index of poverty or socioeconomic status, or measures of achievement.

The use of formulas for allocation is not confined to the formula-based categorical grant. For example, GRS and several of the block grants (e.g., CETA) allocate funds by formula. In addition, many of the categorical project grants, after allocating funds on the basis of proposal awards, require or allow the use of a formula in the next level of allocation.

A useful distinction can be made between allocation formulas and eligibility formulas. An allocation formula is used to disperse funds from one level of government to another. An eligibility formula is used to screen the potential recipients of the program or funds. Medicaid, for example, makes use of both types. Funds are allocated to states on the basis of the formula relating state and national per capita income. Beneficiary eligibility is determined by a separate formula based on age and disability. A similar distinction can be made for the allocation

procedures used for Title I. The final state in the allocation process is to determine eligible participants. Thus, as in Title I, the use of a cutoff score on an achievement test can be considered an eligibility formula which forms the last step of the funding allocation process.

In most cases it is the Congress who clearly and specifically state the allocation formula and the factor weightings. Several classifications have been offered of the types of variables used in allocation formulas. One group (A.C.I.R., 1978) suggests three major types of measures, those related to population, financial need, and program need. Specific measures are used in different ways depending upon the grant. Measures related to population, for example, can be used as continuous measures or as "qualifiers." Thus, the allocation of funds could be based on the number of persons residing in each defined geographical area (e.g., county, city, school district), or funds could be dispersed to all areas having greater than a specific percentage of the national population (e.g., urban areas). Measures of financial need are usually related to income. Funds could be dispersed on the basis of the per capita income in a geographical area. Program need is sometimes reflected through a measure of income and at other times is a more direct measure of need. For example, a program for boating safety might use the number of vessels registered while a medical aid program might rely on

estimates of the number of available hospital beds. Sometimes population or demographic measures are presumed to reflect program need if only because they tend to define the constituency of interest. Thus, funds for an adult education program might be restricted on the basis of the number of persons over 18 years of age. Another classification of formula variables (U.S. Department of Commerce, 1978), divides the types of measures into the categories of need, capability, and effort. Here, need refers to population and demographic measures as well as more direct indicators of need. Capability refers to the amount of revenues which a local or state agency might be expected to provide from their own taxes and income. Effort relates to the amount which is already being spent for a given program from non-federal sources. However these measures are classified, it is clear that quantified indicators are often used in the allocation process.

The formulas which are used differ considerably in structure. As an example, one can consider the formula which is used to allocate GRS funds. The formula is essentially a ratio of "effort" (the taxes collected) to a measure of "capability" (per capita income) multiplied by total populations (which is presumably reflective of need). Generally a given variable will be used either directly in the formula or as a "qualifier" or "constraint." These are used to define the target populations and/or to protect

against sudden fluctuations in allotment from one year to the next. One example of this can be seen in what is commonly termed a "hold-harmless" provision. Here, funds are allocated in a given year by means of a formula, but any recipient who does not qualify for the current year and who received funds in the previous year may be allowed to continue receiving funds or to have these phased out gradually over a period of several years. Provisions of this nature are included because of political considerations or a desire to provide continuity in the programs which are offered from one year to the next.

Examples of Coupling Regression-Discontinuity with Allocation Formulas. It is apparent that many of the formulas used to allocate funds meet one or more of the requirements for the regression-discontinuity designs. Many rely on continuous quantified measures or combinations of measures to determine the amount of funds which are allocated. Some specify a cutoff value for allocation or eligibility while others have multiple cutoffs or "cutoff ranges" within which the amount of funds varies continuously.

To illustrate how the regression-discontinuity design could be coupled with allocation formulas consider the allocation of Title I funds from the school district to individual schools. Here, a measure of poverty (presumed also to be reflective of educational disadvantage) is used. This might be a measure like the percentage of free lunches

which are served in each school or the number of children in families receiving AFDC. Schools are ranked by this measure and designated as eligible for Title I funds on the basis of a cutoff. Clearly, several of the requirements for the regression-discontinuity design are met here. The pre-program measure can be the estimate of free lunches and schools divided into "treatment" or "comparison" groups solely on the basis of a cutoff score. The post-program measure could be an estimate of average achievement test scores for each school.

This example illustrates some of the difficulties which will arise when trying to couple the regression-discontinuity design with allocation formulas. One problem concerns the fact that not all eligible schools receive the same amount of funds and only a small percentage of a school's population will receive services. As a result, a post-program measure which reflects the average achievement for the school will only be partially reflective of the effect of Title I services. While this might appear to be a serious difficulty it is simply a variation of the problem which normally occurs in trying to estimate the degree of treatment implementation. Just as the amount of service given to each student will differ even within the same Title I program and classroom, we expect that the amount of Title I services differs from school to school. It would be necessary in this hypothetical design to estimate a "weighting" factor which indicates the

degree of Title I service provided at each school. This is analogous to estimating the amount of service which a given student receives. For example, one might consider the percentage of students in a given school who are enrolled in a Title I program as a suitable covariate in the analysis.

A second example can be constructed to illustrate how the regression-discontinuity design could be used to evaluate the Airport Development Aid Program (Public Law 91-258). The objectives of this law are "to assist public agencies in the development of a nationwide system of public airports adequate to meet the needs of civil aeronautics" (O.M.B., 1979). Generally, funds are used for "constructing, improving, or repairing a public airport or portion thereof." All public airports are eligible for funds, but the level of funding differs depending upon a cutoff value on a pre-funding measure. The measure used to determine the amount of funding is the number of enplanements at each airport. The federal government will pay not more than 75% of the cost of a project at an airport which enplanes one-fourth of 1% or more of all passengers enplaned at such airports and, will pay 80% for projects at all other airports. Thus, the cutoff point is "one-fourth of 1%" of all enplanements.

This example illustrates a different set of difficulties than the previous one. Here, one is not testing the difference between a "treatment" and "control" group, but rather between two levels of funding. In addition, the

levels of funding are not all that different and one might not, depending upon the program of interest, predict that a difference of 5% in the funding level will result in differences in post-funding measures. In this example, the pre-program measure is the number of enplanements, the airports are divided into different levels of treatment by means of a sharp cutoff value on this measure, and the post-program measure might also be the number of enplanements in a subsequent year or some measure related to airport activity. As in the previous example, it would be useful to search for "weighting" variables which reflect the amount of funding or treatment implementation at each airport.

Another example can be constructed using the regression-discontinuity design to evaluate the Comprehensive Employment and Training Act (CETA) programs. The objectives of CETA programs are "to provide job training and employment opportunities for economically disadvantaged, unemployed, and underemployed persons to assure that training and other service lead to increased earnings and enhanced self-sufficiency by establishing a flexible decentralized system of Federal, State, and Local programs" (O.M.B., 1979). The allocation formulas for the CETA program are extremely complicated due to a large number of qualifiers and exceptions. In addition, the grant is divided into six major parts or "Titles" and each part has its own set of formulas. Of special interest here is Title VI which provides public service employment in areas

of high unemployment. The allocation of funds for Title VI includes the following provisions:

> Not less than 85% of the funds appropriated are alloted as follows: (a) 50% in proportion to each area's share of all unemployed persons; (b) 25% in proportion to the area's share of all unemployed persons in excess of 4.5% of the labor force; and (c) 25% among areas of substantial (6.5% for three consecutive months) unemployment as defined in the CETA regulations (O.M.B., 1979).

This formula appears to have two cutoff points, one at 4.5% unemployment and the other at 6.5% unemployment. Those areas below 4.5% unemployment receive the least amount of funds, those above 6.5% receive the most, and those between the two cutoffs are allocated funds on the basis of a separate formula. Here, we have a situation analogous to the airport development grant where we are dealing with different levels of treatment or funding. The pre-program measure would be the indicator of unemployment, units are assigned to groups on the basis of the two cutoffs, post-program measures of unemployment and other variables could be included and, as above, estimates of the amount of funding or the degree of program implementation could be incorporated as covariates.

These three examples serve to illustrate how the regression-discontinuity design could be coupled with allocation formulas. All have in common the existence of a continuous quantified pre-program measure and a specified cutoff value or values which are used, at least in part, to determine program funding or eligibility. Many of the problems of design implementation discussed in Chapter II

will also tend to occur in these applications. In addition, problems related to the context of federal allocation procedures will also occur and are discussed in the next section.

Problems in Coupling Regression-Discontinuity with Allocation Formulas. Every research context and in fact, every federal grant, is likely to pose its own set of problems for any methodology. It is partially because of this that the idea of design implementation analysis was advocated in Chapter II. In this section, the context of federal allocation formulas is generally examined for likely potential problems for the application of the regression-discontinuity design. It is important to keep in mind that the grants administered under the federal government cover an impressive range of purpose and form and that the use of the regression-discontinuity design or any other should be examined carefully before it is applied to any particular probject.

A major problem in using the regression-discontinuity design coupled with allocation formulas concerns the number of units which are available for analysis. The first level of allocation in many grants is from the federal government directly to the state and trust territories. It would be difficult indeed to have confidence in an analysis based on between 50 and 60 data points, although low variance in these values could allow for a reasonable analysis. In many cases

the state subsequently allocates funds to smaller units of government within their boundaries. This may or may not be done by a predetermined formula. Often, the states are given guidelines for this second level of allocation but are not required to restrict themselves to a predetermined formula. It will sometimes be the case therefore that there will be too few units to justify a regression-discontinuity analysis. The possibility of "skipping" an allocation level and doing an analysis on sub-state units will often be ruled out because of inconsistencies in formulas or data between states.

A second problem which will often occur concerns the distributions of the variables which are available for the analysis. The statistical procedures used to analyze the regression-discontinuity design for the most part assume that the distributions for the variables are normal or that they can be transformed into normally distributed variables. Previous experience indicates that many of these variables, especially those related to income, are not likely to be so distributed. While procedures for analyzing non-normal distributions are being explored (Sacks and Ylvisaker, 1976), distributional problems may rule out the use of the design for certain projects at this time.

Federal funds are often dispersed under certain "matching" considerations. In these cases, the state and/or local government is required to provide a certain percentage of

the funding. The total amounts of funds which are available are determined in part by these local governments. The most common matching ratio is 50-50, or half federal funds and half funds from other sources. Problems generated by matching provisions may be able to be at least partly accounted for in regression-discontinuity analysis by the inclusion of a variable which measures a level of program funding in each unit.

Another potential difficulty concerns the fact that at least at certain levels of allocation, there are no comparison groups. Especially when the federal government allocates to the state it is usually the case that few if any states are denied at least some funding. This is probably due in part to political considerations. We are left, in the absence of comparison groups, with a strategy which attempts to test differences between various levels and types of programs rather than between the program and no program at all. This was illustrated in two of the hypothetical examples outlined above.

Numerous constraints or exceptions are often allowed for a given formula. These are dictated by the political necessities involved in order to get the bill through Congress, but nevertheless pose some difficulty for the application of the regression-discontinuity design. Sometimes it takes the form of allowing the director of the appropriate federal department to have certain discretionary

powers in allocating funds while at other times it is evidenced by the inclusion of specific subgroups who would not necessarily qualify by means of the formula but should be included anyway. These exceptions result in a class of problems similar to those posed by the exclusions allowed by Title I evaluations as indicated in Chapter II. The same approaches which are advocated there could be applied here. This might include separate analyses for different subgroups of interest to determine whether estimates of program effect differ, as well as exploratory analyses designed to assess the degree of this difference.

Another problem which holds some similarities with problems evidenced in Title I evaluation concerns the existence of competing programs to which a particular governmental unit could apply if they do not get federal funding from the government. Thus, state and local governments may already have their own programs for airport development, compensatory education, or manpower training. It may even be the case that these programs exist primarily to provide funds to governmental units which are denied funds from the federal government because of their placement by the formula. When similar programs exist for purposes of equalizing funding to governmental units these can act to degrade a comparison between treatment and comparison groups or between different levels of the treatment.

A variety of measurement problems can occur depending

upon the program and the variables which are used. A major problem results from the fact that many of the variables used in allocation formulas, especially those related to population and income, are obtained from census data which is collected every ten years. Sometimes an attempt is made to provide updated estimates, but these are based upon fallible prediction models and must be considered suspect. Some work has been done on developing continuously measured indices, such as the cost of living index and the unemployment index, and such developments should improve measurement qualities of the regression-discontinuity design.

Another problem stems from the fact that the regression-discontinuity design requires the pre and post-program measurement of comparison subjects or units. It will often be the case that the information of greatest interest when determining the effect of a federal program will only be collected for the treatment group. For example, in the case of Medicaid programs, relevant indicators of the effects of medical care may only be collected (or even if collected, may only be aggregated) for the treatment group because that is the group of prime interest.

A federal grant is simply an offering from the federal government to other governmental agencies or individuals. Participation in programs or even application for such programs is by no means mandatory. There are many cases in which an eligible unit will not apply for or receive the

program. This may occur if the governmental units recognize in advance that they will not qualify, if they are unwilling to submit to federal requirements and paper work, or if they are simply unaware of the existence of the program. In any event, this self-selection process is likely to have an effect on the constitution of the groups of interest and, subsequently, on the analysis itself.

One of the most difficult problems in applying the regression-discontinuity design to allocation formulas concerns the placement of cutoffs and the assignment to conditions. Given the political factors which it must consider, Congress is hard put to develop formulas which are simply constructed just for the sake of expediency or methodological clarity. In addition, it is not clear that the sharp cutoff for assignment which is a requirement for the basic regression-discontinuity design is always politically desirable or feasible. To get an idea of the difficulty involved in decisions concerning cutoff points consider a passage from the Report of Statistics for Allocation of Funds (U.S. Department of Commerce, 1978):

> Undesirable discontinuities may be introduced into an allocation system by cutoffs, especially by cutoffs. For example, if an area must have an unemployment rate of 5% before it can receive any funds, a very trivial error in the estimation of the unemployment rate can easily throw an area from under 5% or from over 5% into the other group. Here a very small error can make a tremendous difference and lead to continual complaints about the accuracy of the data on the part of the governmental units which

feel the cutoff operates to their disadvantage.

A common solution to controversies over cutoffs is to provide alternative cutoffs and to permit each jurisdiction to select the formula which is most advantageous. While this works moderately well, it has the disadvantage of making it difficult to predict in advance (and budget for) the amount required for the process if no fixed overall sum to be allocated is specified. If no overall sum is specified but each jurisdiction may choose which formula it will use in determining its share (with computed amounts totaled over all competing jurisdictions so that the per cent of the total allocated to each jurisdiction can be determined), one gets a floating cutoff point where the amount one jurisdiction gets depends upon the decisions made by other jurisdictions.

For eligibility cutoffs it is almost always possible to devise a formula such that there is a gradual approach to zero (or to some cutoff point lower than the existing absolute cutoff). Here, small errors in the data lead only to small changes in the allocation and the tendency to prolonged (and insoluble) arguments over minor errors is removed. Of course, major errors will and should continue to be the subject of controversy but one will be spared the waste of time and effort involved in the use of a formula which requires data of unattainable accuracy.

The use of a formula based on a continuous allocation of funds implies, at least for some cases, that there will be multiple cutoff points. All units scoring below the lower cutoff point get one level of treatment, those scoring above the highest cutoff point get another level, while all those scoring between these two get a continuously varying level of treatment. This was illustrated in the example above for the CETA program. The regression-discontinuity design or a variation thereof could be appropriate for these cases. What must be avoided is the use of a formula which allocates continuously different levels of funding across

the entire range of the pre-program measure. As long as there are clear cutoff points which differentiate between the level of funding or treatment in adjacent groups (as measured on the pre-program variable), one expects that if these levels of treatment or program have an effect this will show up as a discontinuity in the relationship between the pre and post-program measures.

We can see from this discussion that, just as in Title I evaluation, there are a large number of potential problems which can act to degrade the analysis of program effect by means of the regression-discontinuity design. Nevertheless, the sheer magnitude of federal domestic aid, the need to evaluate the effectiveness of these programs, and the potential usefulness of the regression-discontinuity design in this regard, warrants a more detailed explanation of the appropriate uses for the design in this type of context.

V. CONCLUSIONS AND RECOMMENDATIONS

The previous chapters describe the major problems surrounding the regression-discontinuity design. These fall into three general categories: those related to the theory of the design, its implementation, and its analysis. This chapter sets forth several recommendations which, if adopted, are likely to improve the quality of research which is conducted using regression-discontinuity.

It is important to keep in mind that there is no such thing as perfect research or the perfect evaluation. These are merely ideals which the research community can approach but never reach. The important question is not whether regression-discontinuity "works," but rather, how well or how poorly it is used. It is tempting to argue that, in general, school districts lack the sophistication necessary for carrying out methodologically competent evaluations and hence, should not be encouraged or required to do so. A more optimistic viewpoint would hold that districts are carrying out research, however difficult that task might be, and that over time the quality of research can be improved by isolating the major problems and eliminating or minimizing them. It is in this spirit that the following recommendations are offered.

Recommendations for Title I Evaluation

The recommendations presented here are not comprehensive but instead are directed to the most important problems which affect the design. Not every school district could be expected to incorporate any or all of these suggestions but to the degree that they can, one would expect the quality of research to be improved.

1. Select measures so as to minimize problems related to curvilinearity in the data. Specifically, select tests with lower chance levels and which are least likely to yield floor or ceiling effects.

The most serious problem in regression-discontinuity analysis involves specifying the true pretest-posttest relationship. When the data appear to be non-linear it is often impossible to determine whether this is the true shape or the result of measurement problems. Ideally, one wishes to eliminate potential for the latter, but in practice this is difficult to achieve. More realistically, the recommendation holds that, all things being equal, it is better to select a test which has the lowest chance level possible. In addition, while it is desirable to avoid floor and ceiling effects, if this is not possible it is preferable in Title I contexts to allow ceiling effects rather than floor effects. This is because in most Title I evaluations there are more comparison than treated students. Non-linearities in the extreme of the comparison group distribution are not confounded with the treatment and can be

more easily investigated with auxiliary analyses which eliminate extreme cases or weight them less heavily.

2. The cutoff point should be chosen to be as distant from the extremes of the pretest distribution as is feasible.

This recommendation is also directed to the problem of curvilinearity. If, for example, it is likely that there will be a floor or chance level effect in the treated group, it is desirable to have as many cases as possible in this group which do not exhibit these effects. Selection of a higher cutoff value will reduce the effects of such problems on estimates of gain and allow more flexible analysis based on weighting cases near the cutoff more heavily.

3. Challenges (i.e, misassigned cases) should not be allowed. When some challenges are unavoidable, the treatment effect should be estimated both with and without these cases.

The challenge procedures common to most districts violate the assumption of sharp assignment which is central to the regression-discontinuity design. If some discretion in assignment is desired, this should be accomplished through the use of composite pre-program measures which incorporate teacher's or administrator's judgments about the need for service. In the event that some misassignment does occur, one should conduct multiple analyses to determine the effect of the challenges.

4. The data should be graphed in as many ways as are feasible, both with and without any subgroups which are considered candidates for exclusions.

Always plot the treatment and comparison cases on the same graph.

Visual inspection of the data is essential for determining the accuracy of the data, the potential for degradation posed by various subgroups, and the existence of curvilinearity. The common Title I practice of plotting the treatment and comparison groups separately should be avoided because it makes it nearly impossible to view the overall shape of the data and visually determine whether a discontinuity occurs at the cutoff point.

5. The analytic model discussed in Chapter III should be used to obtain multiple estimates of the treatment effect at the cutoff.

This recommendation holds only if it is reasonable to apply a polynomial model to the data or transformations of the data. Probably the best strategy would be to compare the estimate of effect when trying to exactly-specify the true model with the one obtained when deliberately overfitting the likely true model. The former is efficient but may be biased, the latter is less efficient but unbiased. If the two estimates are similar, one can feel more confident in accepting the exactly-specified one. If they differ, one might prefer to use the over-specified estimate or to report a range of estimates. While the use of multiple estimates of effect which may differ is an unsatisfying approach, it explicitly acknowledges the limitations of currently available analytic strategies and offers a range of

values within which the true value is likely to fall.

Several changes in the typical Title I analysis are advisable. First, wherever possible, raw test scores should be used in order to avoid scaling problems caused by nationally standardized test scores. If different levels or forms of a test are used within grade, local standard scores are generally preferable to national ones. Estimates of gain can be converted to NCE units after being calculated from raw or local standard scores.

Second, use of the general model outlined in Chapter III is preferable to the typical Title I model. The assumption of linearity which is generally made is unjustified in many cases. In addition, the model in Chapter III subsumes the Title I analysis as one step in the procedure and hence, can be considered a more general and flexible approach.

Finally, the practice of estimating the effect at the treatment group mean should be abandoned. This estimate was originally offered because it was thought to be comparable to estimates from the other two models and reflective of the gain of the "average" treated student. In fact, there is reason to believe that the estimates from the three models should not be comparable (given the bias which exists in Model A and the different implementation problems which can affect each model) and that use of the pretest treatment group mean to designate the "average" treated student is conceptually unsound (given that one expects the treatment group pretest

distribution will always be skewed because it is truncated at the cutoff). In addition, the estimate at the mean involves extrapolation from the comparison group regression line. The further one extrapolates into the treatment group region the greater the chances for biased estimates of effect.

6. Conduct multiple analyses to assess the potential for degradation due to problems like floor and ceiling effects, exclusions, and so on.

A conscientious analysis demands that groups not be excluded from the analysis without determining the effects of the exclusion. Even if analyses are conducted with and without scores from a group which is considered for exclusion, it may be impossible to avoid distorted estimates of effect. This was illustrated in Chapter II for "challenge" cases where either including or excluding them is likely to yield biased estimates.

7. Incorporate random assignment wherever possible.

The regression-discontinuity design can be strengthened considerably by incorporating random assignment, especially in the vicinity of the cutoff score. This practice should be encouraged for Title I evaluation because it capitalizes on the methodological advantages of Model B (i.e., of Model B when random assignment is used) and the slightly greater feasibility of Model C.

8. Select a subset of all programs for more intensive study.

Given that it may be difficult for many districts to

maintain high research quality throughout, it is recommended that one or more programs in the district be targeted as test sites and extensively monitored for design implementation problems. This serves two useful purposes. First, it enables a more exact specification of the major difficulties of design implementation. Second, estimates of effect can be compared to district-wide estimates to assess the effect of these difficulties. For example, a district could select a relatively small Title I program (e.g., perhaps a hundred students, twenty-five of which are treated) and could directly observe test administration, individually check coding errors, monitor the Title I class, and so on. Sites might be selected on the basis of the previous year's results, that is, the "worst" programs of the previous year could be monitored in order to search for reasons for their poor performance. While such an approach is susceptible to reactive effects due to intensive monitoring, it provides a useful mechanism for encouraging incremental improvement of research quality.

9. Document the research process as thoroughly as possible.

A major problem in this investigation was the lack of sufficient documentation for examining the effects of serious implementation problems. For example, challenges are often not noted, and if noted, the reasons for them are seldom distinguished; the reasons for missing scores are seldom divided into classes like absentees, migrants, computer

coding errors, and so on; the classification of treated versus untreated is sometimes not explicitly included in the data but rather is defined by means of the cutoff score. A thorough analysis of the effects of design implementation problems requires that they be recorded accurately.

10. For national aggregation, the possibility of requiring an estimate of gain based on overfitting the likely true model (as described in Chapter III) should be considered.

Despite the fact that overfitting yields less efficient estimates of effect, these estimates are less likely to be biased than others. This is important for national aggregation where an unbiased portrayal of effect is desirable and where individual inefficient estimates are more tolerable because of the gain in efficiency resulting from averaging across estimates. Thus, if every district reports gains based on, say, a fifth-order polynomial model (i.e., all polynomials and interactions from first to fifth-order as described in Chapter III) the national aggregation of these is likely to be an accurate portrayal of the effect of Title I programs. Individual districts may be able to justify local use of a slightly more efficient estimate while reporting the overfitted one for aggregation purposes.

Recommendations for Evaluation Methodology

The recommendations for methodologists are more general

in nature because they are directed at the design itself rather than a specific context. They are comprised of five major topics related to the design which would be useful to study in greater detail.

1. Design implementation problems need to be addressed more directly by the methodological community.

The ideals set forth by methodologists are not attained in practice. The idea of design implementation analysis as described in Chapter II offers an overall strategy for investigating the discrepancy between theory and practice. Because so little work has addressed these problems, practicing researchers have been on their own in devising solutions. While it is impossible to anticipate every potential problem, the most important ones should be expected and strategies for dealing with them should be developed.

2. The appropriateness of polynomial models for common regression-discontinuity design settings needs to be determined and the effects of analyzing non-polynomial functions with polynomial models needs to be assessed.

The analytic strategy suggested in Chapter III rests on the assumption that the true model (or transformations of it) is adequately described by polynomials in x. If this is not true, any analysis based on polynomial models is likely to be biased. Two questions are important. First, how frequently in practice is a polynomial model appropriate? Second, in practice how great a bias can be expected when analyzing non-polynomial functions with polynomial models?

3. More appropriate modes of analysis need to be developed.

The statistical analysis of the regression-discontinuity design is by no means a closed issue. The strategies outlined earlier are unsatisfying because they require multiple analyses which are likely to yield estimates of effect which differ. At the least, it would be useful to determine the degree to which estimates from competing analyses can be expected to differ in practice. In addition, it is important to investigate the likelihood of achieving biased estimates when attempting to exactly-specify the true model and the seriousness of the loss of precision involved in deliberately overfitting the likely true model.

4. Strategies for interpreting multiple estimates of effect based on competing analyses need to be developed.

While it may be more appropriate to report a range of effects rather than a single estimate, little work has been done on how to interpret such a range. Differences in the estimates of gain yielded by different analyses are due to the assumptions which are made. Sensitivity analysis might be useful for exploring the effects which different assumptions have on the estimates, although it is not likely to be feasible in many research settings. Suggestions for interpreting multiple estimates while not ignoring the inherent complexities of the analyses are in order.

5. Variations of the regression-discontinuity design and its analysis need to be

developed if the design is to be made more practically useful.

One arena of particular promise is discussed in Chapter IV and involves the coupling of the regression-discontinuity design with federal allocation formulas. In addition, analytic strategies appropriate for "fuzzy" regression-discontinuity are being developed (Trochim, 1980) and, if successful, will increase the range of applicability of the design. As the utility of the design is increased more attention will be given it by methodologists and practitioners and the quality of research is likely to improve.

This study demonstrates that the regression-discontintuity design is used, although its usefulness is diminished due to serious problems of implementation and analysis. What remains to be seen is whether the design, if implemented as well as is practically possible, is still seriously degraded by unidentified or uncontrolled problems. This judgment depends on future research: study of the theory of the design and its variations, the problems of implementation and strategies for statistical analysis.

REFERENCES

Advisory Commission on Intergovernmental Relations (ACIR). Categorical Grants: Their Role and Design. U.S. Government Printing Office, Washington, D.C., 1977.

Advisory Commission on Intergovernmental Relations (ACIR). A Catalog of Federal Grant-in-Aid Programs to State and Local Governments: Grants Funded FY 1978. The Intergovernmental Grant System: an assessment and proposed policies. U.S. Government Printing Office, Washington, D.C., 1978.

Barnow, B.S., Cain, G.C. and Goldberger, A.S. Issues in the analysis of selection bias. Unpublished manuscript, preliminary draft, August, 1978.

Bentler, P.M. and Woodward, J.A. A Head Start reevaluation: Positive effects are not yet demonstrable. Evaluation Quarterly, 1978, 2, 493-510.

Blair, C. Personal Communication, 1980.

Boruch, R.F. Regression-discontinuity designs revisited. Unpublished manuscript, Northwestern University, 1973.

Boruch, R.F. Regression-discontinuity designs: A summary. Presentation at the Annual Meeting of the American Educational Research Association, May, 1974.

Boruch, R.F. On appropriateness and feasibility of randomized tests of social programs. In L. Sechrest (Ed.) Evaluating emergency medical services. Washington, D.C.; U.S. Government Printing Office, 1978.

Boruch, R.F. and DeGracie, J.S. Regression-discontinuity evaluation of the Mesa reading program: Background and technical report. Unpublished Draft Manuscript, Northwestern University, Evanston, Ill., November, 1975.

Boruch, R.F. and Gomez, H. Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 1977, 411-443.

Boruch, R.F. and Wortman, P.M. Progress report: Project in secondary analysis. Psychology Department, Northwestern University, Evanston, Illinois, 1977.

Bridgeman, B. A practical alternative to Model A. Unpublished manuscript. Educational Testing Service, Princeton, N.J., 1979.

Campbell, D.T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Campbell, D.T. and Boruch, R.F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C.A. Bennett and A.A. Lumsdaine (Eds.), Evaluation and Experiment. New York: Academic Press, 1975.

Campbell, D.T. and Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory education: A national debate. New York: Brunner/Mazel, 1970.

Campbell, D.T., Reichardt, C.S. and Trochim, W. The analysis of the "fuzzy" regression-discontinuity design: Pilot simulations. Unpublished manuscript, Northwestern University, 1979.

Campbell, D.T. and Stanley, J.C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1966.

Chow, G.C. Tests of equality between sets of coefficients in two linear regressions. Econometrika, 1960, 28, 591-605.

Conner, R.F. Selecting a control group: An analysis of the randomization process in twelve social reform programs. Evaluation Quarterly, 1, 2, 195-244, 1977. Excerpted in Cook, T.D., DeRosario, M.L., Hennigan, K.M., Mark, M.M. and Trochim, W. (Eds.) Evaluation Studies Review Annual, Volume 3, Sage Publications: Beverly Hills, California, 1978.

Cordray, D.S. Making the Case for the Use of "Patchwork" Analyses in Quasi-experimental Evaluation Research. Unpublished Doctoral Dissertation, Claremont University, 1978.

Crane, L.R. and Maye, R.O. Effects of correctable errors on Title I NCE gain estimates and implications for statistical quality control. Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts, April 7-11, 1980.

Director, S.M. Underadjustment bias in the quasi-experimental evaluation of manpower training. Ph.D. Dissertation, Northwestern University, Evanston, Illinois, 1974.

Divgi, D.R. Choosing polynomials in the regression-discontinuity design. Unpublished manuscript, Syracuse University, Syracuse, N.Y., 1979.

Echternacht, G. A summary of the special meeting on Model C held in Atlanta in January 1978. Unpublished manuscript, Educational Testing Service; Princeton, New Jersey, 1978.

Echternacht, G. The choice of covariates. Unpublished manuscript. ETS: Princeton, New Jersey, 1978.

Echternacht, G. The comparabioity of different methodologies for ESEA Title I Evaluation. Paper presented at the Annual Meeting at the American Psychological Association, New York, 1979.

Echternacht, G. Model C is feasible for Title I evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, Mass., April 10, 1980.

Echternacht, G. and Swinton, S. Getting straight: Everything you always wanted to know about the Title I Regression Model and curvilinearity. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, 1979.

Educational Consulting Services, Title I Compensatory Education, Document C-1. Unpublished advertisement, Trenton, N.J., 1980.

Eidenberg, E. and Morey, R.D. An Act of Congress: The Legislative Process and the Making of Educational Policy. W.W. Norton & Company: New York, 1969.

Fleming, P.D. and White, J.W. Comparison of several schemes to estimate the treatment effect under Model C of the ESEA Title I Evaluation System. Paper presented at the annual meeting of the American Educational Research Association in Boston, Massachusetts, April 7-11, 1980.

Florida Department of Education. Title I Evaluation Report, Tallahassee, Fla., 1978.

Florida Department of Education. Title I Evaluation Report, Tallahassee, Fla., 1979.

Gilbert, J.P., Light, R.J., and Mosteller, F. Assessing social innovations: An empirical base for policy. In C.A. Bennett and A.A. Lumsdaine (Eds.), Evaluation and Experiment. New York: Academic Press, 1975.

Glass, G.V. Memorandum to technical assistance center project directors, March 8, 1978.

Goldberger, A.S. Selection bias in evaluating treatment effects: Some formal illustrations. Discussion Papers, #123-72. Madison: Institute for Research on Poverty, University of Wisconsin, 1972.

Gujarti, D. Use of dummy variables in testing for equality between sets of regression coefficients. American Statistician, 1970, 24, 50-52.

Hansen, J.B. Problems and suggested procedures related to the regression model for Title I evaluation (Model C). Paper presented at the Technical Assistance Center Director's Conference, Atlanta, Georgia, January 19-21, 1977.

Hill, R. Personal Communication, 1979.

Hocking, R.R. The analysis and selection of variables in linear regression. Biometrics, 32, March, 1976, 1-49.

Horwitz, S., Long, J.V. and Pellegrini, A.D. An empirical investigation of the ESEA Title I Evaluation System: A comparison of Models A and C. Paper presented at the American Educational Research Association Meeting, San Francisco, Calif1, April, 1979.

House, G.D., A comparison of Title I achievement results obtained under USOE Models A1, C1 and a mixed model. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, Calif., April 12, 1979.

Humphreys, L.G. Investigations of the simplex. Psychometrika, 1960, 25, 4, 313-323.

Johnson, R.T. and Thomas, W.P. User experiences in implementing the RMC Title I Evaluation Models. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, Calif., April 7-13, 1979.

Linn, R.L. Measurment of Change. Presented at the Second Annual Johns Hopkins University National Symposium on Educational Research entitled "Educational Evaluation Methodology: The State of the Art," Washington, D.C., November 2, 1979.

Linn, R.L. Validity of inferences based on the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, 2, 1979. 23-32.

Lloyd, E. Personal Communication, 1979.

Louisiana State Department of Education, Analysis of the application of ESEA, Title I, Model C. Unpublished manuscript. Baton Rouge, Louisiana, 1980.

Magdison, J. Towards a causal model approach for adjusting for pre-existing differences in the non-equivalent control group situation: A general alternative to ANCOVA. Evaluation Quarterly, 1977, 1, 399-420.

McNeil, K. NTS/TAC Responses to Model C questions raised at Region III Model C workshop on October 21, 1977. Unpublished memorandum, 1977.

McNeil, K. and Findlay, E.A. Evaluating Title I Early Childhood Programs: Problems, the Applicability of Model C, and several evaluation plans. Unpublished draft manuscript, NTS Research Corporation, July, 1979.

Murray, M. Models A and C: Theoretical and practical concerns. Paper presented at the Florida Educational Research Association Convention, Daytona Beach, Fla., January 21, 1978.

Murray, S. An analysis of regression effects and the equipercentile growth assumption in the norm referenced evaluation model. Paper presented at the Washington Educational Research Association Annual Meeting, Seattle, 1978.

National Center for Education Statistics (NCES). Quick Survey on Title I Evaluation Models, 1979.

National Institute of Education, Administration of Compensatory Education. Washington, D.C.: NIE, 1977.

Office of Management and Budget (OMB). 1979 Catalog of Federal Domestic Assistance. U.S. Government Printing Office, Washington, D.C., 1979.

Parkes, J. Personal Communication, 1980.

Providence School Department, 1978-1979 ESEA Title I Evaluation, Providence, Rhode Island, July, 1979.

Reichardt, C.S. The design and analysis of the non-equivalent group quasi-experiment. Unpublished Doctoral Dissertation, 1979.

Riecken, H.W., Boruch, R.F., Campbell, D.T., Caplan, N., Glennan, T.K., Pratt, J.W., Rees, A. and Williams, W. Social Experimentation: A Method for Planning and Evaluating Social Intervention. New York: Academic Press, 1974.

RMC Research Corporation, Out-of-Level Testing. Unpublished collection of materials, Portsmouth, N.H., 1979.

Roberts, A.O.H. Regression to the mean and the regression effect bias. Unpublished Draft Manuscript, RMC Research Corporation, Mountain View, Calif., February, 1980.

Rubin, D. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics. 2, 1, 1-26, 1977.

Sacks, J. and Ylvisaker, D. Linear estimation for approximately linear models. Discussion paper Number 9, Center for Statistics and Probability, Northwestern University, October, 1976.

Spiegelman, C.H. Ph.D. Thesis, Northwestern University, 1976.

Spiegelman, C.H. A technique for analyzing a pretest-posttest nonrandomized field experiment. Florida State University, Statistics Report M435, 1977.

Spiegelman, C.H., Estimating the effect of a large scale pretest posttest social program. Proceedings of the Social Statistics Section, American Statistical Association, 370-373, 1979.

Stewart, B.L. Investigating the technical adequacy of Model C in Title I evaluation. Paper presented at the annual meeting of the American Educational Research Association in Boston, Massachusetts, April 7-11, 1980.

Stewart, B.L. The regression model in Title I evaluation. Unpublished Draft Manuscript. RMC Research Corporation, Mountain View, Calif., April, 1980b.

Strand, T. Who said what about Title I evaluation at the AERA/NCME Annual Meeting, San Francisco, April, 1979: Summaries of selected papers. Unpublished manuscript. Educational Testing Service, Evanston, Ill., August 1, 1979.

Sween, J.A. The experimental regression design: An inquiry into the feasibility of nonrandom treatment allocation. Ph.D. Dissertation, Northwestern University, 1971.

Sween, J.A. Regression discontinuity: Statistical tests of significance when units are allocated to treatments on the basis of quantitative eligibility. Unpublished draft, April, 1977.

Talmadge, G.K. Cautions to evaluators. In M.J. Wargo and D.R. Green (Eds.) Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. McGraw-Hill, New York, 1976.

Talmadge, G.K. and Horst, D.P. A procedural guide for validating achievement gains in educational projects. Number 2 in a series of Monographs on Evaluation in Education, U.S. Department of Health, Education and Welfare, Washington, D.C.; 1976.

Talmadge, G.K. and Wood, C.T. User's Guide: ESEA Title I Evaluation and Reporting System. RMC Research Corporation: Mountain View, California, 1978.

Thistlethwaite, D.L. and Campbell, D.T. Regression discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 51, 309-317, 1960.

Trochim, W. An introduction to the analysis of the regression-discontinuity design. Unpublished Manuscript, Northwestern University, 1979.

Trochim, W. The relative assignment variable approach to selection bias in pretest-posttest group designs. Unpublished Manuscript, Northwestern University, 1980.

Trochim, W. Locating data for secondary analysis, in Boruch, R.F., Wortman, P.M. and Cordray, D.S. (Eds.) Secondary Analysis in Applied Social Research. San Francisco: Jossey-Bass, in press.

U.S. Department of Commerce, Report on Statistics for Allocation of Funds: Statistical Policy Working Paper 1, Prepared by the Subcommittee on Statistics for Allocation of Funds, Federal Committee on Statistical Methodology, U.S. Government Printing Office, March, 1978.

U.S. Office of Education (USOE) Policy Manual: Subpart F -- Evaluation. Unpublished draft manuscript, October 20, 1979.

U.S. Office of Education. ESEA Title I Annual Report, 1979.

Visco, R. Personal communication, 1980.

Wick, J.W. Title I Elementary and Secondary Education Act: Formation, Function and Purposes, 1965-1978. Unpublished manuscript, City of Chicago, Department of Education, August, 1978.

Winbush, N. Personal Communication, 1980.

Wood, C. Personal Communication, 1979.

Figure 1.
Hypothetical Data for Regression-Discontinuity in Compensatory Education

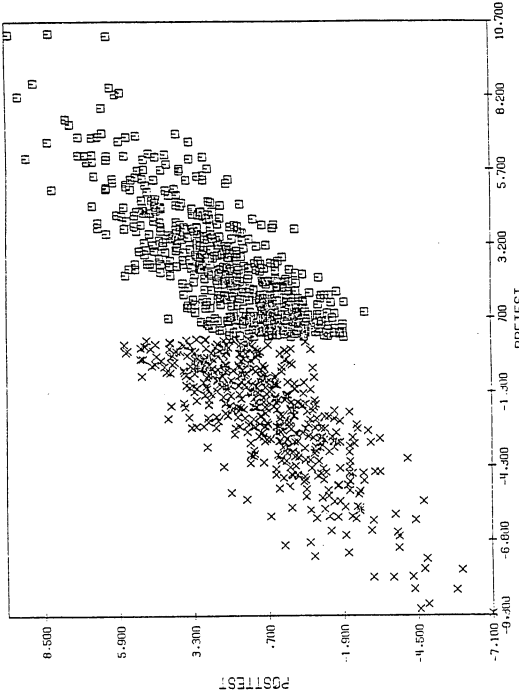


Figure 2.
Regression Lines for Hypothetical Regression-Discontinuity Data in Figure 1

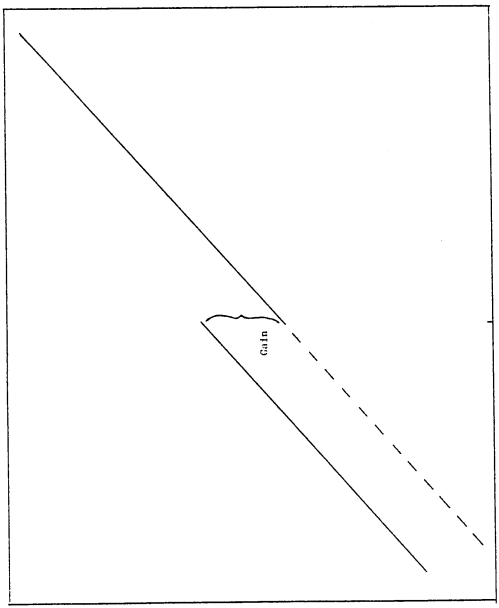


Figure 3.

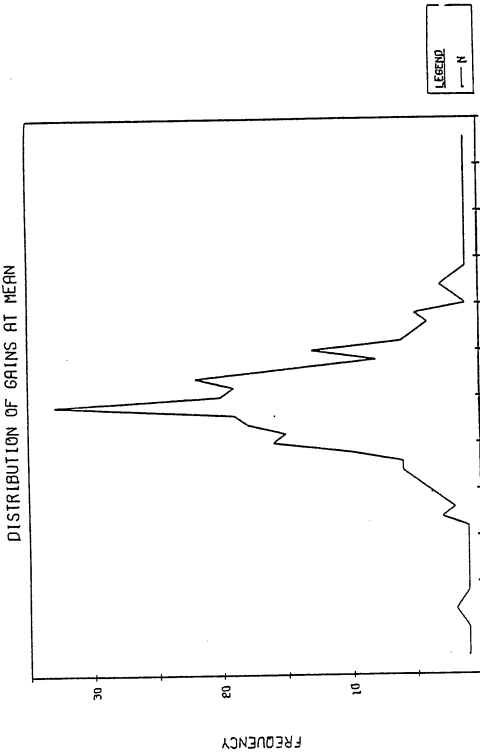


Figure 4.

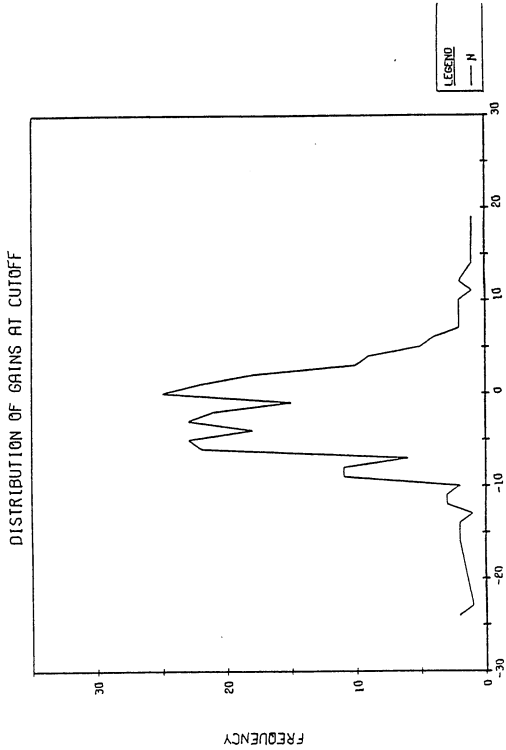


Figure 5.

Regression Artifact with Two Variables

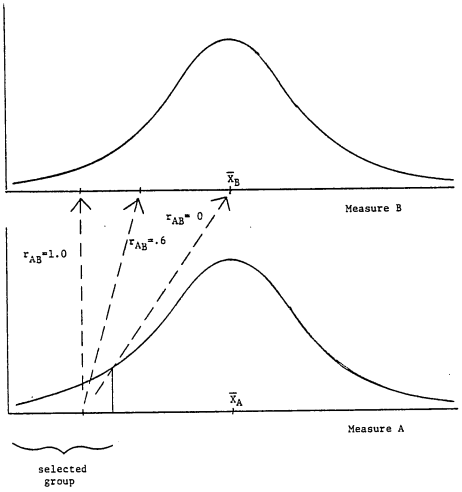


Figure 6.

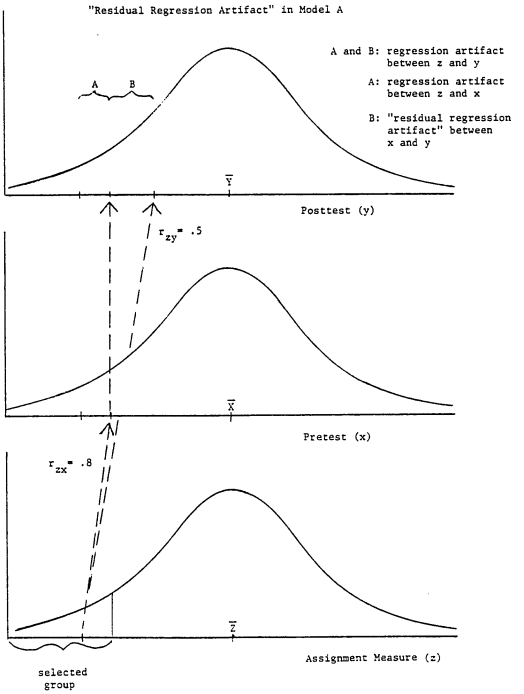


Figure 7.

Effect of Pretest Truncation of Bivariate Distribution on Estimates of Slope

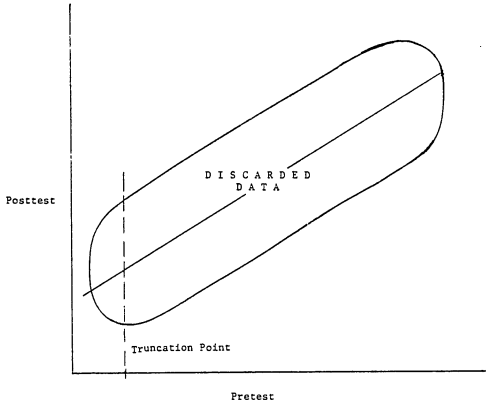


Figure 8.

Regression-Discontinuity when the Pretest and Posttest are Uncorrelated

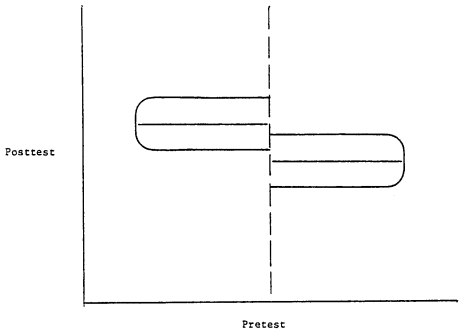


Figure 9.

The Effect of Challenges on
Within-Group Slopes

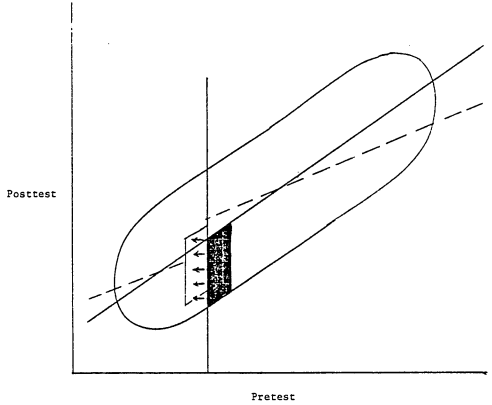


Figure 10.
Hypothetical Pretest-Posttest Relationships for Floor or Ceiling Effect on Either Measure

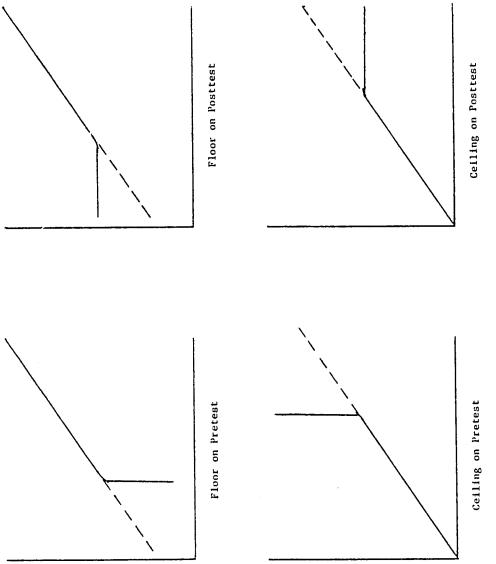


Figure 11.

Hypothetical Bivariate Distribution with Chance Level Effect

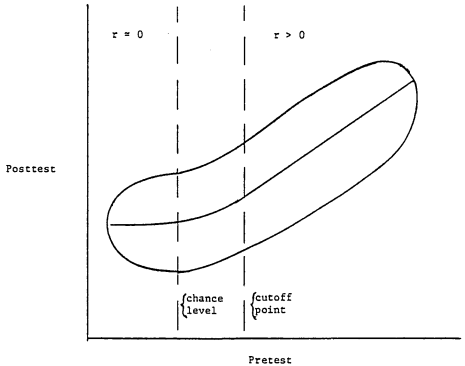


Figure 12.

Effect of Excluding Title I Repeaters

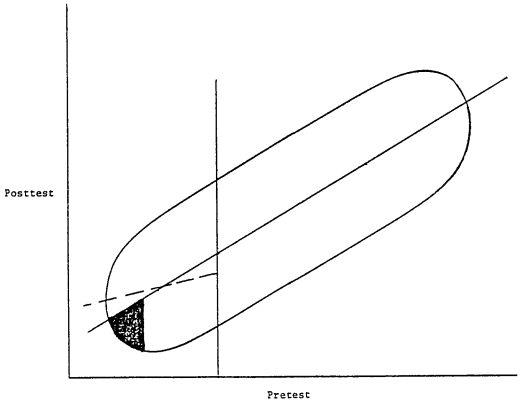


Figure 13.

Pseudo-Effects Resulting from Fitting a
True Quadratic Function with a Linear Model

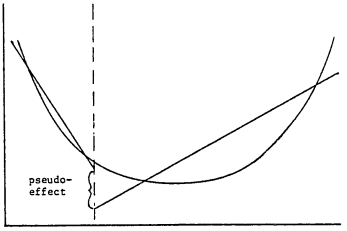
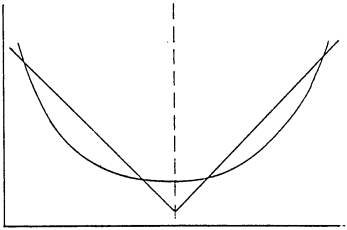


Figure 14.
Symmetry of Function around Cutoff
Results in no Pseudo-Effect when Underfitting



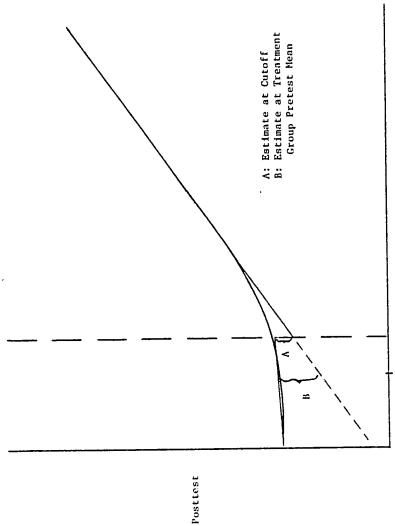


Figure 15.
Pseudo-Effects with Title I Model C Analysis when the True Functional Relationship is Non-Linear

Figure 15a

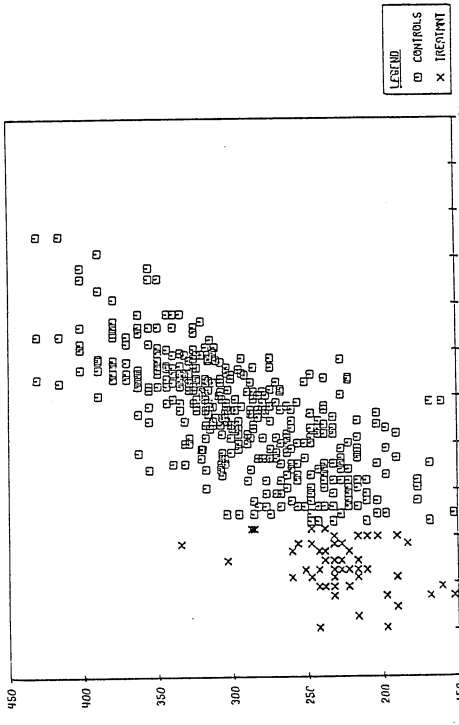


Figure 15b

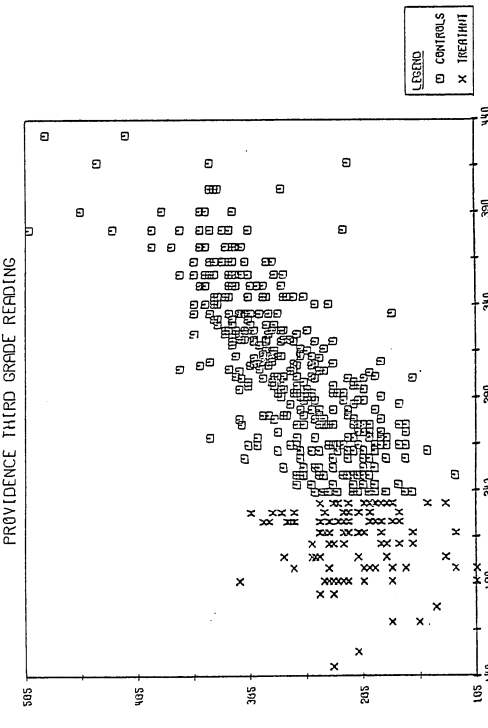


Figure 15c

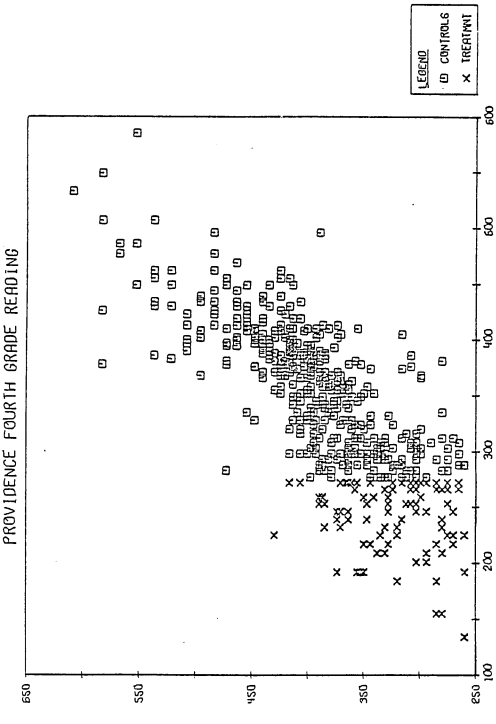


Figure 15d

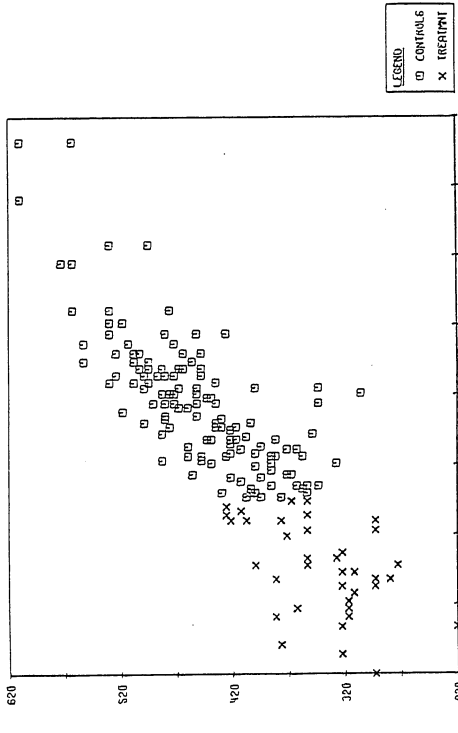


Figure 15e

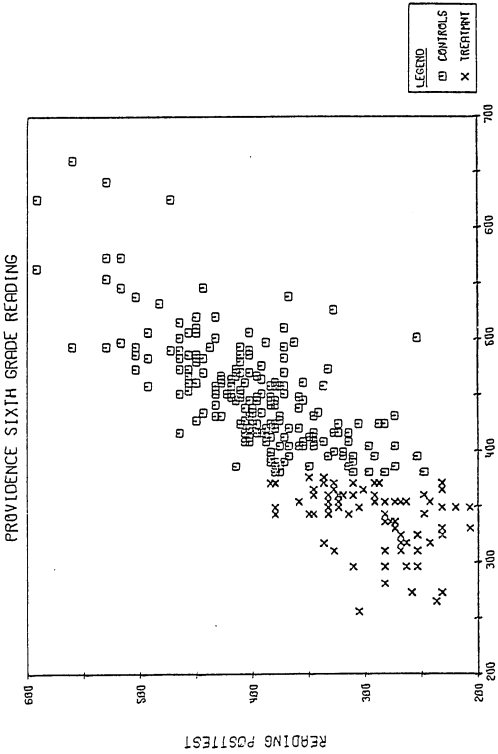
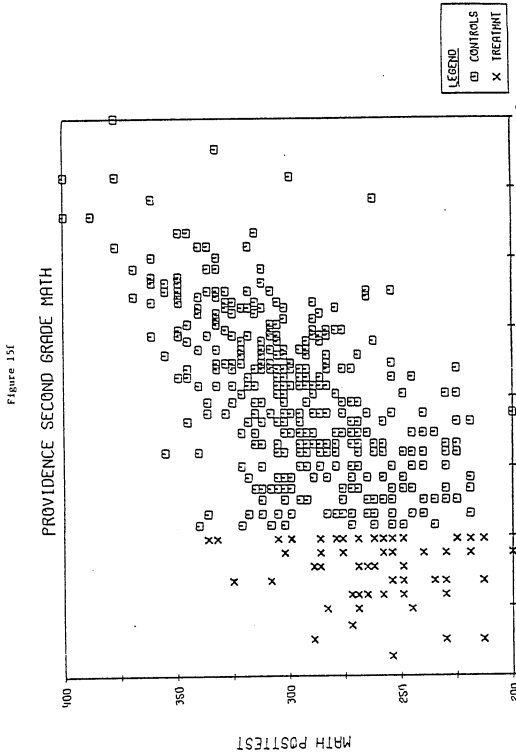


Figure 15f



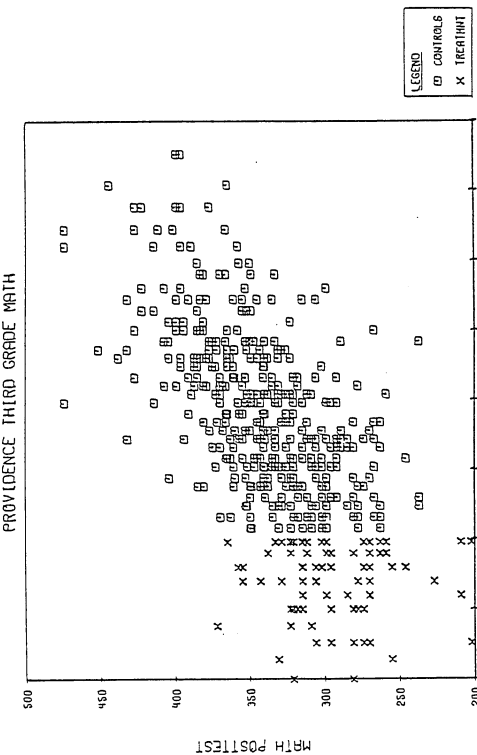


Figure 15g

Figure 15h

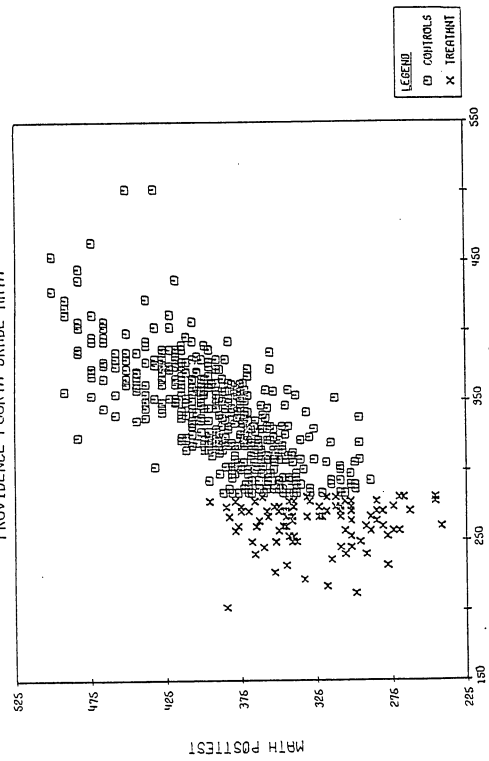


Figure 16a

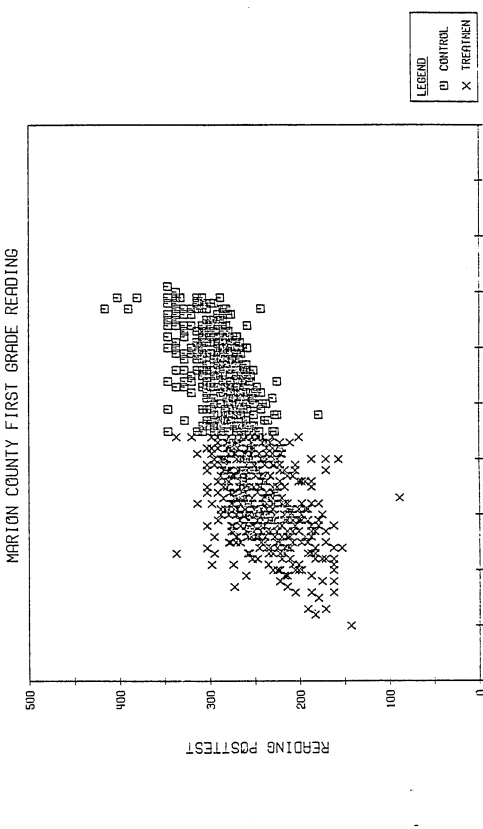


Figure 16b

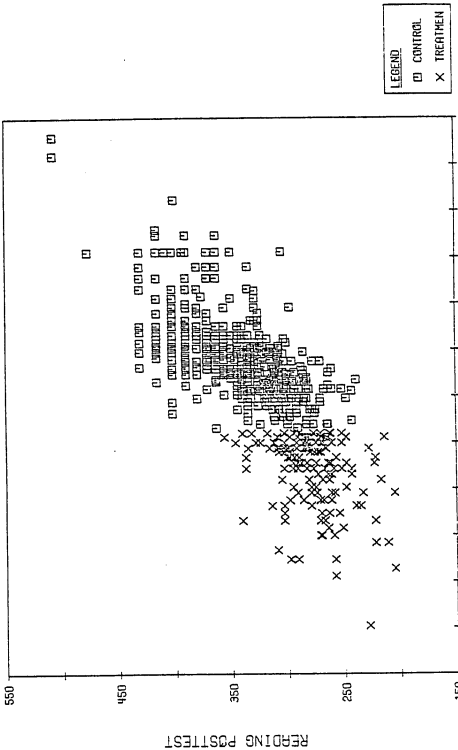


Figure 16c

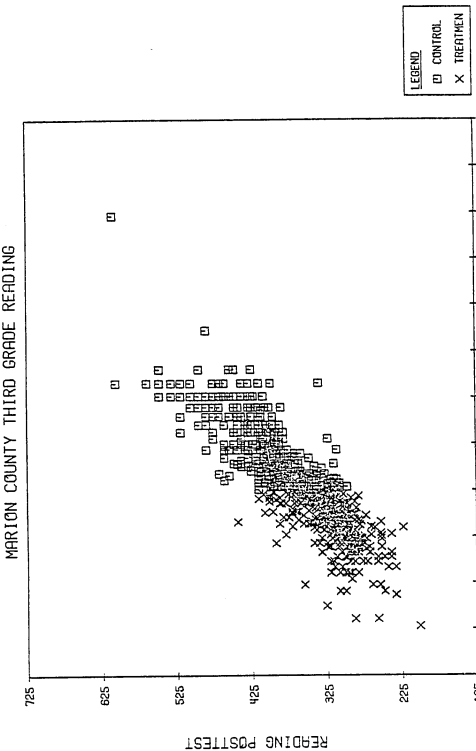


Figure 16d

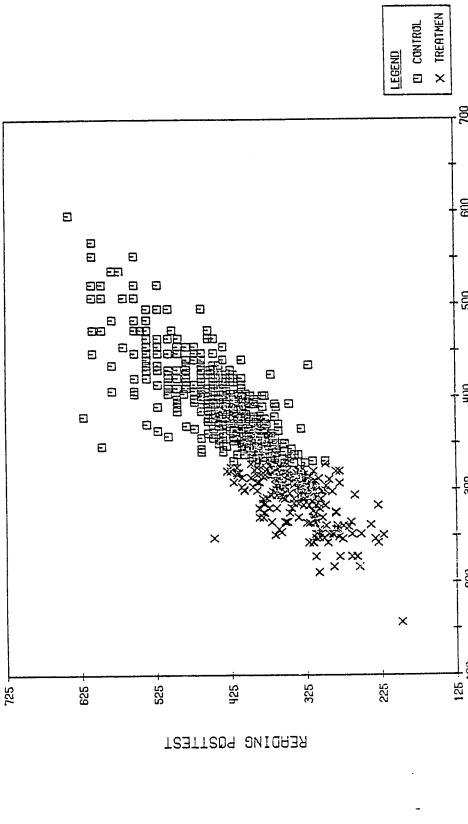


Figure 16e

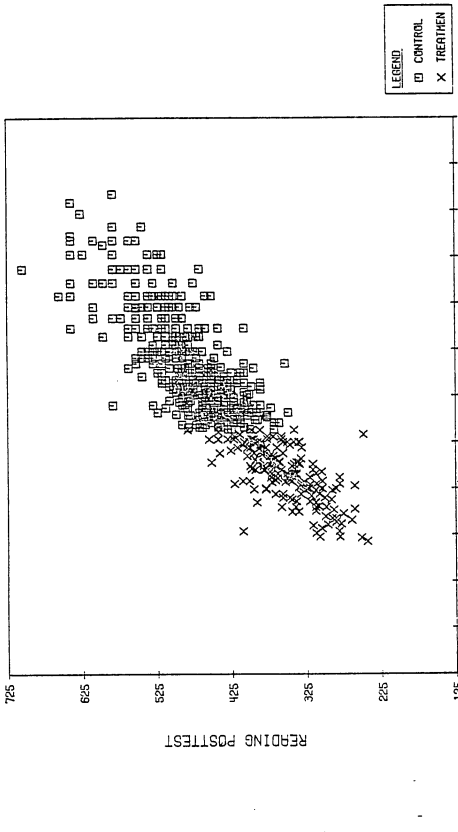


Figure 16f

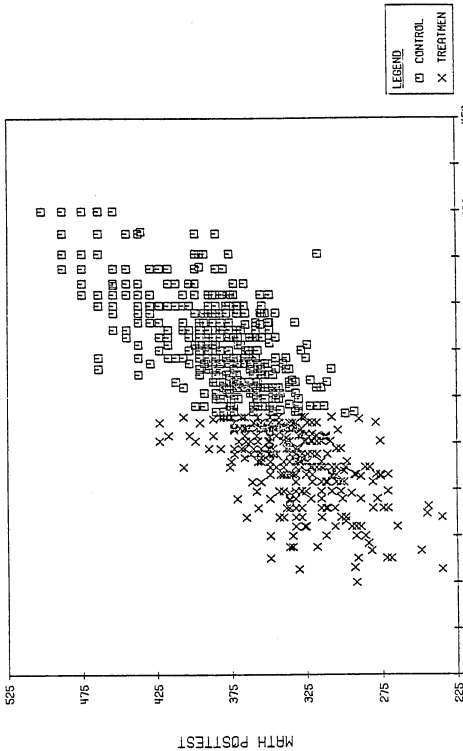


Figure 16h

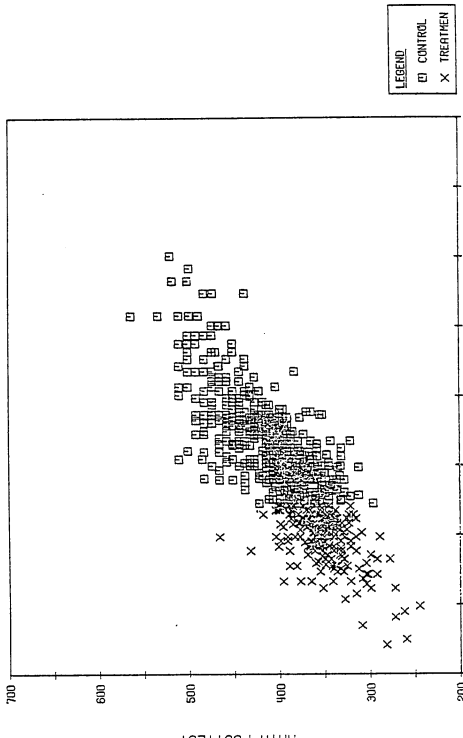


Figure 16h

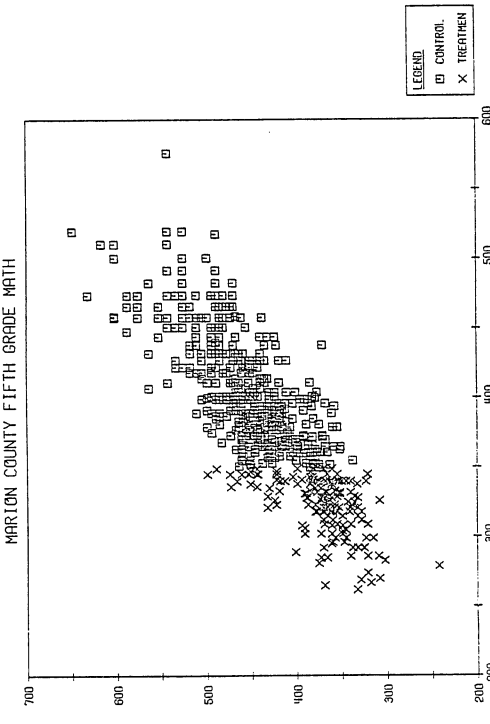


Figure 17

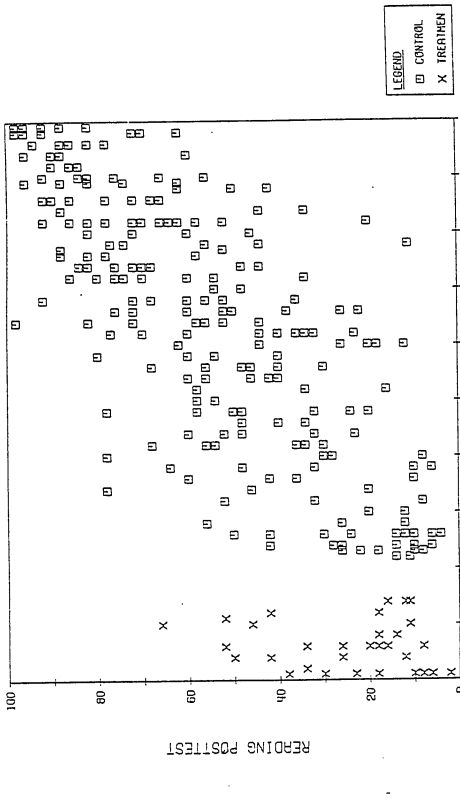


Figure 18a

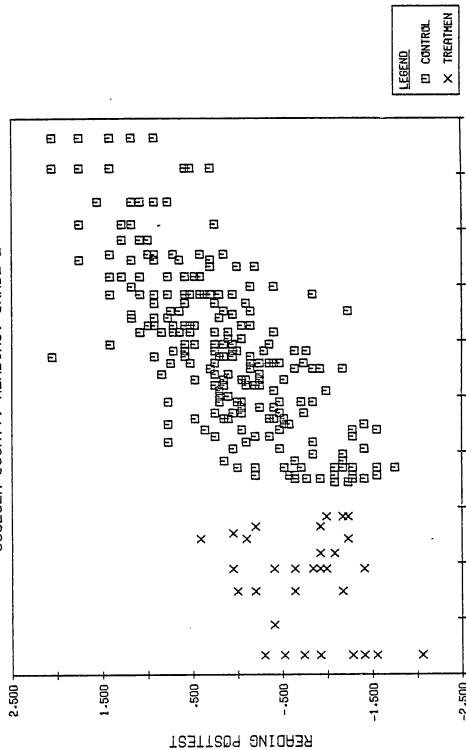


Figure 18b

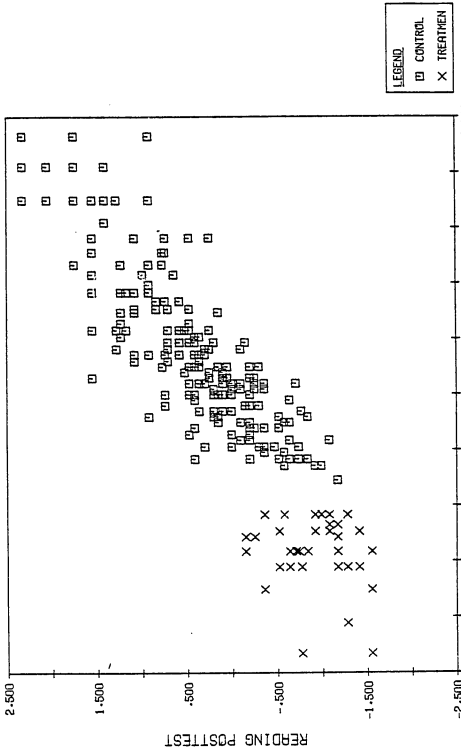


Figure 18c

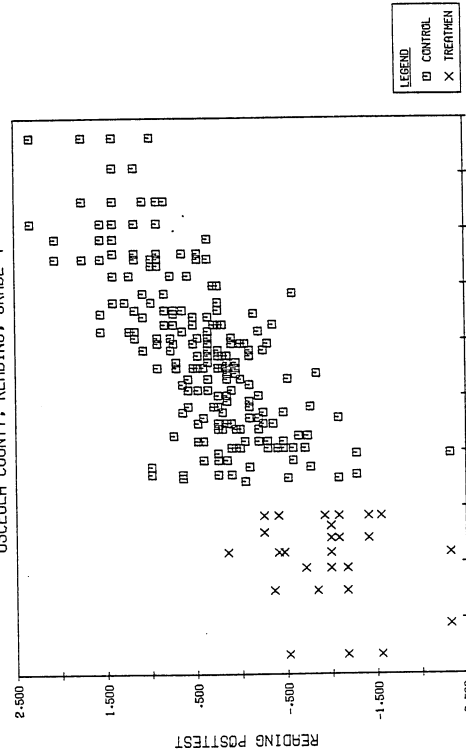
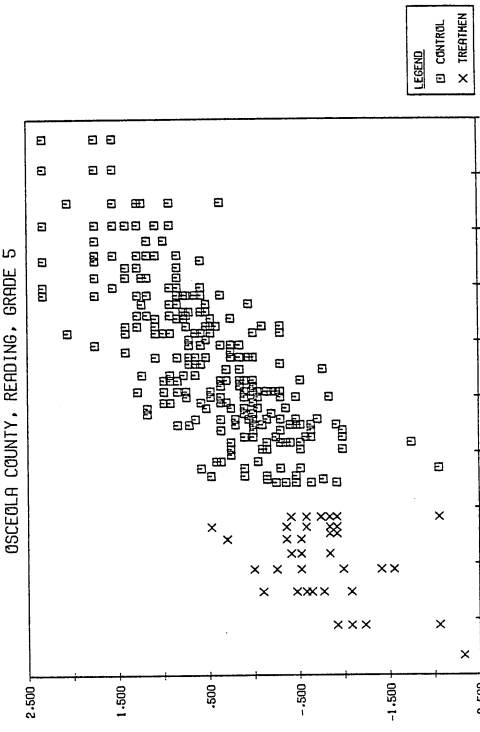


Figure 18d



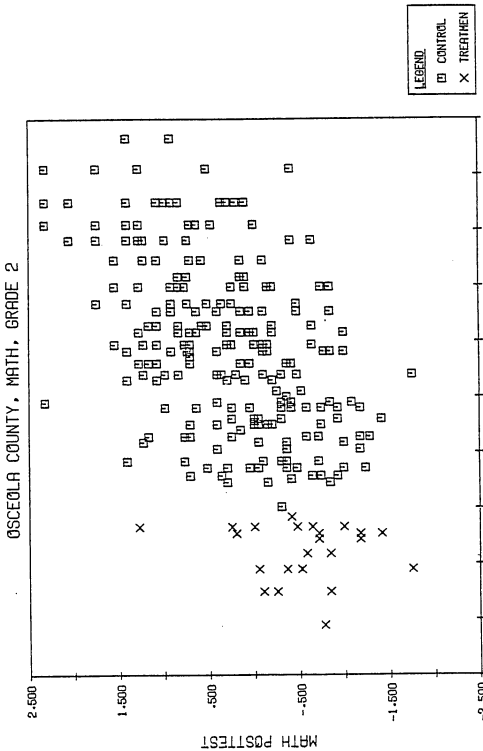


Figure 18e

Figure 16f

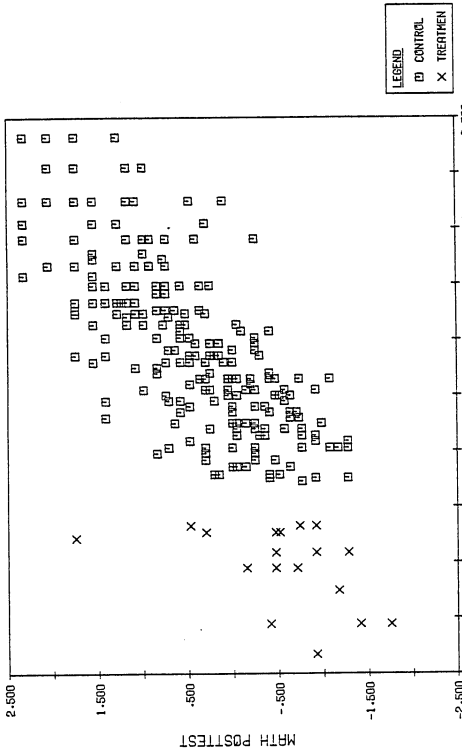


Figure 10g

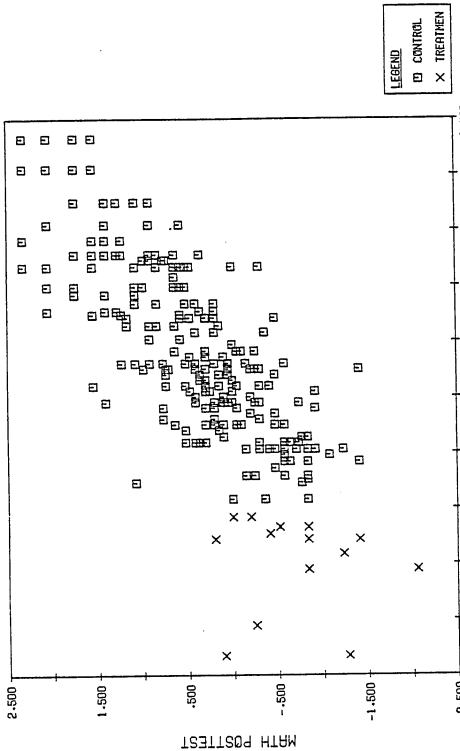


Figure 18h

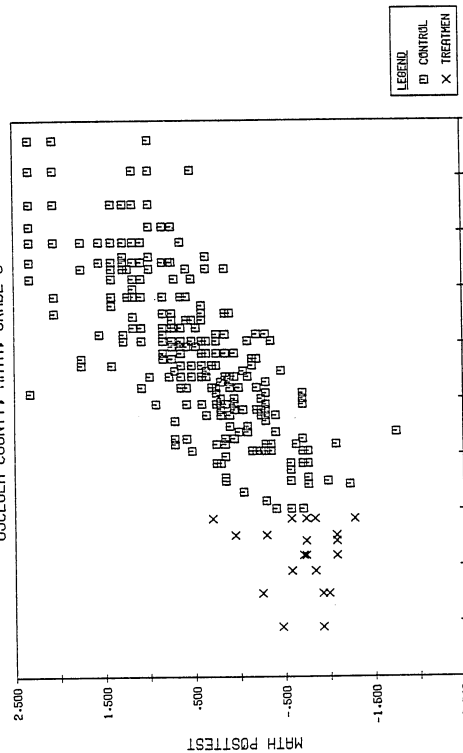


Figure 19a

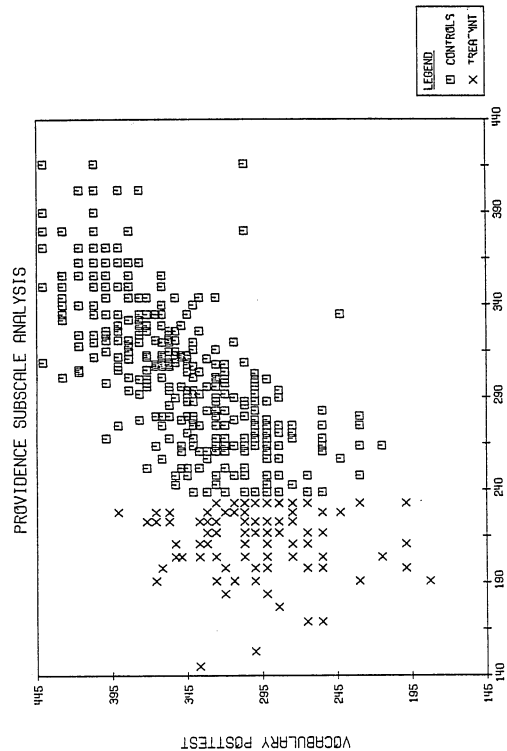


Figure 19b

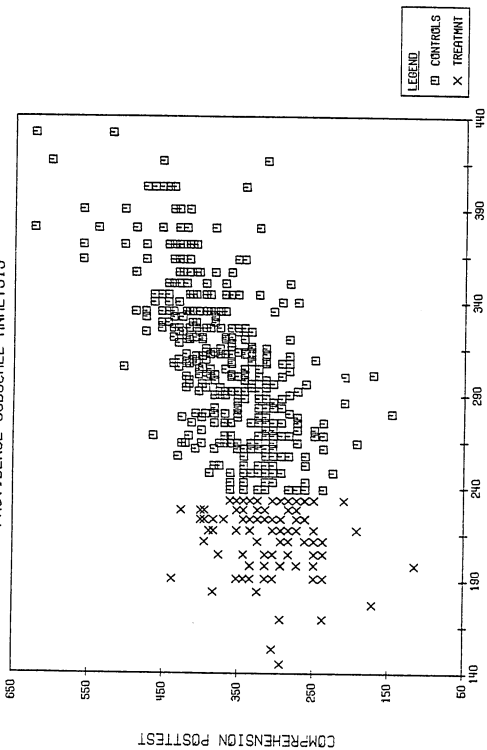


Figure 20a

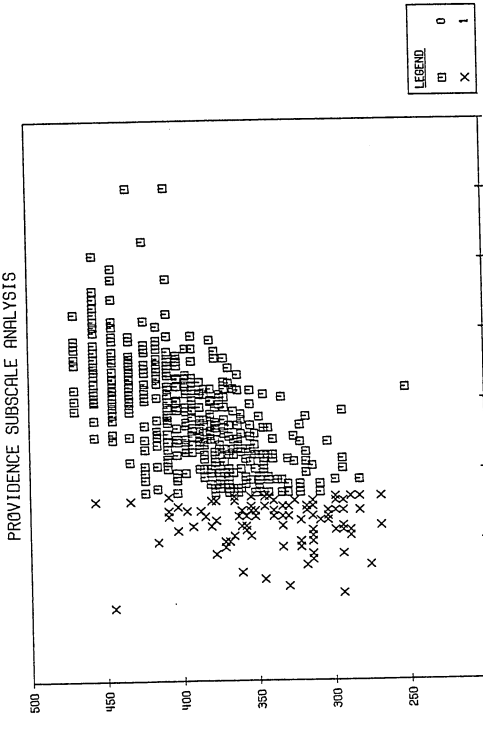


Figure 20b

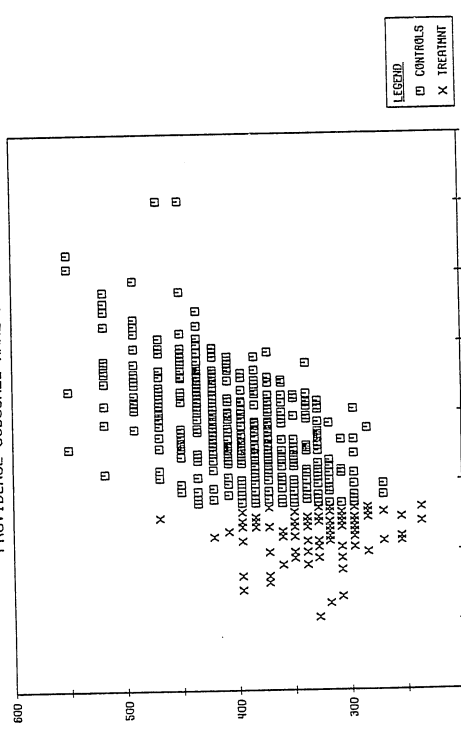


Figure 20c

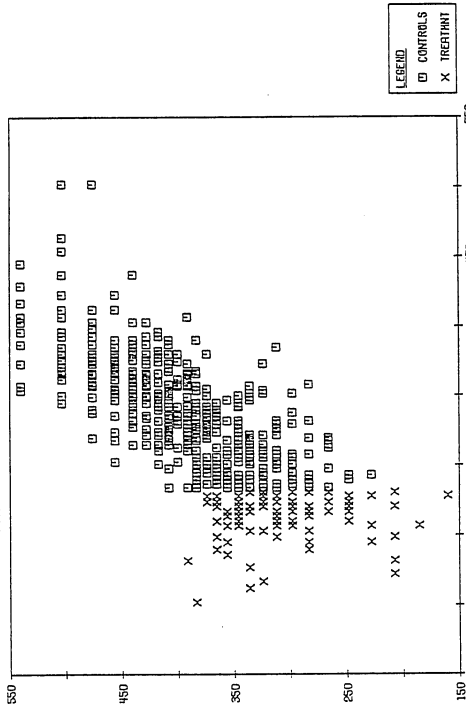


Table 1

Estimates of Slope and Intercept Under Varying Restrictions of Pretest Range

Run 1: $\sigma_t = 3$, $\sigma e_x = 1$, $\sigma e_y = 1$, $\rho = .9$

Pretest Range Restricted to Scores Below

	-2	-1	0	+1	+2
$\hat{\beta}_1$	.91	.89	.87	.87	.86
$\hat{\beta}_o$	.03	.04	.06	.08	.09

Run 2: $\sigma_t = 3$, $\sigma e_x = 1$, $\sigma e_y = 1$, $\rho = .9$

Pretest Range Restricted to Scores Below

	-2	-1	0	+1	+2
$\hat{\beta}_1$	.90	.89	.89	.87	.86
$\hat{\beta}_o$	-.06	-.02	.04	.02	.03

Run 3: $\sigma_t = 3$, $\sigma e_x = 2$, $\sigma e_x = 2$, $\rho = .69$

Pretest Range Restricted to Scores Below

	-2	-1	0	+1	+2
$\hat{\beta}_1$	.70	.72	.75	.68	.70
$\hat{\beta}_o$	.01	.02	-.01	.09	.09

Run 4: $\sigma_t = 3$, $\sigma e_x = 2$, $\sigma e_y = 2$, $\rho = .69$

Pretest Range Restricted to Scores Below

	-2	-1	0	+1	+2
$\hat{\beta}_1$	.72	.75	.76	.80	.80
$\hat{\beta}_o$	-.02	.01	-.02	-.06	-.08

Table 1
(Continued)

Run 5: $\sigma_t = 3$, $\sigma e_x = 3$, $\sigma e_y = 3$, $\rho = .5$

Pretest Range Restricted to Scores Below

	-2	-1	0	+1	+2
$\hat{\beta}_1$	.49	.43	.46	.41	.45
$\hat{\beta}_o$	.11	.05	.08	.11	.09

Run 6: $\sigma_t = 3$, $\sigma e_x = 3$, $\sigma e_y = 3$, $\rho = .5$

Pretest Range Restricted to Scores Below

	-2	-1	0	+1	+2
$\hat{\beta}_1$	.50	.53	.49	.50	.48
$\hat{\beta}_o$	.06	.07	.05	.06	.07

Table 2

Statistics Under Various Restrictions of Pretest Range
($\sigma_t = 3$, $\sigma_{e_x} = 1$, $\sigma_{e_y} = 1$, $\rho = .9$)

Pretest Range Restricted to Scores Below	S_y	S_x	S_{xy}	n	r	S_y/S_x	$\hat{\beta}$
All Data	3.2129	3.2081	9.275	1000	.89992	1.001	.901
+2	2.5909	2.3185	5.081	754	.84590	1.12	.945
+1	2.4572	2.1274	4.312	634	.82484	1.16	.953
0	2.2584	1.9289	3.492	506	.80172	1.17	.938
-1	2.0993	1.7465	2.816	384	.76814	1.20	.923
-2	1.9485	1.5739	2.199	278	.71719	1.24	.888

Table 3

Parameter Values for Illustrative Simulations of Under and Overfitting the True Function

	σ_t	σ_{e_x}	σ_{e_y}	x_0
Linear				
g = 0	3	1	1	-1.50
g = 5	3	1	1	-1.50
Quadratic				
g = 0	2	1	1	-1.25
g = 5	2	1	1	-1.25
Cubic				
g = 0	1	1	1	-1.00
g = 5	1	1	1	-1.00

Table 4

Models Estimated at Each Step in the Illustrative Computer Simulation Analyses

Step 1: $y = \beta_0 + \beta_1 x' + \beta_2 z + e$

Step 2: $y = \beta_0 + \beta_1 x' + \beta_2 z + \beta_3 x'z + e$

Step 3: $y = \beta_0 + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^{-2} + \beta_5 x'^{-2} z + e$

Step 4: $y = \beta_0 + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^{-2} + \beta_5 x'^{-2} z + \beta_6 x'^{-3} + \beta_7 x'^{-3} z + e$

Step 5: $y = \beta_0 + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^{-2} + \beta_5 x'^{-2} z + \beta_6 x'^{-3} + \beta_7 x'^{-3} z + \beta_8 x'^{-4} + \beta_9 x'^{-4} z + e$

Step 6: $y = \beta_0 + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^{-2} + \beta_5 x'^{-2} z + \beta_6 x'^{-3} + \beta_7 x'^{-3} z + \beta_8 x'^{-4} + \beta_9 x'^{-4} z + \beta_{10} x'^{-5} + \beta_{11} x'^{-5} z + e$

Step 7: $y = \beta_0 + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^{-2} + \beta_5 x'^{-2} z + \beta_6 x'^{-3} + \beta_7 x'^{-3} z + \beta_8 x'^{-4} + \beta_9 x'^{-4} z + \beta_{10} x'^{-5} + \beta_{11} x'^{-5} z + \beta_{12} x'^{-6} + \beta_{13} x'^{-6} z + e$

Step 8: $y = \beta_0 + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^{-2} + \beta_5 x'^{-2} z + \beta_6 x'^{-3} + \beta_7 x'^{-3} z + \beta_8 x'^{-4} + \beta_9 x'^{-4} z + \beta_{10} x'^{-5} + \beta_{11} x'^{-5} z + \beta_{12} x'^{-6} + \beta_{13} x'^{-6} z + \beta_{14} x'^{-7} + \beta_{15} x'^{-7} z + e$

Table 5

Estimates of Gain and Standard Errors for
True Linear Function, Gain = 0

STEP 1	.15937244	.14345886
STEP 2	.19229522	.15193996
STEP 3	.22355321	.20089430
STEP 4	-.41882406E-01	.25799966
STEP 5	-.30271201	.31388807
STEP 6	.10967221	.36537469
STEP 7	-.69973307E-01	.42478725
STEP 8	.22321649	.48897469

STEP 1	.25857282	.14266355
STEP 2	.24864566	.15248773
STEP 3	.19505811	.20645407
STEP 4	.48702403	.26297753
STEP 5	.48164727	.32057747
STEP 6	.55852917	.37669312
STEP 7	.65555652	.42872512
STEP 8	.28449175	.48097897

STEP 1	.29290099	.14602812
STEP 2	.36693815	.15550268
STEP 3	.46296535	.21523065
STEP 4	.38993168	.27474425
STEP 5	.38010305	.33433035
STEP 6	.13162441	.39383123
STEP 7	.18152651	.45549697
STEP 8	.17131259	.51979992

STEP 1	.21425867	.14340927
STEP 2	.24331347	.15385772
STEP 3	.32842042	.20992387
STEP 4	.37729627	.26367844
STEP 5	.31463261	.32411148
STEP 6	.95687652E-02	.39226485
STEP 7	-.36087552	.46121415
STEP 8	-.34907027	.53931705

STEP 1	-.24969710	.14418639
STEP 2	-.24574819	.15016634
STEP 3	-.25529789	.20563240
STEP 4	-.59304006	.26454416
STEP 5	-.56063583	.32300514
STEP 6	-.46089338	.39035426
STEP 7	-.52359883	.46061831
STEP 8	-.86398485	.53412428

STEP 1	.17563911	.14770736
STEP 2	.11441807	.15888857
STEP 3	-.67893737E-01	.21419577
STEP 4	-.31199813	.26701349
STEP 5	-.14538137	.32330371
STEP 6	-.27477265	.37132306
STEP 7	-.46920913	.41814009
STEP 8	-.59008173	.46294482

STEP 1	-.12589623	.14945837
STEP 2	-.16130224	.15684676
STEP 3	-.25684285E-01	.21360386
STEP 4	-.14764004	.27358858
STEP 5	-.47173931	.33139824
STEP 6	-.30624377	.38833649
STEP 7	-.32352733	.44970592
STEP 8	-.45842853	.50413727

STEP 1	.10181058	.14111817
STEP 2	.58991229E-01	.15337498
STEP 3	-.25855784E-01	.21279103
STEP 4	.79310461E-01	.27082366
STEP 5	-.45209554E-01	.32949255
STEP 6	-.47392462	.39329624
STEP 7	.73390221E-01	.46396651
STEP 8	-.34888288E-01	.53958060

STEP 1	.24164848	.14602472
STEP 2	.23125895	.15413720
STEP 3	.28118026	.21327688
STEP 4	.13993972	.27093917
STEP 5	-.53044037E-01	.33452669
STEP 6	.13713733	.39617958
STEP 7	.45484653E-01	.45481368
STEP 8	.60589461E-01	.51324267

STEP 1	.11448574	.14195295
STEP 2	.50791180E-01	.14855164
STEP 3	.70463466E-01	.19757013
STEP 4	-.30962969	.25084156
STEP 5	-.24674413	.30901359
STEP 6	.20705203	.36665798
STEP 7	.19324432	.42825834
STEP 8	-.21583938E-01	.49324111

Table 5
(continued)

Estimates of Gain and Standard Errors for
True Linear Function, Gain = 5

STEP 1	4.8861732	.14367704	STEP 1	5.1507535	.14196989
STEP 2	4.8484194	.15279109	STEP 2	5.1716720	.14962921
STEP 3	4.6112907	.21078632	STEP 3	5.0650711	.20419159
STEP 4	4.6617914	.26890971	STEP 4	4.9896202	.25633902
STEP 5	5.0168601	.33152204	STEP 5	4.9199097	.31238472
STEP 6	5.1534195	.40320846	STEP 6	4.7115319	.37097341
STEP 7	5.1519125	.47828929	STEP 7	4.7247456	.43020383
STEP 8	4.9943011	.54964928	STEP 8	4.9155290	.48974445
STEP 1	5.1432340	.14640321	STEP 1	5.2231853	.14041259
STEP 2	5.1294626	.15459488	STEP 2	5.2591464	.14879063
STEP 3	5.0056153	.20995200	STEP 3	5.0395925	.19795973
STEP 4	5.2887233	.26319560	STEP 4	5.0831105	.24891769
STEP 5	5.3141909	.31518349	STEP 5	5.0814644	.30171714
STEP 6	5.4101971	.37286783	STEP 6	4.8913775	.35882270
STEP 7	5.6541111	.43266562	STEP 7	4.9802425	.42000503
STEP 8	5.5446662	.49435268	STEP 8	5.0232382	.48214740
STEP 1	5.1556522	.14914881	STEP 1	4.9893151	.14221486
STEP 2	5.2038289	.15327098	STEP 2	5.0621880	.15351361
STEP 3	5.2148368	.21138665	STEP 3	4.8183819	.20903764
STEP 4	5.3889228	.27613390	STEP 4	5.0419820	.26632493
STEP 5	5.6357885	.34265466	STEP 5	5.0179331	.33045358
STEP 6	6.0203501	.41297714	STEP 6	4.7571685	.39279991
STEP 7	6.2293830	.47709414	STEP 7	4.7712478	.45589399
STEP 8	6.2145560	.54626779	STEP 8	4.6911176	.52978485
STEP 1	5.0281147	.15317723	STEP 1	4.7930892	.14972988
STEP 2	5.0203469	.16060395	STEP 2	4.7875390	.15935508
STEP 3	5.0481433	.21745551	STEP 3	4.5505612	.22034742
STEP 4	4.9664353	.28232858	STEP 4	4.3235050	.28070376
STEP 5	4.8115082	.34510463	STEP 5	4.3359037	.34346513
STEP 6	4.9332259	.40427629	STEP 6	3.9910319	.40710947
STEP 7	4.4740946	.45847812	STEP 7	4.0205751	.46929092
STEP 8	4.3156000	.51464726	STEP 8	3.8751492	.52698078
STEP 1	4.7999090	.14323643	STEP 1	5.1973933	.14426121
STEP 2	4.7558906	.15320572	STEP 2	5.1838439	.15340224
STEP 3	4.8835753	.20879684	STEP 3	5.1139154	.20613199
STEP 4	4.8195776	.26796992	STEP 4	5.1357563	.26267716
STEP 5	4.6579485	.32210773	STEP 5	5.2091237	.31553413
STEP 6	4.7427096	.37726831	STEP 6	5.1372058	.37537786
STEP 7	4.4530145	.43318676	STEP 7	5.1904694	.43494604
STEP 8	4.6382841	.48470652	STEP 8	5.5674623	.49137073

Table 6

Simulation Sampling Means and Standard Errors for True Linear Function
($n=10$, $\sigma_t=3$, $\sigma_{e_x}=1$, $\sigma_{e_y}=1$, cutoff=-1.5)

Gain = 0

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
1	.118	.055
2	.110	.061
3	.119	.069
4	.007	.112
5	-.065	.113
6	-.036	.105
7	-.060	.116
8	-.158	.122

Gain = 5

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
1	5.037	.052
2	5.042	.058
3	4.935	.069
4	4.970	.098
5	5.000	.113
6	4.975	.166
7	4.965	.201
8	4.978	.212

Table 7
(continued)

Estimates of Gain and Standard Errors for
True Quadratic Function, Gain = 5

STEP 1	11.243563	.58273880	STEP 1	12.252422	.60112413
STEP 2	6.5052747	.46220462	STEP 2	7.0084934	.45783641
STEP 3	5.7693072	.38942075	STEP 3	4.5508265	.55265092
STEP 4	3.7509826	.74359227	STEP 4	5.3050136	.70305577
STEP 5	5.3203451	.89426191	STEP 5	5.6113061	.86194718
STEP 6	4.0556522	1.0541751	STEP 6	5.6493038	1.0226546
STEP 7	4.1637775	1.2075381	STEP 7	5.3902693	1.1844210
STEP 8	4.1353346	1.3509380	STEP 8	5.5647332	1.3429612
STEP 1	11.530092	.56193980	STEP 1	10.775081	.52652382
STEP 2	6.8855200	.42861742	STEP 2	6.1174795	.41447956
STEP 3	5.3436685	.52573253	STEP 3	5.5305536	.50927471
STEP 4	5.5852189	.67811228	STEP 4	5.5912634	.65450277
STEP 5	5.3338366	.84377630	STEP 5	5.5549942	.80189072
STEP 6	5.0629253	1.0191839	STEP 6	5.0488542	.95771448
STEP 7	6.4841415	1.1924723	STEP 7	5.1175368	1.1208440
STEP 8	4.4815903	1.3484217	STEP 8	5.7081970	1.2866924
STEP 1	12.234677	.58576171	STEP 1	11.971827	.56651622
STEP 2	6.5517058	.48649843	STEP 2	7.2734757	.45740365
STEP 3	5.0021891	.54436858	STEP 3	5.9638111	.54262139
STEP 4	5.1194968	.68823908	STEP 4	4.5448250	.68560281
STEP 5	5.7540766	.83160201	STEP 5	6.2073853	.83391275
STEP 6	5.1492846	.98585433	STEP 6	4.6332774	.96156493
STEP 7	5.9745122	1.1420174	STEP 7	6.1008726	1.0932817
STEP 8	5.3406048	1.2902842	STEP 8	6.1643401	1.2173904
STEP 1	11.845448	.59503759	STEP 1	11.906179	.62869233
STEP 2	7.1429195	.46706092	STEP 2	6.5501419	.48583017
STEP 3	5.2681941	.56459482	STEP 3	4.9338877	.61473095
STEP 4	5.6586245	.71566412	STEP 4	4.8006893	.78940122
STEP 5	5.2562694	.88422130	STEP 5	4.5874209	.95706336
STEP 6	4.2322775	1.0444928	STEP 6	3.8438973	1.1159027
STEP 7	4.5595647	1.2191716	STEP 7	5.3509858	1.2680923
STEP 8	3.0841016	1.3929677	STEP 8	5.8988427	1.4218973
STEP 1	11.981407	.56870634	STEP 1	11.491774	.63599572
STEP 2	7.5407055	.43419260	STEP 2	6.0016634	.48696981
STEP 3	4.4194547	.52862201	STEP 3	3.7249884	.56688432
STEP 4	4.7690199	.66379669	STEP 4	4.5512792	.71513374
STEP 5	4.2085160	.81821353	STEP 5	5.0820513	.86887874
STEP 6	3.4786298	.97768229	STEP 6	2.5246634	1.0262595
STEP 7	5.9234028	1.1220406	STEP 7	5.2329081	1.1922497
STEP 8	3.6767618	1.2690185	STEP 8	3.9403922	1.3783932

Table 7

Estimates of Gain and Standard Errors for
True Quadratic Function, Gain = 0

STEP 1	6.4143674	.56367628	STEP 1	7.1185522	.58359656
STEP 2	1.9025779	.44215887	STEP 2	2.6959386	.48351252
STEP 3	-.12599175	.50204279	STEP 3	-.53594443	.56846496
STEP 4	-.52740613	.64444199	STEP 4	.87487159	.70750847
STEP 5	-1.5149543	.78534294	STEP 5	-.85856944	.86309768
STEP 6	-1.2755742	.93141566	STEP 6	.76240732E-01	1.0255973
STEP 7	-1.8654433	1.0865625	STEP 7	-.53706572	1.1827541
STEP 8	-.41372511E-01	1.2343492	STEP 8	-.77048570	1.3493162
STEP 1	7.0999267	.53934046	STEP 1	6.9255438	.58665315
STEP 2	2.5110231	.45869356	STEP 2	2.0189631	.48148344
STEP 3	-.77662865E-01	.58914866	STEP 3	.13932528	.59277449
STEP 4	-.12387344	.76812243	STEP 4	-1.0197738	.74422478
STEP 5	1.3772557	.94824735	STEP 5	-.35795142	.93342315
STEP 6	-.55191524	1.1404907	STEP 6	-1.0380386	1.0998907
STEP 7	.45160997E-01	1.3457303	STEP 7	-.80064413	1.2735803
STEP 8	1.2825323	1.5546731	STEP 8	-.79530434	1.4530898
STEP 1	7.9004913	.60205207	STEP 1	6.5415920	.59938781
STEP 2	2.1724583	.48365895	STEP 2	1.9206948	.49303909
STEP 3	.32700168	.59470179	STEP 3	-.26806737	.62082449
STEP 4	.44403251	.75382639	STEP 4	.74538150	.79854599
STEP 5	-.69738770	.91989983	STEP 5	1.6144852	.99092069
STEP 6	1.8623757	1.0658054	STEP 6	.65090334E-02	1.1951888
STEP 7	-.18996660	1.2272882	STEP 7	.30636956	1.4070103
STEP 8	2.3105323	1.3803508	STEP 8	.63820788	1.6243337
STEP 1	8.2859865	.59944098	STEP 1	5.1299366	.59475208
STEP 2	2.7503752	.49853366	STEP 2	1.6143104	.44770220
STEP 3	-.36637874	.55535069	STEP 3	-.28268157	.55883871
STEP 4	.51320123	.69733996	STEP 4	-1.1266596	.68896565
STEP 5	-.49532898	.84608056	STEP 5	-.90315206	.83217435
STEP 6	-1.6415920	.99929893	STEP 6	-.48506992E-01	.97514829
STEP 7	-.28025262	1.1556352	STEP 7	.52980609	1.1220245
STEP 8	.38256861	1.3411213	STEP 8	.19100021	1.2731654
STEP 1	8.1419331	.58779880	STEP 1	5.3979921	.60416932
STEP 2	3.5173899	.48267435	STEP 2	.26629454	.45380790
STEP 3	.68940307	.55588689	STEP 3	-.15061762	.55058357
STEP 4	.52586255	.69308884	STEP 4	-.41507468	.69075888
STEP 5	.79699197E-01	.84660829	STEP 5	-.90193548	.84051601
STEP 6	-.77752427E-01	1.0032787	STEP 6	-1.1617883	.97803287
STEP 7	-.20765235	1.1724908	STEP 7	-.85529543	1.1084727
STEP 8	.61572593	1.3350747	STEP 8	-1.1005321	1.2354910

Table 8

Simulation Sampling Means and Standard Errors for True Quadratic Function ($n=10$, $\sigma_t=2$, $\sigma_{e_x}=1$, $\sigma_{e_y}=1$, cutoff=-1.25)

Gain = 0

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
1	6.896	.338
2	2.137	.271
3	-.065	.114
4	-.011	.232
5	-.266	.322
6	-.385	.317
7	-.385	.216
8	.271	.328

Gain = 5

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
1	11.723	.147
2	6.758	.158
3	5.051	.214
4	4.968	.191
5	5.292	.181
6	4.370	.296
7	5.530	.246
8	4.799	.338

Estimates of Gain and Standard Errors for
True Cubic Function, Gain = 0

STEP 1	.42644841	.40854407	STEP 1	1.1749171	.36004070
STEP 2	.49292700	.47173786	STEP 2	2.1446768	.38209532
STEP 3	-.48604957	.64715500	STEP 3	-1.3161143	.50033627
STEP 4	.15323656	.81754423	STEP 4	-.36466745	.63483325
STEP 5	1.0016286	1.0123309	STEP 5	-.43236760	.76902454
STEP 6	.20382356	1.2012756	STEP 6	.50487580	.89685066
STEP 7	.15946144	1.3988129	STEP 7	-1.8568375	1.0027613
STEP 8	-.40207884E-01	1.6019713	STEP 8	-1.6101856	1.1240122

STEP 1	.30146748	.36988464	STEP 1	1.0181313	.39011652
STEP 2	.86393570	.43974624	STEP 2	2.1081736	.43008409
STEP 3	-.94820071	.57518439	STEP 3	.20381801	.58188704
STEP 4	-.82699019	.72384112	STEP 4	-.68749012	.76876949
STEP 5	.19278358	.89730231	STEP 5	1.2482760	.96833541
STEP 6	-1.3464464	1.0761021	STEP 6	-.62383681E-01	1.1633067
STEP 7	-.68387977	1.2575412	STEP 7	1.1030425	1.3467307
STEP 8	.38264488	1.4380397	STEP 8	2.1834662	1.5264310

STEP 1	.28151304	.37122241	STEP 1	1.1579288	.37597557
STEP 2	1.3857255	.43220711	STEP 2	1.4267253	.41493031
STEP 3	-.40481502	.56233234	STEP 3	-1.5699932	.52987381
STEP 4	.14551245	.71097169	STEP 4	1.4794622	.65583333
STEP 5	-.38504132	.87168263	STEP 5	-1.3419971	.78271541
STEP 6	-.22138417	1.0386108	STEP 6	-.22947599	.90049516
STEP 7	.33963871	1.2053073	STEP 7	-.74649361	1.0409983
STEP 8	1.4186107	1.3594446	STEP 8	-.11916426	1.1831189

STEP 1	.62497013	.35257583	STEP 1	-.61378690	.34947524
STEP 2	1.0057399	.39202649	STEP 2	-.36636080	.39695681
STEP 3	-.88231367	.51778802	STEP 3	-.70990276	.53567584
STEP 4	-1.3173083	.64603999	STEP 4	-.92146324	.67230187
STEP 5	-.98811069	.77831753	STEP 5	-.55524552	.80942465
STEP 6	-.45972793	.90718824	STEP 6	-1.2535907	.94581492
STEP 7	-1.1633819	1.0346532	STEP 7	.56215605	1.0801324
STEP 8	-1.8810562	1.1777101	STEP 8	.26555274	1.2330809

STEP 1	1.3018277	.39980864	STEP 1	.70869266	.39106205
STEP 2	1.8871713	.44546691	STEP 2	1.4578222	.44822833
STEP 3	-.34745659	.58192665	STEP 3	-.18640137	.59193926
STEP 4	-.78302220E-01	.74009157	STEP 4	-.54217831	.73897726
STEP 5	-.19244376	.89013121	STEP 5	.45031994	.88960014
STEP 6	-1.1040090	1.0363043	STEP 6	-.67334672	1.0293489
STEP 7	-.27342422	1.1924351	STEP 7	.97619660	1.1755913
STEP 8	-.95241408	1.3472315	STEP 8	-.30913698	1.3051197

Estimates of Gain and Standard Errors for
True Cubic Function, Gain = 5

STEP 1	5.8704069	.33082103	STEP 1	5.9339600	.37964978
STEP 2	7.1287397	.36461442	STEP 2	6.4635922	.40775244
STEP 3	3.6984101	.43966365	STEP 3	5.0687288	.55551634
STEP 4	4.4025426	.54922560	STEP 4	5.1027161	.70429384
STEP 5	4.7859538	.66290964	STEP 5	4.6694976	.85659590
STEP 6	6.7426233	.78807070	STEP 6	5.1036320	.99565701
STEP 7	4.1232774	.90522565	STEP 7	5.7751211	1.1420221
STEP 8	5.5953935	1.0352208	STEP 8	5.0058274	1.2880840
STEP 1	5.3364985	.35949982	STEP 1	5.6383341	.36696061
STEP 2	6.2447105	.38748614	STEP 2	6.0156303	.42085686
STEP 3	4.7508025	.52737830	STEP 3	4.3028115	.56905430
STEP 4	4.3945865	.66384914	STEP 4	4.2935798	.70615676
STEP 5	4.4927446	.80744713	STEP 5	4.9350049	.84754273
STEP 6	4.8104145	.95819253	STEP 6	6.0599835	.97590332
STEP 7	3.3917281	1.1066781	STEP 7	5.3163340	1.1132374
STEP 8	5.8113450	1.2425822	STEP 8	6.2858511	1.2418815
STEP 1	5.2788071	.34631805	STEP 1	5.7814805	.37203518
STEP 2	6.2101543	.38310060	STEP 2	6.3503709	.41800998
STEP 3	5.8048686	.53243483	STEP 3	3.6168058	.53623335
STEP 4	6.3126136	.67868610	STEP 4	4.0145585	.67508087
STEP 5	6.0508475	.81055051	STEP 5	4.9142022	.79573946
STEP 6	6.6083723	.93516381	STEP 6	2.7935640	.91828258
STEP 7	8.1977972	1.0587372	STEP 7	3.0649865	1.0579202
STEP 8	6.2958526	1.1867226	STEP 8	4.9003552	1.2012128
STEP 1	5.2256367	.47842629	STEP 1	5.2349352	.33313844
STEP 2	6.5578904	.53207442	STEP 2	6.2831987	.36936218
STEP 3	4.4744572	.72476032	STEP 3	3.9808989	.49312754
STEP 4	3.1112863	.92756643	STEP 4	3.9385382	.60795797
STEP 5	5.7816528	1.1327536	STEP 5	4.7080669	.73346812
STEP 6	3.4730819	1.3233036	STEP 6	4.6582847	.86969586
STEP 7	2.0109960	1.5266442	STEP 7	4.0818047	1.0332451
STEP 8	5.6050109	1.7229649	STEP 8	2.8737879	1.2046886
STEP 1	5.2896086	.31986348	STEP 1	5.6117397	.35076010
STEP 2	5.6569092	.34784880	STEP 2	6.6698875	.38784574
STEP 3	4.7265241	.46943391	STEP 3	4.2176775	.51576937
STEP 4	4.4355516	.59961678	STEP 4	5.7208171	.63930302
STEP 5	4.6121075	.73099131	STEP 5	6.3902409	.75722024
STEP 6	2.9400233	.86062053	STEP 6	5.6307829	.88233029
STEP 7	3.4121660	1.0089438	STEP 7	6.2201645	1.0208519
STEP 8	2.9329956	1.1654596	STEP 8	6.7151286	1.1692777

Table 10

Simulation Sampling Means and Standard Errors for True Cubic Function
($n=10$, $\sigma_t=1$, $\sigma_{e_x}=1$, $\sigma_{e_y}=1$, cutoff=-1.0)

Step	Gain = 0 $\hat{\beta}_2$	$SE(\hat{\beta}_2)$
1	.638	.183
2	1.241	.246
3	-.665	.169
4	-.296	.248
5	-.100	.262
6	-.464	.198
7	-.138	.306
8	-.066	.396

Step	Gain = 5 $\hat{\beta}_2$	$SE(\hat{\beta}_2)$
1	5.520	.088
2	6.358	.124
3	4.464	.209
4	4.573	.291
5	5.134	.214
6	4.882	.455
7	4.559	.576
8	5.202	.422

Table 11

Models Estimated at Each Step in the Regression-Discontinuity Analysis

Step 1: $y = \beta_o + \beta_1 x' + \beta_2 z + e$

Step 2: $y = \beta_o + \beta_1 x' + \beta_2 z + \beta_3 x'z + e$

Step 3: $y = \beta_o + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^2 + \beta_5 x'^2 z + e$

Step 4: $y = \beta_o + \beta_1 x' + \beta_2 z + \beta_3 x'z + \beta_4 x'^2 + \beta_5 x'^2 z + \beta_6 x'^3 + \beta_7 x'^3 z + e$

Table 12

Tests and Test Forms for Providence, R.I. Data

Grade	Pretest Test	Pretest Level	Pretest Form	Posttest Test	Posttest Level	Posttest Form
			Reading			
2	CTBS	B	S	CTBS	C	S
3	CTBS	C	S	CTBS	1	T
4	CTBS	1	S	CTBS	1	T
5	CTBS	1	S	CTBS	2	T
6	CTBS	2	S	CTBS	2	T
			Mathematics			
2	CTBS	B	S	CTBS	C	S
3	CTBS	C	S	CTBS	1	T
4	CTBS	1	S	CTBS	1	T

Table 13

Variables in Providence, R.I. Database*

Reading

1. Total Reading Score (pretest)
2. Total Reading Score (posttest)
3. Letter Sounds (level B only)
4. Word Recognition 1 (level B only)
5. Word Recognition 2 (level B only)
6. Comprehension (levels B, 1 and 2)
7. Comprehension- sentences (level C only)
8. Comprehension- passages (level C only)
9. Vocabulary (levels C, 1 and 2)
10. Group Designation (i.e., treatment, control or non-Title I)

Mathematics

1. Total Mathematics Score (pretest)
2. Total Mathematics Score (posttest)
3. Concepts and Applications (levels B and C)
4. Concepts (levels 1 and 2)
5. Applications (levels 1 and 2)
6. Computation
7. Group Designation (i.e., treatment, control and non-Title I)

* Separate datasets for reading and mathematics for each grade level and containing only matched scores were obtained.

Table 14

Descriptive Statistics For Reading Scores of
Second Through Sixth Grade Students in Providence, R.I.

		N	Mean	Variance	Skewness	Kurtosis
Second Grade						
Treatment	x	54	159.9	194.2	-.549	-.321
	y	54	229.9	1032.9	.405	2.022
Control	x	411	239.6	992.9	.121	.147
	y	411	292.1	3444.8	-.217	-.009
Non-Title I	x	239	252.7	1548.1	-.006	-.175
	y	234	323.4	3107.5	-.198	-.456
Third Grade						
Treatment	x	103	212.7	335.3	-1.171	1.603
	y	103	291.1	1698.9	-.109	.323
Control	x	385	307.0	1847.5	.442	-.429
	y	385	351.9	3286.1	.310	.794
Non-Title I	x	320	319.9	3276.8	.124	-.765
	y	324	388.8	5305.2	.503	.289
Fourth Grade						
Treatment	x	87	239.1	963.3	-1.069	.939
	y	87	325.2	1674.4	.281	-.719
Control	x	414	363.5	3090.5	.444	.087
	y	414	398.8	4123.6	.442	.619
Non-Title I	x	280	363.2	5167.6	.450	.553
	y	278	400.5	4626.6	.402	.413

Table 14
(Continued)

Fifth Grade						
Treatment	x	38	300.0	1105.3	-.295	-.754
	y	38	341.0	2254.6	-.018	-.152
Control	x	130	419.5	2384.9	1.229	2.553
	y	130	450.7	3670.8	.092	-.237
Non-Title I	x	261	404.4	5040.3	.163	.241
	y	260	452.6	6038.3	-.066	-.247
Sixth Grade						
Treatment	x	79	338.1	815.2	-1.021	.490
	y	79	370.4	1888.3	.250	-.711
Control	x	209	452.1	2516.6	1.163	2.321
	y	209	481.1	4009.9	.306	.429
Non-Title I	x	368	431.5	6039.9	-.005	.335
	y	365	465.9	7583.1	.204	-.186

Table 15
Descriptive Statistics for Mathematics Scores of
Second Through Fourth Grade Students in Providence, R.I.

		N	Mean	Variance	Skewness	Kurtosis
Second Grade						
Treatment	x	58	188.4	124.5	-1.026	.377
	y	58	259.8	930.3	.357	.057
Control	x	410	259.1	825.1	.536	-.045
	y	410	295.6	1212.9	.066	.016
Non-Title I	x	236	256.4	1176.5	.228	-.280
	y	239	310.8	1635.3	.025	-.116
Third Grade						
Treatment	x	80	234.4	179.7	-.729	-.363
	y	80	294.8	1287.7	-.449	.399
Control	x	401	297.3	795.3	.654	.051
	y	401	342.2	1730.6	.195	.177
Non-Title I	x	320	304.5	1524.1	.233	-.234
	y	317	369.5	2307.7	.533	.114
Fourth Grade						
Treatment	x	95	262.6	274.5	-1.409	2.042
	y	95	324.1	1277.9	-.195	-.677
Control	x	442	339.9	1348.4	.740	1.100
	y	442	389.2	1702.1	.343	.048
Non-Title I	x	280	343.2	2143.9	.428	.297
	y	275	390.3	2247.7	.265	-.032

Table 16

Estimates of Gain and Standard Errors
Providence Rhode Island

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
	Reading, Grade 2	
1	44.61	7.50
2	35.84	10.07
3	37.45	13.55
4	21.69	17.25
	Reading Grade 3	
1	29.73	6.12
2	16.96	6.94
3	16.34	9.25
4	-1.59	11.67
	Reading, Grade 4	
1	31.81	6.65
2	15.78	7.53
3	3.79	9.33
4	-3.75	11.10
	Reading, Grade 5	
1	.44	10.88
2	-8.86	13.08
3	18.60	17.28
4	22.74	22.03
	Reading, Grade 6	
1	-13.70	8.79
2	-22.13	9.94
3	-.81	14.38
4	.10	19.21

Table 16
(continued)

	Math, Grade 2	
1	8.17	4.91
2	4.32	5.84
3	-1.99	7.26
4	-10.82	8.68
	Math, Grade 3	
1	4.77	5.64
2	-4.31	6.60
3	-3.07	8.59
4	-8.20	10.78
	Math, Grade 4	
1	-4.98	4.56
2	-19.03	5.31
3	-10.84	7.01
4	-28.47	8.60

Table 17

Variables in Marion County, Florida Database

1. Total Reading Score (pretest)
2. Total Reading Score (posttest)
3. Total Math Score (pretest)
4. Total Math Score (posttest)
5. Pretest Grade
6. Pretest Test Level
7. Pretest Form
8. Posttest Grade
9. Posttest Test Level
10. Posttest Form
11. Merge Code (i.e., both tests, pretest only, posttest only)
12. School Type (i.e., Title I school or non-Title I school)
13. Group Designation (i.e., control, reading only, math only, both reading and math).

Table 18

Descriptive Statistics for Reading Scores of First Through Fifth Grade Students in Marion County, Florida

		N	Mean	Variance	Skewness	Kurtosis
First Grade						
Treatment	x	462	32.33	50.97	-.268	-.497
	y	462	242.56	1196.39	-.520	.526
Control	x	384	56.17	47.05	.136	-1.002
	y	384	288.98	804.17	.592	1.868
Second Grade						
Treatment	x	148	208.69	439.07	-1.449	2.354
	y	148	277.94	1009.47	-.493	1.774
Control	x	411	273.71	596.76	1.143	1.968
	y	411	346.72	2019.42	.246	-.126
Third Grade						
Treatment	x	288	261.65	815.78	-.988	.984
	y	288	317.16	1465.69	.209	.365
Control	x	369	349.65	1471.59	1.210	3.641
	y	369	409.01	2776.36	.719	.978
Fourth Grade						
Treatment	x	224	298.58	804.11	-1.083	1.480
	y	224	335.98	1822.38	-.216	.257
Control	x	446	399.21	2073.49	1.004	1.407
	y	446	445.26	3645.97	.706	.303

Table 18
(Continued)

Fifth Grade						
Treatment	x	173	315.45	882.75	-.423	-.580
	y	173	352.28	2054.21	-.021	-.201
Control	x	410	438.14	2959.86	.860	.085
	y	410	484.17	3527.23	.547	.433

Table 19

Descriptive Statistics for Mathematics Scores of Third Through Fifth Grade Students in Marion County, Florida

		N	Mean	Variance	Skewness	Kurtosis
Third Grade						
Treatment	x	297	263.24	409.96	-.839	.011
	y	297	332.22	1018.08	-.101	.707
Control	x	368	330.35	657.85	.573	-.283
	y	368	385.50	1492.70	.668	.255
Fourth Grade						
Treatment	x	187	296.10	377.57	-1.163	1.650
	y	187	349.52	1012.40	-.191	1.369
Control	x	458	370.25	1110.75	1.152	1.442
	y	458	412.60	2079.98	.307	-.339
Fifth Grade						
Treatment	x	156	321.99	434.34	-.922	.133
	y	156	370.42	1570.28	.759	1.365
Control	x	427	407.93	1529.45	.795	.378
	y	427	458.44	2741.26	.528	.555

Table 20

Estimates of Gain and Standard Errors
Marion County, Florida

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
	Reading, Grade 1	
1	2.07	3.88
2	1.96	3.92
3	-13.57	5.73
4	-5.41	7.72
	Reading, Grade 2	
1	.85	5.06
2	-3.44	5.12
3	4.20	7.19
4	26.10	9.42
	Reading, Grade 3	
1	-8.91	4.32
2	-11.89	4.40
3	.32	5.63
4	6.27	7.34
	Reading, Grade 4	
1	-1.31	4.95
2	-4.82	5.30
3	-6.99	7.30
4	-9.00	9.62
	Reading, Grade 5	
1	-28.83	5.27
2	-22.16	6.31
3	-19.38	9.26
4	-26.63	12.50

Table 20
(continued)

	Math, Grade 3	
1	8.56	3.89
2	8.07	3.90
3	5.86	5.61
4	10.58	7.55
	Math, Grade 4	
1	7.66	3.99
2	5.45	4.44
3	7.48	6.11
4	2.21	7.96
	Math, Grade 5	
1	-4.81	4.92
2	-4.27	5.63
3	8.83	8.05
4	8.60	10.48

Table 21

Tests and Test Forms for Osceola County, Florida Data

Grade	Test	Pretest Level	Form	Test	Posttest Level	Form
2	SAT	Pri 1	A	SAT	Pri 2	A
3	SAT	Pri 2	A	SAT	Pri 3	A
4	SAT	Pri 3	A	SAT	Int 1	A
5	SAT	Int 1	A	SAT	Int 2	A

Table 22

Variables in Osceola County, Florida Database

1. Total Battery
2. Total Reading
3. Total Mathematics
4. Total Auditory
5. Listening Comprehension
6. Vocabulary
7. Reading, Part A (Pri 1 and 2 only)
8. Reading, Part B (Pri 1 and 2 only)
9. Reading Comprehension
10. Word Study Skills
11. Math Concepts
12. Math Computation
13. Math Applications (except Pri 1)
14. Spelling (except Pri 1)
15. Language (except Pri 1 and 2)
16. Social Science (except Pri 1)
17. Science (except Pri 1)
18. Grade
19. Age
20. School
21. Reading Condition (i.e., treatment, control, holdover)
22. Math Condition (i.e., treatment, control, holdover)

Table 23

Reading Sample Sizes Obtained in Osceola County and Secondary Analysis

	Group	Osceola Analysis	Secondary Analysis
Reading			
Grade 2	T	23	33
	C	331	245
Grade 3	T	44	39
	C	304	145
Grade 4	T	29	28
	C	271	207
Grade 5	T	38	37
	C	300	260

Table 24

Descriptive Statistics for Reading Scores of
Second Through Fifth Grade Students in Osceola County, Florida

		N	Mean	Variance	Skewness	Kurtosis
Second Grade						
Treatment	x	33	-1.681	.210	-.370	-1.335
	y	33	-.788	.354	-.139	-.274
Control	x	245	.452	.622	.480	-.309
	y	245	.134	.667	.002	-.404
Third Grade						
Treatment	x	39	-1.389	.098	-1.604	2.754
	y	39	-.938	.165	.347	-.878
Control	x	195	.427	.528	.820	.106
	y	195	.461	.571	.407	-.148
Fourth Grade						
Treatment	x	28	-1.471	.153	-1.101	.408
	y	28	-.972	.344	-.518	.449
Control	x	207	.460	.595	.455	-.471
	y	207	.484	.553	.003	.635
Fifth Grade						
Treatment	x	37	-1.476	.117	-.692	-.372
	y	37	-.764	.345	-.687	1.268
Control	x	260	.425	.567	.562	-.312
	y	260	.540	.636	.114	.009

Table 25

Descriptive Statistics for Mathematic Scores of Second Through Fifth Grade Students in Osceola County, Florida

		N	Mean	Variance	Skewness	Kurtosis
Second Grade						
Treatment	x	25	-1.387	.062	-1.012	.411
	y	25	-.493	.390	.614	1.687
Control	x	246	.521	.617	.223	-.856
	y	246	.403	.740	.187	-.452
Third Grade						
Treatment	x	18	-1.533	.134	-.925	-.326
	y	18	.544	.637	1.376	3.106
Control	x	227	.455	.591	.571	-.317
	y	227	.463	.746	.190	-.621
Fourth Grade						
Treatment	x	17	-1.488	.183	-1.134	-.162
	y	17	-.597	.399	-.712	.093
Control	x	229	.446	.590	.457	-.516
	y	229	.432	.662	.185	-.348
Fifth Grade						
Treatment	x	21	-1.411	.095	-.846	-.193
	y	21	-.708	.157	1.067	.954
Control	x	269	.441	.548	.430	-.317
	y	269	.518	.592	.327	.007

Table 26

Estimates of Gain and Standard Errors
Osceola County, Florida

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
	Reading, Grade 2	
1	.71	.14
2	.52	.18
3	.39	.30
4	.53	.45
	Reading, Grade 3	
1	.15	.10
2	-.06	.12
3	.12	.20
4	.21	.32
	Reading, Grade 4	
1	-.11	.14
2	-.30	.17
3	-.54	.25
4	-1.16	.36
	Reading, Grade 5	
1	.26	.12
2	.?0	.16
3	-.22	.23
4	-.31	.34

Table 26
(continued)

	Math, Grade 2	
1	.27	.19
2	.19	.30
3	.66	.62
4	.54	1.19
	Math, Grade 3	
1	.71	.18
2	.75	.27
3	1.21	.49
4	.44	.85
	Math, Grade 4	
1	.54	.17
2	.22	.22
3	.96	.37
4	.95	.54
	Math, Grade 5	
1	.24	.13
2	-.02	.19
3	.16	.28
4	.09	.40

Table 27

Estimates of Gain and Standard Errors
for 11-Step Analysis
Providence, Math, Grade 4

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
1	-4.98	4.56
2	-19.04	5.31
3	-10.84	7.01
4	-28.47	8.60
5	-25.49	10.03
6	-34.94	11.33
7	-39.06	12.82
8	-42.42	14.49
9	-40.96	14.61
10	-41.81	16.27
11	-42.24	16.33

Table 28

Effects of Pretest Case Weighting on Estimates of Effect
Truncation on Pretest (0,1) Weighting
Providence, Math, Grade 4

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
	All Data (n=537)	
1	-4.98	4.56
2	-19.04	5.31
3	-10.84	7.01
4	-28.47	8.60
	Truncate at x = 475 (n=535)	
1	-2.07	4.58
2	-16.60	5.27
3	-15.68	7.09
4	-26.33	8.82
	Truncate at x = 450 (n=533)	
1	-2.06	4.63
2	-16.47	5.30
3	-16.48	7.18
4	-25.99	8.94
	Truncate at x = 425 (n=529)	
1	-2.12	4.68
2	-16.34	5.32
3	-17.65	7.23
4	-24.65	9.11
	Truncate at x = 400 (n=516)	
1	-3.77	4.79
2	-17.20	5.37
3	-16.97	7.43
4	-26.00	9.39

Table 28
(continued)

	Truncate at x = 375 (n=464)	
1	-5.83	5.12
2	-17.55	5.54
3	-17.92	7.75
4	-25.83	9.89
	Truncate at x = 350 (n=369)	
1	-10.93	5.73
2	-18.03	5.78
3	-19.03	8.16
4	-26.89	10.76
	Truncate at x = 325 (n=270)	
1	-19.46	6.79
2	-17.79	6.64
3	-24.41	9.78
4	-25.66	13.79
	Truncate at x = 300 (n=159)	
1	-27.07	7.45
2	-21.79	10.06
3	-34.03	17.89
4	-30.42	37.15

Table 29

Effects of Pretest Weighting of Estimates of Effect when Weighting is based on Absolute Distance from Cutoff
Providence, Math, Grade 4

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
	Linear Weight	
1	-4.24	4.98
2	-16.74	5.60
3	-17.47	7.47
4	-26.04	9.18
5	-27.66	10.73
6	-37.26	12.21
7	-39.97	13.93
8	-42.96	15.90
	Quadratic Weight	
1	-6.18	5.36
2	-17.03	5.92
3	-18.75	7.85
4	-26.19	9.56
5	-29.20	11.14
6	-37.85	12.71
7	-40.44	14.51
8	-43.57	16.64

Table 30

Descriptive Statistics for Posttest Subscales in Providence, R.I.

	N	Mean	Variance	Skewness	Kurtosis
Third Grade Reading					
Vocabulary	488	341.15	2905.40	.089	-.129
Comprehension	488	356.37	5078.55	.296	.806
Fourth Grade Mathematics					
Computation	537	386.10	1989.51	-.290	-.391
Concepts	537	386.91	3467.97	.180	-.266
Applications	533	379.29	4801.45	-.073	-.137

Table 31

Estimates of Gain and Standard Errors
Posttest Subscales
Providence, Rhode Island

Step	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$
	Reading Vocabulary Subscale, Grade 3	
1	.27.02	5.59
2	14.36	6.32
3	22.66	8.40
4	-1.00	10.52
	Reading Comprehension Subscale, Grade 3	
1	31.94	8.21
2	19.95	9.38
3	9.47	12.47
4	-7.48	15.79
	Math Computation Subscale, Grade 4	
1	-5.41	5.04
2	-13.65	6.03
3	5.61	7.96
4	-12.03	9.93
	Math Concepts Subscale, Grade 4	
1	-8.79	6.78
2	-28.46	8.02
3	-29.98	10.76
4	-38.93	13.48
	Math Applications Subscale, Grade 4	
1	.13	7.10
2	-21.08	8.39
3	-4.64	11.17
4	-27.21	13.95

Dissertation Contact Persons

Federal

Ron Fishbein, Office of Education*

Judith Anderson, Office of Education*

Technical Assistance Center Personnel

Everett Barnes, Region I***

Rich Hill, Region I***

Steve Horowitz, Region I

Gary Echternact, Region II*

Joan Troy, Region III

Keith McNeil, Region III

Roy Hardy, Region IV

George Temp, Region IV

Ted Storlie, Region V*

Jerry Goldman, Region V

Laura Crane, Region V*

Paul Rasseld, Region VI

Carmen Finley, Region VII

Dick Bennett, Region VIII

Gary Estes, Region IX

Barbara Williams, Region X

State Title I Personnel

Richard Hodges, New Hampshire

Richard Zussman, Massachusetts

Patrick Proctor, Connecticut

Juanita Parkes, Florida**

Clyde Bezanson, Minnesota*

Carl Evans, Ohio*

Jean Hall, Indiana*

David Krueger, Wisconsin*

Rose O. Maye, Illinois*

Paul Novak, Michigan*

Claudia Merkel-Keller, New Jersey*

Eileen Kelley, New York*

Maria DelGado, Puerto Rico*

Kurt Komines, Virgin Islands*

Local Title I Personnel

Ronald Visco, Providence, RI***

Lou Perrullo, Boston, MA

Elspeth Lloyd, Escambia County, FL

Naomi Winbush, Osceola County, FL**

John Drewes, Osceola County, FL

Arthur Roberson, Duval County, FL

Art Roberts, Duval County, FL

Mille Murray, Hillsborough County, FL**

Mary Anne DeLong, Marion County, FL

Mel Lucas, Marion County, FL**

Steve Iachinni, Pinellas County, FL

Ricky Nations, Sarasota County, FL

Others

Kast Talmadge, RMC Corporation, Mountain View, CA

Chris Wood, RMC Corporation, Mountain View, CA

Beverly Helms, Panhandle Area Educational Cooperative, FL

Cliff Blair, University of South Florida**

Bruce Hall, University of South Florida**

Notes

The persons listed above were contacted in telephone interviews which lasted anywhere from five minutes to over an hour or, if indicated, were contacted directly in interview or meetings as described below.

* Participated in Title I Evaluators' Conference for Regions II and V, Carson Inn/Nordic Hills Conference Center, Itasca Illinois, January 22 - 23, 1980.

** Interviewed in Florida Site Visits, February 4 - 8, 1980.

*** Interviewed in Region I Site Visit, December 26 - 27, 1979.

Regression-Discontinuity TAC Interview

TAC Experience

Does your TAC assign particular individuals to RD or train generally? (Personnel)

What type of training (how much, method of presentation) do you provide to TAC personnel? (Training)

What is the role of the technical advisor in recommending or instructing SEAs and LEAs on the choice of research design? (Role)

What types of training for RD (how much, method, etc.) does your TAC offer to SEAs and LEAs? (Training)

What is the role of TAC personnel in helping SEAs or LEAs implement the RD? (Role)

What role do TAC personnel have in statistically analyzing data from the RD? (Role)

Views of SEA and LEA experiences

How often is RD considered, especially relative to other RMC models? (Feasibility)

Why, if at all, do people at SEAs/LEAs say they would like to use RD? (Applicability)

Why do they say they don't want to use RD? (Feasibility)

How often has RD been attempted in your region? (Use)

Have the attempts to implement RD been successful? (Implementation)

Are there particular settings (e.g., certain types of Title I programs) which are seen as especially appropriate for RD? (Applicability)

What aspects of RD seem to cause the most difficulty to those who try to implement it? (Implementation)

Specifically, how much are each of the following a problem:

- getting data for control subjects
- using pretest score as only assignment variable - (is it the general view that RD requires the pretest or a single pre-measure?)
- adhering to the cutoff in assigning (fuzzy RD)
- assumption of linearity
- costs

Request for Materials

Materials used for training staff members.
Materials used for technical assistance to SEAs/LEAs.
List of LEAs which have tried RD.
Names of TAC personnel who have personal experience with RD.
Names of SEA/LEA people who have had experience with RD.

Site Visit - Explore for potential usefulness.

The Florida Department of Education Title I proposal form is a rather detailed 24 page packet which each district supplies annually as part of their application for funding. The following is a partial list of variables collected:

1. General district description information.
2. School Eligibility Survey - This includes a ranking of each school in the district on a measure of economic need (e.g., percent of students receiving school lunches, percent of AFDC children, etc). This rank is used to determine which schools are eligible for Title I funding.
3. Needs Assessment - A listing by grades of the numbers of children who score below some national norm on an achievement test.
4. Information on how the district plans to conduct Title I activities with activities of other agencies.
5. Project Planning: Plan of dissemination of Title I information to teachers and administrators.
6. Project Planning: List of relevant personnel, their roles, and expected hours.
7. Plan for supplemental use of funds in the regular school program and the state compensatory program.
8. Number of participants by grade levels for public and private schools.
9. Description of Project including:
    a. Subject matter.
    b. Setting.
    c. Grades served.
    d. Instructed staff/student participation.
    e. Days per week each student is served.
    f. Instructional period length.
    g. Weekly student instructional hours.
    h. Estimated per pupil expenditures.
    i. Description of selection criteria.
    j. Number of students identified in district.
    k. Objectives of its program.
    l. Project narrative.
10. Project Evaluator Description including:
    a. Grades served.
    b. Evaluator model.
    c. Pretest name, level and date of administration.

d. Posttest name, level and date of administration.
e. Evaluation plan narrative

11. Program information by site.

12. Itemized Estimated Expenditures.

13. Equipment and Uses.

14. Budget

15. Personnel Job Descriptions.

16. Inservice Training Activities.

17. Parent Advisory Council Activities

18. Location and Use of Title I Funded Portable Classrooms.

19. Plan for dissemination of information to teachers, parents and the general public.

20. Justification for Capital Outlay.

21. Statements of Assurance.

22. Approval Forms.

The most relevant item for this work were the description of the project (Item 9) and the description of the evaluation (Item 10).

The Florida Department of Education Title I report details the results of the annual evaluation including:

1. General district description information.
2. Numbers of school age children and numbers of Title I participants (unduplicated count).
3. Number of students in Title I programs (duplicated count) by subject area of program.
4. Staff assigned to Title I programs.
5. Parent Activities.
6. Program Effectiveness Data
   a. Subject matter.
   b. Setting.
   c. Grades served.
   d. Instructional staff/student Ratio.
   e. Student Instructional hours.
   f. Testing cycle.
   g. Grade
   h. Explanations for any data not reported.

In addition, the following is reported by model

Model A:
   a. Number of students in treatment and comparison group.
   b. Pretest, posttest and adjusted posttest standard score for treatment and comparison group.
   c. Pretest, posttest and adjusted posttest NCF for treatment and comparison group.
   d. NCF gain.
   e. Weighted gain.
   f. Mean weighted gain.

Model C:
   a. Number of students in treatment and comparison group.
   b. Mean pretest treatment NCF.
   c. Observed and predicted posttest means.
   d. NCF intercepts at cutoff for treatment and comparison groups.
   e. NCF gain at the treatment group pretest mean.
   f. NCF gain at the cutoff.
   g. Weighted NCF gain.
   h. Weighted mean NCF gain.

7. Program Effectiveness Data when using other than the three models.
8. Outcomes for other objectives.

Of primary interest for this study is the Program effectiveness data (Item 6) especially that which is reported for Model C.

Regression-Discontinuity Sites

KEY: It is very difficult to obtain accurate information about who is using regression-discontinuity except by contacting the school district directly. Therefore, the following is a list which is as complete as possible by contacting primarily regional Technical Assistance Center (TAC) people. The key below gives the best available information short of direct contact. The source for each citation is also listed.

C - district is currently using regression-discontinuity

D - district expressed desire (usually in print) but never did it

U - district used regression-discontinuity but abandoned

P - district is planning to use regression-discontinuity

M - district mentioned as possibly using (or having used)

A - district did regression-discontinuity analysis on data collected by the norm-referenced model (Model A)

B - district implemented regression-discontinuity (Model C) and the norm-referenced model simultaneously

Region I

Massachusetts

Attleboro	C
Boston	A
Mansfield	D
New Bedord	D
North Adams	D
Pittsfield	D
Williamsburg	D

Maine

Augusta	D
Bath	D
Portland	D
Van Buren	D

New Hampshire

Londonderry	D
Portsmouth	D
Peterborough	D
Franklin	D
Bristol	D
Stratford	D
Wolfboro	U

Rhode Island

Providence	C
Woonsocket	P
Cheriho	A

Connecticut

Norwalk	D

Vermont

none

Region II

New York

(several, contact SEA)	C

Virgin Islands

none

Puerto Rico	C

New Jersey

Burlington Twp.	C
Vineland	C
Union City	C
Bayonne	C
Hopatcong	C
Bridgewater-Raritan	C
Nutley	C
Essex County	C
South Brunswick	C
Milltown	C
Piscataway	C
South River	C
Jackson	C
South Regional	C
Sussex-Wantage Regional	C
Phillyssburg	C

Region III

Pennsylvania

Phillipsburg	C

Delaware

none

Maryland

none

Washington, D.C.	M

West Virginia

none

Virginia

Portsmouth County	U
Loudon County	U
Fairfax County	U
Carroll County	C
Accomack County	C
Giles County	C
Westmoreland County	C
Glouster County	C
Powhatan County	C
Isle of Wight	C

Region IV

Georgia

(two or three)	C

North Carolina

(three or four)	C

Florida

Duval County	C
Hillsborought County	C
Jackson County	C
Levy County	C
Marion County	C
Osceola County	C
Polk County	C
Sarasota County	C
Calhoun County	U
Franklin County	U
Gulf County	U
Holmes County	U
Manatee County	U
Walton County	U
Washington County	U
Escambia County	U,B

South Carolina - none

Mississippi - none

Alabama - none

Kentucky - none

Tennessee - none

Region V

Illinois - none

Indiana - none

Michigan - none

Minnesota - none

Ohio - none

Wisconsin - none

Region VI

Arkansas - none

Bureau of Indian Affairs - none

Oklahoma - none

Texas - none

Louisiana - (three)

New Mexico - none

Region VII

Iowa - none

Kansas - none

Missouri - none

Nebraska - none

Region VIII

Montana - (one)

Colorado - (one)

Utah - (two)

South Dakota - none

North Dakota - none

Wyoming - none

Region IX

Arizona

Tucson	C
Wilcox	C
Balz	C
Mesa	U

California

Cajon Valley	U
Monterey	U
Lakeport	U
Monroe Elementary	U

Guam C

Northern Mariana Islands C

Samoa U

Hawaii - none

Nevada - none

Trust Territory - none

Region X

Idaho - none

Alaska - none

Oregon

Springfield	U
Portland	C
Roseburg	P

Washington

Spokane	C
Takoma	U
Seattle	U

VITA

William M. K. Trochim

18 Faculty Circle
Lake Forest
Illinois 60045
(312) 234-3253

Division of Methodology and
Evaluation Research
Department of Psychology
Northwestern University
Evanston, Illinois 60201
(312) 492-5292

EDUCATION

1977 - present	Ph.D., Methodology and Evaluation Research (anticipated June, 1980) Division of Methodology and Evaluation Research Department of Psychology Northwestern University Evanston, Illinois
1973 - 1975	M.A. with Honors, General - Experimental Psychology Department of Psychology Southern Connecticut State College New Haven, Connecticut
1970 - 1973	B.A., Psychology Mundelein College Chicago, Illinois

EMPLOYMENT

1976 - 1977	Associate in Research Yale University and the Connecticut Mental Health Center New Haven, Connecticut Duties included the development and management of a computerized time-series evaluation system for psychiatric data.
1975 - 1976	Assistant in Research Social Psychology Program Department of Psychology Yale University New Haven, Connecticut Duties included the management and statistical analysis of social psychological experiments in impression formation, attitude change and inter-racial relations.

TEACHING EXPERIENCE

1975 - 1977 Lecturer
Department of Psychology
Southern Connecticut State College
New Haven, Connecticut

Courses taught:
Introductory Psychology (4)
Social Psychology (2)
Experimental Psychology
Introductory Statistics

WORKSHOPS CONDUCTED

May, 1979 Time Series Analysis in Human Service Evaluation
Department of Human Services
State of Maine
Augusta, Maine

June, 1979 The Statistical Analysis of Time-Series Data
Maine Health Information Service
Augusta, Maine

CONSULTATION

1979 - 1980 Research Associate
Robert F. Boruch and David S. Cordray,
Co-Principal Investigators
"A comprehensive study of practices and procedures in Federally funded education programs."

1979 - 1980 Computer Consultant
The CAUSE Project: Comprehensive Assistance to Undergraduate Science Education
Lake Forest College
Lake Forest, Illinois

1979 - 1980 Evaluation Consultant
Donnelley Library
Lake Forest College
Lake Forest, Illinois

1978 - 1979 Evaluation Consultant
Project WILL: Women in Leadership Learning
Barat College
Lake Forest, Illinois

PROFESSIONAL AFFILIATIONS

Professional Advisory Commission to the Mental Health Association of Evanston (Illinois)

Evaluation Research Society

PUBLICATIONS

Trochim, W. Locating data for secondary analysis. In Boruch, R.F., Wortman, P.M. and Cordray, D.S., eds. Secondary Analysis in Applied Social Research. San Francisco: Jossey-Bass, in press.

Cook, R.D., DelRosario, M., Hennigan, K., Mark, M., and Trochim, W., eds. Evaluation Studies Review Annual. Beverly Hills, Calif.: Sage Publications, 1978.

Trochim, W. The three-dimensional graphic method for quantifying body position. Behavior Research Methods and Instrumentation, V. 8 (1976) p. 1-4.

PAPERS

Campbell, D.T., Reichardt, C.S., and Trochim, W. The analysis of the "fuzzy" regression-discontinuity design. Northwestern University, 1979.

Trochim, W. Some economic issues in residential peak-load pricing. Northwestern University, 1979.

Trochim, W. Using quantitative methods in decision-making. In Cochran, N., Gordon, A., and Campbell, D.T., eds. Bureaucracies, Social Experimentation and Evaluation, Vol. II. Unpublished monograph, Northwestern University, 1978.

TECHNICAL REPORTS

Trochim, W. An introduction to the analysis of the regression-discontinuity design. NSF Grant BNS 7827810, Northwestern University, 1979.

Trochim, W. Proposal for a time-series based evaluation system for a state mental health center. NIMH Grant 5 T32 MH15113-02, Northwestern University, 1978.

Trochim, W. The analysis of the interrupted time-series reversal design. NIMH Grant 5 T32 MH 15113-02, Northwestern University, 1978.

Trochim, W. The effect of welfare recovery laws: considerations for a secondary analysis. NIE-C-74-0115, Northwestern University, 1977.

MAJOR RESEARCH INTERESTS

Methodology and evaluation research; secondary analysis; applied social psychology; evaluation policy and standards; statistical analysis; evaluation in mental health, human services, energy and education.

REFERENCES

Dr. Robert F. Boruch
Director, Division of Methodology and Evaluation Research
Department of Psychology
Northwestern University
Evanston, Illinois 60201

Dr. Donald T. Campbell
Maxwell School
Syracuse University
Syracuse, New York 13210

Dr. Thomas D. Cook
Division of Social Psychology
Department of Psychology
Northwestern University
Evanston, Illinois 60201

Dr. David S. Cordray
Division of Methodology and Evaluation Research
Department of Psychology
Northwestern University
Evanston, Illinois 60201