DOCUMENT RESUME

ED 135 869 95 TM 006 101

AUTHOR Boruch, R. F.; De Gracie, J. S.
TITLE The Use of Regression Discontinuity Model with Criterion-Referenced Testing in the Evaluation of Compensatory Education.
INSTITUTION Northwestern Univ., Evanston, Ill.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [Apr 77]
CONTRACT NIE-C-74-0015
NOTE 56p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *Compensatory Education Programs; *Criterion Referenced Tests; Elementary Education; Hypothesis Testing; *Mathematical Models; Measurement Techniques; Multiple Regression Analysis; *Program Effectiveness; *Program Evaluation; *Reading Programs; School Districts; Statistical Analysis; Testing Programs
IDENTIFIERS Arizona (Mesa); Elementary Secondary Education Act Title I; Mesa Public Schools AZ; *Regression Discontinuity Model

ABSTRACT

The results of a study which used the regression discontinuity model in the analysis of criterion-referenced test data are presented and inherent drawbacks in the use of the regression discontinuity model with this type of data are discussed. The discussion is based on the results of an analysis of data collected in the Mesa, Arizona School District which uses criterion-referenced testing solely. The problems include celling effects and non-linearity of the data. Problems inherent in implementation of an evaluation design at the school level are also described. (Author/MV)

THE USE OF REGRESSION DISCONTINUITY MODEL WITH CRITERION-REFERENCED TESTING IN THE EVALUATION OF COMPENSATORY EDUCATION

R.F. Boruch
J.S. DeGracie

Presented at the:
American Educational Research Association Annual Meeting
April 1977

Evaluation Research Program
NIE Project on Secondary Analysis
NIE-C-74-0115

Northwestern University
Evanston, Illinois 60201

TABLE OF CONTENTS

THE USE OF REGRESSION DISCONTINUITY MODEL WITH CRITERION-REFERENCED TESTING IN THE EVALUATION OF COMPENSATORY EDUCATION

R.F. Boruch
J.S. DeGracie

1. Introduction

It would appear that within the next year and a half a relatively uniform procedure for the evaluation of Title I programs at the local educational agency (LEA) level will be mandated. There certainly can be no argument against the need for more comprehensive evaluation at the LEA level. An overwhelming argument in support of this need is the findings of Talmadge in 1974. At that time he directed a search which encompassed some 2,000 projects conducted at LEAs, all of which had received some form of official recognition for success. Under close scrutiny only 6 of the 2,000 could be found to meet the selection criteria of effectiveness, cost, availability and replicability. These findings plus the intuitive observations of the local education agencies are not lost on the LEAs. If a survey was to be conducted of the LEAs, I am sure that overwhelming support for the need for more comprehensive evaluation at the local level would be evidenced.

This paper, then, is not intended as a testimonial for the exclusion of such a mandate, but rather as an instructive guide to some of the problems and pitfalls which can be encountered at the local level when highly sophisticated statistical methods are used to investigate program effectiveness at a LEA. Again, this is not to say that these problems and pitfalls cannot be overcome, but they should certainly be investigated to the fullest extent possible before a general mandate is made.

Within the last year representatives from RMC Research Corporation and the Northwest Regional Laboratory have made presentations in Arizona concerning the forthcoming mandate. These presentations drew heavily on the work by Talmadge published both in the monograph, "A Practical Guide to Measuring Project Impact on Student Achievement," and "State ESEA Title I Reports: Review and Analysis." In both of these presentations the models which they felt will be mandated were discussed. In general, there was direct agreement between the two presentations. The only disagreement was in which model seemed to be the one which would be most used by the LEAs. The representatives from RMC seemed to concur with some of the previous writings of Talmadge that Model B, the Control Group Model, would be extremely difficult to implement since control groups for Title I students are not normally available. I would have to tend to agree with this feeling. Here, however, I find myself at odds with my co-author Dr. Boruch, who addressed this topic in discussing Talmadge's work seemed to reject the idea of randomized experimental tests of compensatory programs out of hand, suggesting that their rarity is due by and large to a low feasibility level. At that time, and also documented in Campbell and Boruch (1975), he discussed and gave examples of a number of successful and unsuccessful randomized studies in education. He strongly reinforced the point concerning the utility and feasibility of randomized field tests of compensatory programs and of program elements. The representatives from Northwest Regional Educational Laboratory seemed to agree with Dr. Boruch. They felt and strongly urged that Model B, the Control Group Model, be the one used. Here, however, both authors find ourselves at odds with their reasons for selecting the Control Group Model. It was stated in the presentation that the Control Group Model could be used with equivalent or

nearly equivalent groups. It was further stated that if the groups were not equivalent enough analysis of covariance could be used to match the groups. In this case both authors must agree with Talmadge, who stresses the need for the treatment and comparison group to be sufficiently similar that they can be considered as random samples from a single population. The technique of analysis of variance is a technique for reducing the error variance in an experiment and it is not a technique which can be used to balance groups which are not similar. The theory behind the analysis of covariance is that the groups are randomly selected from the same population. Further discussions of this point can be found in Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child, New York: Brunner/Mazel, 1970, where Campbell discusses regression artifacts in quasi-experimental evaluations.

The above discussion is included to underline the fact that the proposed evaluation models are difficult to select and even more so to implement. If the original proposer of the models, the consultant hired to assist the LEAs in selection and implementation of the model, and other experts in the field cannot agree, it can be seen what problems will face the local educational agencies when the mandate is implemented.

The area that is specifically addressed in this paper is the selection and implementation of one of the evaluation models when the local educational agency is heavily committed to their own criterion-referenced objective-based program. A number of LEAs in recent years have made significant commitments to locally-developed objective-based criterion-referenced testing, which form the backbone for district generated classroom management systems. The Mesa Public School District is an example of such a LEA. For 5 years they have been in the process of moving from a total standardized testing program to a criterion-referenced testing program. These programs

on objectives for the given programs developed by task forces of district personnel including classroom teachers, midlevel administrators and district curriculum people. Over the past 5 years, it is felt that this effort has yielded a viable classroom management system which leads to the accomplishment of the specific goals and objectives which are felt to be important, not only at the local school level, but all the way up to the superintendency and the board of education. The Mesa Public Schools, then, as are other local educational agencies that have similarly developed their own criterion-referenced testing programs, are faced with the problem of selecting one of the three proposed evaluation models and, once selected, to use this model for the evaluation of their programs.

To get a jump on the Federal mandate, the Mesa Public Schools with the assistance of the Evaluation and Research program, and NIE Project on Secondary Analysis at Northwestern University, attempted to select an appropriate recommended evaluation model and to implement this model in its program evaluation. In its selection of a model the underlying assumption was that the district-generated criterion-referenced testing should furnish the necessary test data for the evaluation. This, then, left a choice between Model B, the Control Group Model, and Model C, the Regression Model. It was felt that in the Mesa Public Schools it was simply not feasible to select students randomly for the program. And it was also felt that because of the uniqueness of the students being served, no control group was available which could be considered as a random sample from the same population as that of the treatment group. Therefore, through a process of elimination Model C or the Regression model utilizing the theory of regression discontinuity analysis was selected. In the next pages, then, you will find the results of that analysis. In addition, some background concerning the exact program is also furnished.

As stated, this is a report of an evaluation of the Mesa Compensatory Reading Program supported by Title I funds in the Mesa Public School District, Mesa, Arizona. The approach used here to estimate the program's effects on children's reading ability is the regression-discontinuity (RD) design proposed in 1960 by Thistlethwaite and Campbell. Partly because this is a novel application of a promising but underutilized approach to program evaluation, we dedicate particular attention to both substantive estimates of program effect and to the credibility and usefulness of the RD approach.

In the following remarks, we first provide background information on the Mesa Compensatory Reading Program (Section 2), and on the basic evaluation design (Section 3). Section 4 summarizes data on reading test scores collected under the design. Succeeding sections cover the results of alternative competing analyses: conventional approaches based on linear models (5) and less conventional approaches based on nonlinear models (Section 6). Section 7 is a summary, not so much of the findings of the analysis, but more toward the concerns that must be expressed as a result of the analysis.

2. Background: The Mesa Compensatory Reading Program

The Mesa Public School District has operated a Reading Classroom Management System since 1970 to diagnose, prescribe, and monitor individual reading skills at all grade levels. The system has terminal goals, program goals, and behavioral objectives for each skill at each grade level. Diagnostic assessment tests are administered early and late within each grade levels; criterion-referenced tests are used for formative evaluation.

Although reliability data has not been collected on the assessment instruments, the test items have undergone an iterative method of item analysis using the responses from over 12,000 students. The District's Office of Research maintains that the resulting tests have evolved over this period of time, 1970-1975, into valid and reliable instruments for measuring student achievement.

Reading services are provided to 25 elementary schools by 20 reading resource teachers, 25 district reading paraprofessionals, and 35 Title I supported reading paraprofessionals. Of the 25 schools, 11 have been designated as Title I schools and receive additional services, i.e., in addition to receiving a district paraprofessional, the 35 Title I paraprofessionals are divided among the 11 schools. All paraprofessionals are trained with a 20-hour competency-based program. The major goal of the reading program is to alleviate reading problems by concentrating resources at the primary grades with first grades receiving the top priority. Therefore, more students are provided individual attention at first grade than any other grade. Second grades receive more services

than third graders, etc., with each subsequent grade receiving less individual services.

Those students that are identified as being educationally deprived on the basis of the Mesa-developed criterion-referenced tests are given assistance by the paraprofessionals for approximately one-half hour a day for four days each week. This assistance is given in groups no larger than five students. The students are either removed from the classroom, or in some cases where the classroom is in an open space, the students are moved to a separate section of the classroom area. The total time during the school year that the students receive assistance is approximately 28 weeks. The first two weeks at the beginning of the school year are taken up by training the paraprofessionals and the last few weeks at the end of the school year the paraprofessionals are released as this time is not usually exclusively instructional time and the paraprofessionals are specifically employed to directly impact the students. The identified students at the first-grade level usually spend the entire year with the paraprofessionals. At the other grade levels the students are more apt to be placed back into the classroom setting as soon as they accomplish their specific deficiencies that were identified through the use of the criterion-referenced tests. On the average, approximately half of the students above grade one spend the entire year with the Title I program, with the other half spending approximately one-half of the school year with the program.

2.1 Selection of Students for Special Assistance

All students are pretested in September with the criterion-referenced tests created by the school district's reading program staff, with

exception of first graders who are administered the Murphy-Durrell Reading Readiness test. . First graders take the district criterion-referenced test in January for further refinement of subskill needs.

The criteria for selecting students for individual instruction in the compensatory reading program is based on test scores. Specifically, a student beginning first grade must score 71 or lower on the Murphy-Durrell Reading Readiness test. In grades 2 to 6, a student must score 50% or lower on the school district's criterion-referenced tests to be eligible for special assistance from district and Title I resources.

At the 11 Title I schools, students identified as needing extra assistance are assigned to a reading paraprofessional. Using the test results, the reading resource teacher prescribes appropriate activities which the paraprofessional implements. The reading resource teacher monitors this instruction weekly and adjusts according to student progress. The Title I student thereby receives additional services above and beyond the classroom and district resources.

3. Evaluation Design

The main substantive objectives of the evaluation is to determine

whether the Mesa Compensatory Reading Program exerts a notable effect on children's reading ability and to estimate the magnitude of the effect. Reading ability here is defined operationally as scores achieved by students on the tests. The main methodological objectives are to better understand the benefits and limitations of the regression-discontinuity

The basic RD design was developed for those cases in which some treatment (an award, a program) is offered to individuals on the basis of a meritocratic criteria and there is some need to estimate effects. This eligibility criteria must be a measurable continuum, such as economic need, educational need, and so forth. And, in the simplest application, individuals must be assigned strictly on the basis of this eligibility; e.g., those children scoring below a certain score on a reading test receive the program, those scoring above the cutting point do not. The preprogram eligibility must be related in a known way (e.g., linearly) to the post-program score in the absence of any program effects.

The post-program score is the dependent variable in such analyses. Assuming that the regression of this dependent variable on eligibility is linear in the absence of any program, one then looks for a discontinuity in the observed regression to infer program effects. That is, the regression of post-program reading scores on pre-program scores for the program recipient group will differ from the corresponding regression line for the nonrecipient group, provided that the program has an effect (Figure 1b). If there is no effect, both groups will be described well by the same regression line (Figure 1a).

4. The Data: Descriptive Statistics and Adherence to the Design

Reading test scores were available for students in the first, third, and fifth grades on the instrument described in Section 2. The statistical summary of these data is given in Table 1 and includes mean, variance, skewness, and kurtosis for each sample of recipients and nonrecipients within grade level, for pre-program test scores (X) and post-program scores (Y).

4.1 Skewness

Though Thistlethwaite and Campbell (1960) do not seem to recognize it, the skewness statistic serves as a check on the process generating the RD data. For if a sharp cutting point is used for the pretest, we would expect scores of program recipients to be negatively skewed (bunched near the cutting point) and we would expect nonrecipient scores to be positively skewed, if the overall distribution is symmetric and roughly normal in shape.

For the posttest, we would expect scores which are initially skewed negative to become less negatively skewed if the treatment has an effect. That is, students would become more spread out in their ability (this assumes that the effect is not completely additive). The nonrecipient group scores would also become less skewed, unless the posttest has a low ceiling, in which case we might expect positive skew to become more negative.

For the data at hand, we have what one might expect of the distribution of program recipients' scores. They are skewed negatively, suggesting that a sharp cutting point has been used. And the negativity decreases

probably because of treatment effects, random errors of measurement, and other factors. This is true for each grade level examined except the third, where level of skewness does not change.

The sample distributions for the nonrecipients runs counter to expectations, however. With strict adherence to an RD approach and with a roughly symmetric distribution of scores on X one would expect the nonrecipient sample to have a positive skew or possibly little or no skew. But in the data at hand, all samples are negatively skewed at pretest (X); the skew is notable for the first grade, negligible for the fifth. This suggests that some data may be missing, that students are not assigned to program strictly on the basis of eligibility scores, and that the tests may have a low ceiling effect, especially for the fifth-grade group.

4.2 Variances

For truncated distributions of fallible observations, one would expect some increase in variance from pretest to posttest, and indeed this is reflected in the data. The coefficient of variation behaves quite erratically and is uninformative.

4.3 Adherence to the Design

The counter-intuitive statistics for skewness implies a failure to adhere to a strict cutting point, and a need to review the information we have on assignment of children to the program.

Of 25 elementary schools in the district, 11 received Title I funds during 1973-74 on the basis of need. Need here is defined by low average economic income level of families within the district according

to a weighted mean of the number of students identified under free lunch, ADC, and the 1970 U.S. Census.

Within a school, the nominal cutting points for assignment to services are, as indicated earlier, 71 for the first grade, and 50 for the third and fifth grades. However, there was no strict adherence to the cutting point. In fact, some program recipients received very high scores in the pretest, and there were substantial number of individuals who did score above the nominal cutting point who did receive services. The percentage of students in each grade level in each subgroup is given below.

	N	First Grade	Third Grade	Fifth Grade
Eligible Recipients	347	.17	.32	.15
Ineligible Recipients	485	.45	.11	.20
Ineligible Non-Recipients	358	.38	.57	.65

The number of individuals with high scores who receive services suggest that teachers are attempting to service as many students as possible. Most such students actually need the service. However, if need is defined solely in terms of the criterion, many do not. There may be several reasons for this phenomena. Reading program teachers may feel compelled to make contact, however brief, with as many students as possible to satisfy some vague idea that the more students they serve, the most valuable their contribution will appear to be. The "program recipient" label itself may be misleading insofar as some of these students may receive only brief attention--enough, say, to establish further that they need little or no help. At this stage of the research we know little about

which explanation is true. Other data on duration or frequency of service to a student are essential for establishing an explanation, or at least establishing the extent to which ineligible recipients are receiving nominal tutoring.

In any event, we conclude that the adherence to the original design is best for the third grade and fifth grades where respectively 11% and 20% of children whose pretest scores are high are assigned to the program. Adherence is worst for the first grade. The implication of nonadherence to the original design is that the original design models and analysis determined by those models cannot be used without modification.

5. Linear Regression Approach

5.1 Eligible Recipients, Ineligible Nonrecipients

The analyses in this section are based on the X,Y points for eligible program recipients and the corresponding X,Y points for program nonrecipients who were ineligible for the program. Data on children who were ineligible for the program, i.e. scored above the cutting point on their reading pretest, but received the program are put aside temporarily.

One of the simplest approaches to analyzing data from an RD set up is to assume that posttest is linearly related to pretest within each group. That is, one assumes

$$Y_{i1} = \alpha_1 + \beta_{i1} X_{i1} + e_{i1} \qquad e_{i1} \sim I(0,\sigma^2)$$

and

$$Y_{i2} = \alpha_2 + \beta_2 X_{i2} + e_{i2} \qquad e_{i2} \sim I(0,\sigma^2)$$

for the program recipient group (1) and the nonrecipient group (2) respectively. To assess the program's impact, assume that under null conditions, the regression lines are identical:

$$H_0: \alpha_1 = \alpha_2; \beta_1 = \beta_2 | \sigma_1^2 = \sigma_2^2$$

The test of the hypothesis is clear-cut (see Chow, 1960; Gulliksen & Wilks, 1954, for example). If the hypothesis is rejected and the other assumptions hold, we may infer that the program exerted an influence on slope, intercept, or both parameters, and then conduct some other tests on the data. Linear fits are illustrated in Figures 2-4.

5.11. $H_0: \sigma_1^2 = \sigma_2^2$. The preliminary examination suggests that variances differ from group to group. The usual tests of hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2$$

suggests that the conditional variances do indeed differ for recipients and nonrecipient groups. Specifically, we employ the usual tests for equality of variances to find:

First Grade	$F = 3.57$	$df = 58, 155$
Third Grade	$F = 3.48$	$df = 154, 276$
Fifth Grade	$F = 4.91$	$df = 53, 230$

which are each significant (two-tailed F, $p \leq .05$). Variation about the linear regression lines is consistently greater for the program recipient groups. (Note that variance about some other fitted curve may be homogeneous; we consider this possibility below.)

5.12. $H_0: \alpha_1 = \alpha_2; \beta_1 = \beta_2$. If we choose to ignore variance differences in this particular case, we find that the fitted regression lines for recipients differs notably from the line for nonrecipients, regardless of grade. The raw statistics are presented in Table 2. The F statistics based on the null hypothesis given above are:

First Grade F = 8.89 with 2, 213 df;

Third Grade F = 10.69 with 2, 430 df;

Fifth Grade F = 1.52 with 2, 283 df.

We ignore the p levels; they are not as advertised on account of the heterogeneity in variance.

It is clear that slopes within grades differ notably so that effects of a program may not be completely additive as the models here imply. The slope difference for fifth graders is not substantial.

Conditional on the models, using the nonrecipient group as a standard, and ignoring ineligible recipients entirely, we would be forced to conclude the following from this analysis.

First Grade. Students who are low on pretest scores are positively affected by the program; students who are near the cutting point are not affected or affected negatively. This follows from regarding the regression for nonrecipients as a standard, and examining the whole line, not just elevation of the line.

Since slopes differ between groups, using elevation as an indicator of treatment effect is difficult. If we take the mean level of Y of the recipient group and examine it with respect to predicted (from nonrecipient line) we conclude that the effect is positive. If we consider only the elevation of Y for nonrecipients at the cutting point, we must conclude either no effects or negative effects.

Any of these inferences may be wrong due to (a) possible floor/ceiling effects or (b) selection effects on the nonrecipient regression line, or (c) both factors.

Third Grade. Students in the recipient group who are low on pretest scores appear to be negatively affected by the program, if the nonrecipients regression line is taken as a standard. The further away from the cutting point they fall, the worse off they appear. Again, these conclusions follow from considering the whole regression line, not just elevation.

Since slopes again differ between groups, using elevation alone an indicator of program effect is difficult. If we take the mean level of Y for recipients relative to the predicted (from nonrecipients) to estimate effects, we must conclude that the overall program impact is negative. The impact based on prediction at the cutting point is null or negative.

Again, these inferences may be wrong due to (a) ceiling and/or floor effects, or (b) selection affecting the nonrecipient line, or (c) both.

Fifth Grade. Students who are near the cutting point are affected positively or negligibly by the program; those whose pretests are very low are affected negatively. The standard here is the nonrecipient's complete regression line.

Considering elevation only, we conclude that on the average, the mean level of Y is reduced for recipients relative to predictions made from the nonrecipient line. The contrary is true if we focus on the cutting point.

Again, ceiling effects and selection for this data may be critical and may obviate conclusions.

5.13. Within-school Regressions. The pooled linear regression lines for grades 1, 3, and 5 for eligible recipients and nonrecipients are difficult to interpret. If the conventional model and approach is espoused, the program effect would seem to be negative.

One possible problem with the approach is misspecification of the models in the conventional approach. In particular, [illegible] that children are students at some 11 different schools. It is possible that most students in the program recipient group are students of one cluster of schools, and most of those in the nonrecipient group are students in another cluster. If this is the case, even in the absence of any program effect, there may be differences in the regression of pretest on posttest which are a function of school differences rather than program differences. In order to assay the possibility, a within-schools analysis is justified.

The estimates of slope and intercept for each school and for eligible recipients and nonrecipients is given in Table 4 . Also given is the size of the ineligible recipient group.

a. In all but three of the schools, the sample size within the recipient group is marginally adequate. The following inferences stem from the larger sample groups.

b. The slopes for recipient groups always exceed those for nonrecipient groups, suggesting that the misspecification of model because of gross school-related variables is not really the problem. Even within school, slopes differ above and below the cutting point.

c. The higher slope phenomena occurs if there are very few ineligible recipients (as in schools 6 and 8) as well as when there are a number of such recipients (e.g. schools 1 and 9).

d. Items b and c above raise the distinct possibility that the regression line is simply not linear. The nonlinearity might be caused by ceiling effects.

e. Items c and d also suggest that selection may affect slope uniformly and regardless of the number of selectees, i.e. of ineligible nonrecipients. In particular, a peculiar selection strategy used by teachers in assigning ineligible kids to programs may influence the relation between Y and X for the nonrecipients, since points lower on Y and on X are taken out of these groups. It may influence the regression of Y on X for the selected group, and make no different from the regression for the unselected group.

Selection effects cannot influence the eligible recipient group though. All deserving kids get the program.

How can we get a picture of how selection affects the nominal nonrecipient slopes? One option is to compare distributions on X of the ineligible recipients and recipients. With substantial overlap, the effect is likely to be small; with little overlap, the effect could be large. Also, some outside explanation (e.g. by teachers) could be helpful.

f. Finally, it appears that

(i) if a selection effect is operating, it operates for all schools;

(ii) if a ceiling effect is operating, it operates for all schools;

(iii) or both

to produce slope differences.

g. Items a through f assume that the impact of program is strictly additive. If it is not, slopes will change consequence. It must then be nonadditive for each school.

Assuming nonadditivity of effect still makes programs look worse. One would expect an increase slope with more effective treatment and indeed that is what occurs. However, mean levels go down suggesting an overall decline in ability: the better students get better, the worse students get much worse.

5.14. H_0: $\beta_1 = \beta_2$. The slope differences between lines for eligible recipients and nonrecipients are marked within both the third and first grades. The usual F statistic for testing differences between slopes is a crude indicator of the level of that difference:

Third Grade F = 21.5 with 1 and 430.

First Grade F = 17.8 with 1 and 213.

Again, because variances are heterogeneous, the usual alpha levels are not as advertized. If we use the Cochran-Cox approach to testing the difference given heterogeneous residuals, we find these statistics significant.

5.15. Deviations from predicted values at the cutting point. Still another way of appraising the impact of the reading program is to examine performance of program recipients whose pretests scores lay in the vicinity of the cutting point. Presumably, if the program exerts an effect on these individuals and if the regression line for nonrecipients can be taken as a reference, program effects will be reflected by the extent to which actual values of recipients' posttest scores deviate

from predictions based on the regression line.

More specifically, we use as a basis for prediction the regression line based on nonrecipient data:

$$Y_i = \alpha + \beta X_i + e_i \qquad e_i \sim (0,\sigma^2)$$

for which

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

is the prediction equation. We substitute the mean value of X_i for individuals at or near the cutting point into the equation and so estimate the mean Y for this marginal group. To formally assay the difference between a predicted value, $\hat{Y}.$, and an actual value $\bar{Y}.$, we use

$$t = \frac{\bar{Y} - \hat{Y}}{\sqrt{S^2_{\bar{Y}-\hat{Y}}}}$$

with N - 2 degrees of freedom.

Results of testing the hypothesis that predicted and actual values are identical yields no contradictory evidence. The reading programs appear to add nothing to the level of performance of students near the cutting point for first, third, or fifth grades. In particular, the t statistics are:

First	t =	with df =
Third	t = -.048	with df = 276
Fifth	t = .337	with df = 230

Remarks: One reason for scrutinizing recipients whose pretest scores fall near or at the cutting point is the suspicion that these children might receive most attention in a special program. Teachers

might regard them as most promising, more easily rehabilitated or taught, and so forth. This suspicion is reinforced by the interaction effect, i.e. different slopes of the regression lines from group to group. However, the evidence accruing from the tests just described suggests that this "most malleable--best treated" sequence probably does not occur in any grade.

5.16. Deviations from predicted values: Deviations selected a posteriori. Now suppose that program staff can dedicate substantial time and energy to only a few students, given the large number of students in each program. The supposition immediately suggests that, rather than assume that all students labelled as "recipients" got special attention, we should search for outliers, i.e. marked deviations from the regression line. Such outliers might reflect positive effects, when teachers focus much greater attention on a few children; or they might reflect negative effects, when for example teachers ignore some low-calibre students or label them as such.

The third and fifth-grade data do support this view in a limited way. For if we look at deviations from the (control) regression line, we find some significant departures. In the fifth grade a cluster of 7 students with marked positive deviation and a cluster of 5 students with negative deviation yielded the following t ratios (for test of deviation from the line):

$t = 3.64$ with $df = 230$, $\bar{X} = 39.4$, $\bar{Y} = 93.47$;

$t = -4.10$ with $df = 230$, $\bar{X} = 31.28$, $\bar{Y} = 23.80$.

For a cluster of 5 students with positive deviations and a cluster of 9 students with negative deviations, all in the third grade, we have:

$t = 2.27$ with $df = 276$, $\bar{X} = 21.80$, $\bar{Y} = 89.20$;

$t = -5.63$ with $df = 276$, $\bar{X} = 24.44$, $\bar{Y} = 26.67$.

We conclude that at least some students are affected substantially by being in the reading program. A few appear to profit greatly; still others appear to be inhibited markedly by the program.

Remarks. Recall that earlier tests showed that the variance about the (linear) regression line differed from group to group. The outliers chosen here for more intensive examination appear to be important in producing the heterogeneity of variance, those of which they are not the only cause.

Note also that we have attached no particular p value to any of the t statistics above. If the potential deviations had been identified beforehand, then the tests constructed, then p levels would be as advertised in ordinary tables and each statistic would deviate significantly at least at the .025 level (two-tailed). Because the deviations were chosen a posteriori, however, we know that the usual p values are inappropriate. The actual p values are greater than .025 but we are unable to compute them. The test is suspicious in this respect.

5.17. Double extrapolation: Differences between predicted Y's at margin, predictions based on ER, NR, IR, NR + IR. Sween (1971) and Campbell have recommended that in the simplest case one examine predicted values of Y at the cutting point to help establish the existence of an effect. In particular, one predicts a Y from the regression line for the program recipient group (Y/X_{0ER}) at the margin (X_0), and one predicts a Y at the same value of X_0, but using the nonrecipient line (giving Y/X_{0NR}).

We adopt a similar but more elaborate strategy in this section.

Specifically, we compare predictions of Y when predicted values at X_0 are based on equations from

Eligible recipients and nonrecipients

Eligible recipients and ineligible recipients

Eligible recipients and the combination of nonrecipients and ineligible recipients

Algebraically, we examine

$$\hat{Y}_{ER} = \hat{\alpha}_{ER} + \hat{\beta}_{ER} X_0 \quad \text{versus} \quad \hat{Y}_{NR} = \hat{\alpha}_{NR} + \hat{\beta}_{NR} X_0$$

$$\hat{Y}_{ER} = \hat{\alpha}_{ER} + \hat{\beta}_{ER} X_0 \quad \text{versus} \quad Y_{IR} = \hat{\alpha}_{IR} + \hat{\beta}_{IR} X_0$$

$$\hat{Y}_{ER} = \hat{\alpha}_{ER} + \hat{\beta}_{ER} X_0 \quad \text{versus} \quad \hat{Y}_{NR + IR} = \hat{\alpha}_{NR + IR} + \hat{\beta}_{NR + IR} X_0$$

using, where possible, a $\underline{t}$ statistic since the variance of these predicted values can be estimated.

The first comparison is most direct when the design is adhered to perfectly. In the current data, such an estimator is biased to the extent that the nonrecipient sample is biased by selective assignment of ineligible students to the program. The second comparison is of interest in that it can provide us with some information about the effect of selection on regression lines. The third comparision will yield an estimate of the joint effect of treatment and selection.

The result of conducting a test of the equlity of predicted Y and X_0 for each comparison listed above, for each grade level, is unremarkable. In brief, predictions at the margin do not differ. The maximum $\underline{t}$ value of -1.26 (for the first-grade students) is significant at the 20% level; all remaining $\underline{t}$'s are considerably smaller, in the range -.15 to .38.

(Incidentally, the criterion t value is constructed under the prescription given by Cochran and Cox to recognize inequality of variances.)

We infer from all this that the hypothesis of "no program effect" at the cutting point is a tenable one for each grade. Again, the analysis is based on linear models and recognizes no effect on ceiling or floor on test scores.

5.18. An Approach to Analysis which Recognizes Ceiling Effects.

It's clear that ceiling effects can complicate interpretation of RD data. Indeed, ceiling effects, if unrecognized, can lead to analyses which make program effects look harmful when in fact they are negligible (Appendix I). This section offers a tentative approach to data analysis which recognizes ceiling effects and avoids biases in estimates of program effect.

Consider Figure X.1, which represents a null condition. The dotted lines represent fitted regressions; the solid lines represent the relation between posttest and pretest, including a ceiling on posttest. The vertical line again represents the cutting point.

The display emphasizes that in principle, at least, the symptoms of negligible effects are:

(a) small negative difference, $\bar{Y}_R - \bar{Y}_{NR}$, between means (projected

or actual) of R and NR groups at the cutting point.

(b) greater slope in the R group than in the NR group;

(c) lower bound (floor) for the R group is chance-level score.

Note also that the relative slopes of R and NR are predictable under this null condition, if the point of discontinuity in the true regression can be identified and a few assumptions are made. That is, under null conditions, the slope for R and a segment of the true slope for NR will be identical; one is observable (R) and we may assume the segment for NR is the same as that for the R. Given this information and some reasonable guess as to the point of discontinuity (i.e. ceiling) and a few simplifying assumptions, it is a matter of algebra to compute an estimate of the complete regression line for the NR group. If this algebraic estimate differs much from the observed line, then we will know that there is some inconsistency, i.e. that the null condition is not fairly represented by the data.

Consider Figure X.2, which represents a situation in which there is a notable treatment effect exerted at least at the cutting point. In this case, it is possible to discriminate between H_0 and H_A by verifying that the difference between (projected or actual) values of $\bar{Y}_R$ and $\bar{Y}_{NR}$ at the cutting point is positive. It is a weak test in the sense that a substantial negative difference characterizing the null condition may have been overcome by the program to produce a positive effect.

It <u>may</u> be possible to examine $\bar{Y}_R$ relative to an estimated slope segment (solid line) for the NR group. That is, given the observed slope for the NR group, given a reasonable guess as to where the

flattening begins, and a simplifying assumption, it would be possible to compute an algebraic estimate of the true line segment's slope. Comparing the observed $\bar{Y}_R$ to the project of the line segment at the cutting point would produce a more powerful test, but one which is likely to be imprecise.

Figure X.3 illustrates a situation in which the program effect is additive, that exerted uniformly along the full range of pretest scores in the R group. One symptom here is again the elevated position of $\bar{Y}_R$ relative to $\bar{Y}_{NR}$ at the cutting point. A second symptom is the closer match between $\bar{Y}_R$ computed at the midpoint of X_R and the projection that Y based on the regression line computed from the NR group. Again both tests are weak, the first being weak for the reason described in the preceding paragraph. The second is weak because the vertical distance between $\bar{Y}_R$ and an estimated $\hat{Y}_R$ estimated from the NR group at $\bar{X}_R$ depends on the magnitude of the ceiling effect. If most members of the NR group scored at or near the ceiling, that distance would be appreciable unless the treatment effect was quite large.

Figure X.4 illustrates a situation in which treatment effects are exerted only in the lower range of the X variable. The symptom of an effect is elevation of posttest scores for children with low scores on the pretest; the elevation is above chance level scoring, which is one standard for an optimistic (nonconservative test). The slopes are more informative in that if the slope for R is less than that for the NR group, it must follow that the program exerted an influence on low-scoring children. There is no other competing algebraic interpretation, though there may be competing empirical ones.

Figure X.5 represents a situation in which program effects are

strongest for children who score low on pretest and effects are weakest for children scoring high on the pretest. Two symptoms determine the inference. First, the (projected or actual) $\bar{Y}_R$ at the cutting point equals or exceeds $\bar{Y}_{NR}$. Second, the slope for the R group equals or is less than the slope for the NR group. Third, the intercept at X = 0 for the R group regression line equals or exceeds the intercept for the NR group.

The symptoms of negative program effects are demonstrated in Figure X.6. Here the general elevation of the regression is reduced relative to what it would be under null conditions. Estimating what the null condition would be is again possible only if the point at which ceiling effects begin can be guessed at. The same perspective can be used to examine the possibility of negative effects occuring only for the most able children (Figure X.7).

Negative effects on the least able children will be no more detectable especially unless floor effects become influential. From Figure X.8, it's evident that such negative effects will be demonstrated by steep slope for the R group relative to the NR group and a negative (projected or actual) difference $\bar{Y}_R - \bar{Y}_{NR}$ at the cutting point.

Unless one tries to estimate null regression lines using a plausible (guessed) value of ceiling, it is impossible to discriminate between this negative effect and null conditions. It may not be worth the trouble of other evidence or theory suggests that negative program effects are implausible.

5.2 Ineligible Program Recipients

Recall that the original design plan was not carried out completely. Ineligible students did receive services. To understand the implications for analysis, we need to determine how closely the ineligible recipients resemble the other groups.

The ineligible program recipients are much more similar to program nonrecipients than they are to eligible program recipients. Nonetheless, the difference between the ineligibles and the nonrecipients is still notable and consistent from grade to grade.

In particular, pretest means for ineligibles are consistently smaller than, but close to, pretest means for nonrecipients in the first, third, and fifth grades (Tables 1-2). Posttest means show a similar pattern. The differences are significant in each case (using the Cochran-Cox test for equality of means with unequal variances).

Variance of pretest and posttest scores also differ across ineligible recipients, nonrecipients, and eligible recipients of the program. Again, the ineligible recipients and nonrecipients are most similar with respect to variability of scores; however, the differences are still significant and in the same direction regardless of grade level. The variance of pretest scores of ineligible recipients is, except for the first grade, always smaller than the variance for the nonrecipient group; evidently, the ineligibles are being selected on implicit teacher criteria such that they constitute a more homogeneous group. The posttest scores for ineligible recipients always exhibit more variability than the nonrecipients' scores probably because of ceiling effects on the latter.

A visual inspection of the linear regression parameters for the various groups again suggests that ineligible recipients resemble the nonrecipients more closely than they resemble the eligible recipients (Tables 3 and 5). It is clear that in the first grade, however, regression lines for recipient and nonrecipient groups differ, primarily with respect to elevation (the slope differences is small). For the third and fifth grades, the differences are very small. Tests of the hypothesis that the intercept and slope are identical yield F ratios in the 3-8 range, but residual variances differ a bit; residual variances for the ineligibles are consistently about half again as large as the residuals for the nonrecipient data, so the conventional test's alpha level is not as advertised.

The IR, NR, and ER groups differ with respect to simple descriptive parameters and regression parameters. Despite the close (visual similarity of the IR and NR groups, they too differ from one another and from grade to grade. Given that the IR and NR groups are "comparable" in the sense that they resemble one another, and that one group receives the program while the other does not, one might think that a covariance approach might be used to estimate the program's effect. The approach is inappropriate, however, in part because required assumptions about homogeneity of slope and residuals do not hold, and also because the underlying models are almost by definition misspecified. That is, something other than pretest scores is being used as the basis for assigning ineligible students to the program.

Now despite the IR-NR differences, we might choose to pool these data on grounds that any evident differences are entirely a function of

the selection process rather than of any "treatment" effect. Doing so, may permit us to make more powerful tests. If indeed the treatment does influence the IR group, then tests which compare these data with ER data are likely to be weak.

Tests of the hypothesis that residual variance about the regression lines for the eligible recipient group are equal to residual variance for the combined nonrecipient and ineligible recipient group result in the following F ratios for

$$H_0: \quad \sigma_{Y \cdot X}^2 \ (ER) = \sigma_{Y \cdot X}^2 \ (NR + IR):$$

First Grade, F = 2.11 with 58 and 285 degrees of freedom;
Third Grade, F = 2.89 with 54 and 327 degrees of freedom;
Fifth Grade, F = 4.11 with 53 and 299 degrees of freedom.

The F statistics suggest that residuals differ notably, and that variance for eligible recipients is consistently greater than residual variance for the combined NR and IR group. The implications are many and complicated. Heterogeneity of variances may be induced by:

- Program effects on recipients and, to a lesser extent, on the combined group;
- Poor fit of a linear regression to one or both categories of data, i.e. above and below the cutting point;
- A natural relation between variance of observations and the X variable;
- Other reasons.

If we ignore the heterogeneity and proceed to apply a conventional

F test for

$$H_0: (\mu, \gamma)_{ER} = (\alpha, \beta)_{NR + IR}$$

we obtain relatively large F statistics for grades 1 and 3, and an incommunicable F for the fifth grade.

First Grade: F = 10.44 with 2 and 343 degrees of freedom;

Third Grade: F = 7.[illegible] with 2 and 481 degrees of freedom;

6. Results of Fitting Quadratic Functions

First Grade (Figure 5)

The collinearity of X and X^2 make this analysis useless. Within the recipient group and the nonrecipient group, correlations between X and X^2 exceed .99. Graphical results suggest that a quadratic is not a good fit. The interpretation based on the quadratic fit is nonsense.

Third Grade (Figure 6)

The collinearity problem makes this analysis useless. The graphical results appear more sensible. Interpretation suggests that the program is harmful, a conclusion we do not accept at this stage.

Fifth Grade (Figure 7)

Again, collinearity of X and X^2 make this analysis suspect. The graphical results are a bit more sensible looking than the preceding two analyses. Moreover, the chart suggests that the program exerted a positive effect. Again, we do not accept this conclusion for the fifth grade on the basis of this evidence alone.

Evidently, fitting a quadratic does not solve the problem of ceiling effects at all. The fifth-grade data look decent, but the first and third-grade data are confusing. The ineligible recipients may account for the peculiar results.

7. Summary

The point that is made in this paper does not need belaboring. It would appear that the work shown in using regression discontinuity evaluation at a local educational agency is monumental. It is realized, as was pointed out in the presentation by Northwest Regional Educational Laboratory when they presented the models in Arizona, that forms will be designed so that the evaluator at the LEA can simply take from box 3, place in box 6, multiply by box 7, etc. It is difficult to accept that such a process will produce the desired results. If the evaluators at the LEAs are given only those directions and asked to use a technique such as the one presented in this paper, it would be hard to believe that the resulting information would be much better than that Talmadge found in his study in 1974. It would appear from this example that not only is the technique very difficult to implement, but also to interpret. Also, criterion-referenced data by its very nature may not lend itself and certainly does not lend itself easily to the Regression Model. Criterion-referenced tests are designed so that a high proportion of the students can attain mastery at the completion of the program. In the case of the Mesa Public Schools the underlying program is designed around a minimal set of objectives for that specific grade level. It is felt that these are the most important objectives and are the objectives that all of the students should accomplish by the end of the year. Enrichment above this minimal set of objectives is then left up to the classroom teacher. It is realized certainly that the better students go far beyond this minimal set of objectives and in fact may enter a given grade level with these set of objectives mastered. It is, then, up to the teacher through other prescriptive methods to determine an individualized course of instruction for that student. However, because of

this we see that a high proportion of the students do attain a high degree of mastery by the end of the year.

The authors would again like to re-emphasize their commitment to comprehensive and valid evaluation at the local educational agency level. They would also like to acknowledge the significant efforts and work by Mr. Talmadge. As Dr. Boruch stated at the Conference On Minority Group Testing:

> "Mr. Talmadge has attacked an enormously complicated problem with energy and with an awareness of some lessons hard won during the past few years. We admire his fortitude in doing so. The paper itself touches on many of the techniques which have produced biased results in evaluative social research and so deserves recognition for its scope and tentative style. The paper also represents an improvement over the practices of many school districts and contracts in the matter of estimating the impact of educational programs. With a few remarkable exceptions those practices have resulted in a dismal array of evaluation reports and findings which badly undercut rather than enhance school districts' own efforts to improve education in difficult settings."

In conclusion the authors would like to reiterate their concerns for the implementation of the selected evaluation models at the LEA level. This is not to say that the comprehensive evaluation should be precluded at this level, but only that there is a need for valid and reliable information concerning the evaluation of compensatory education from the local school districts.

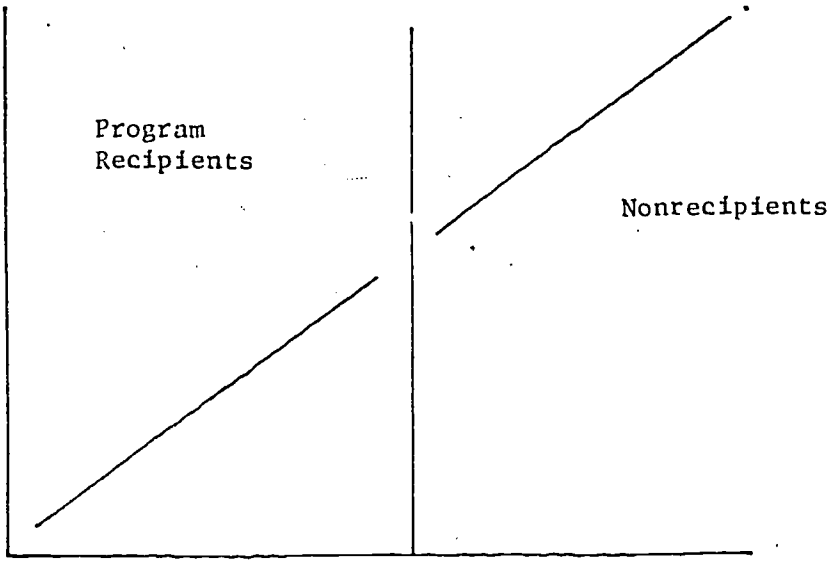


Figure 1b. Null Conditions

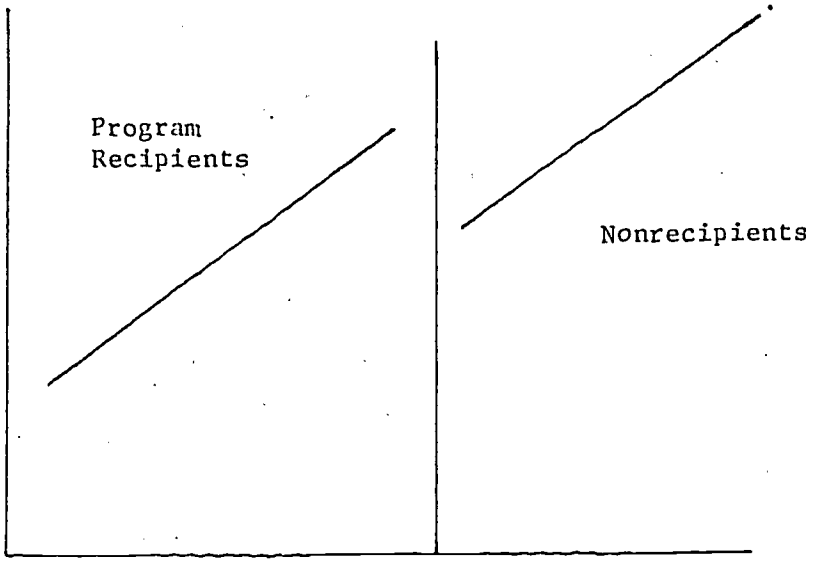
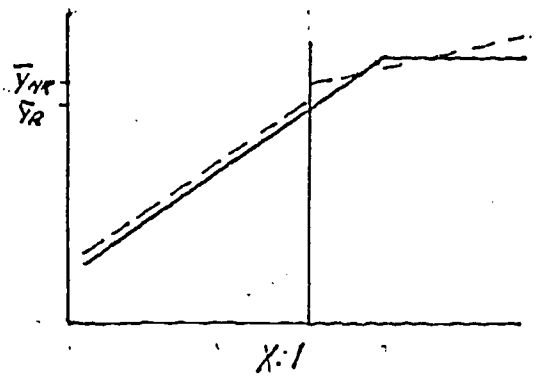
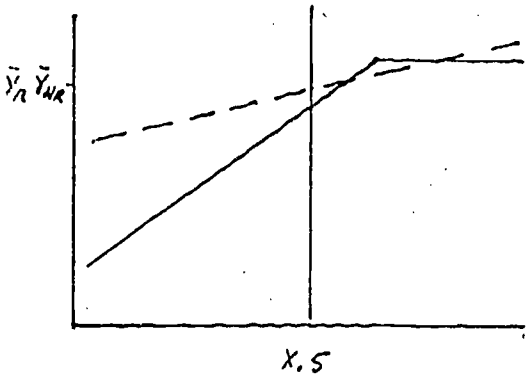


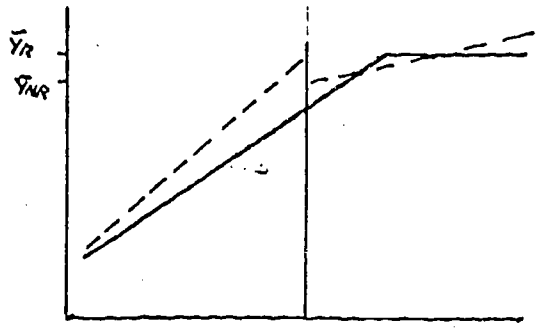
Figure 1b. Treatment Effect, Strictly Additive



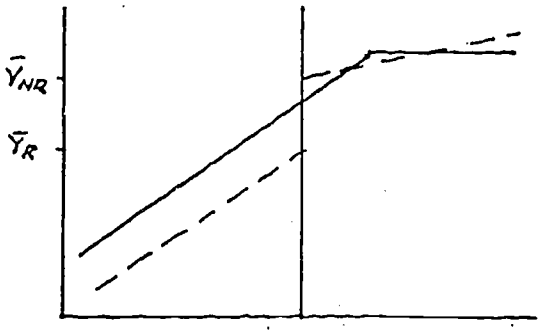
X.1



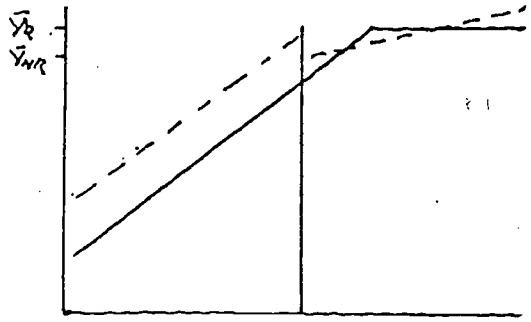
X.5



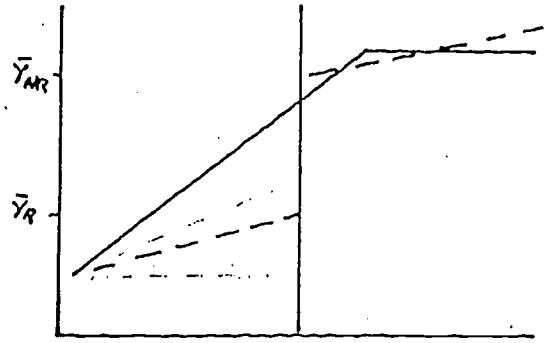
X.2



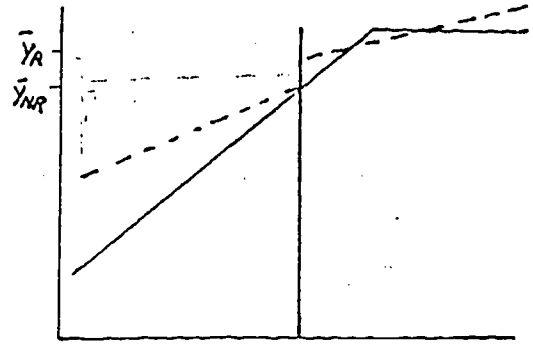
X.6



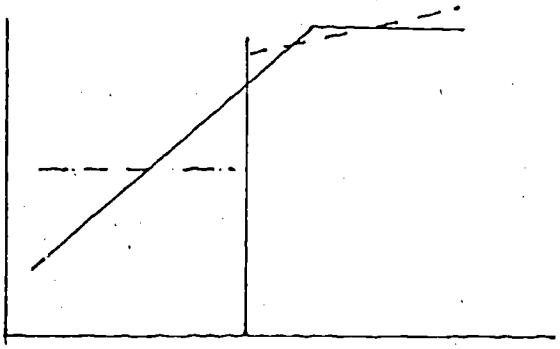
X.3



X.7



X.4



X.8

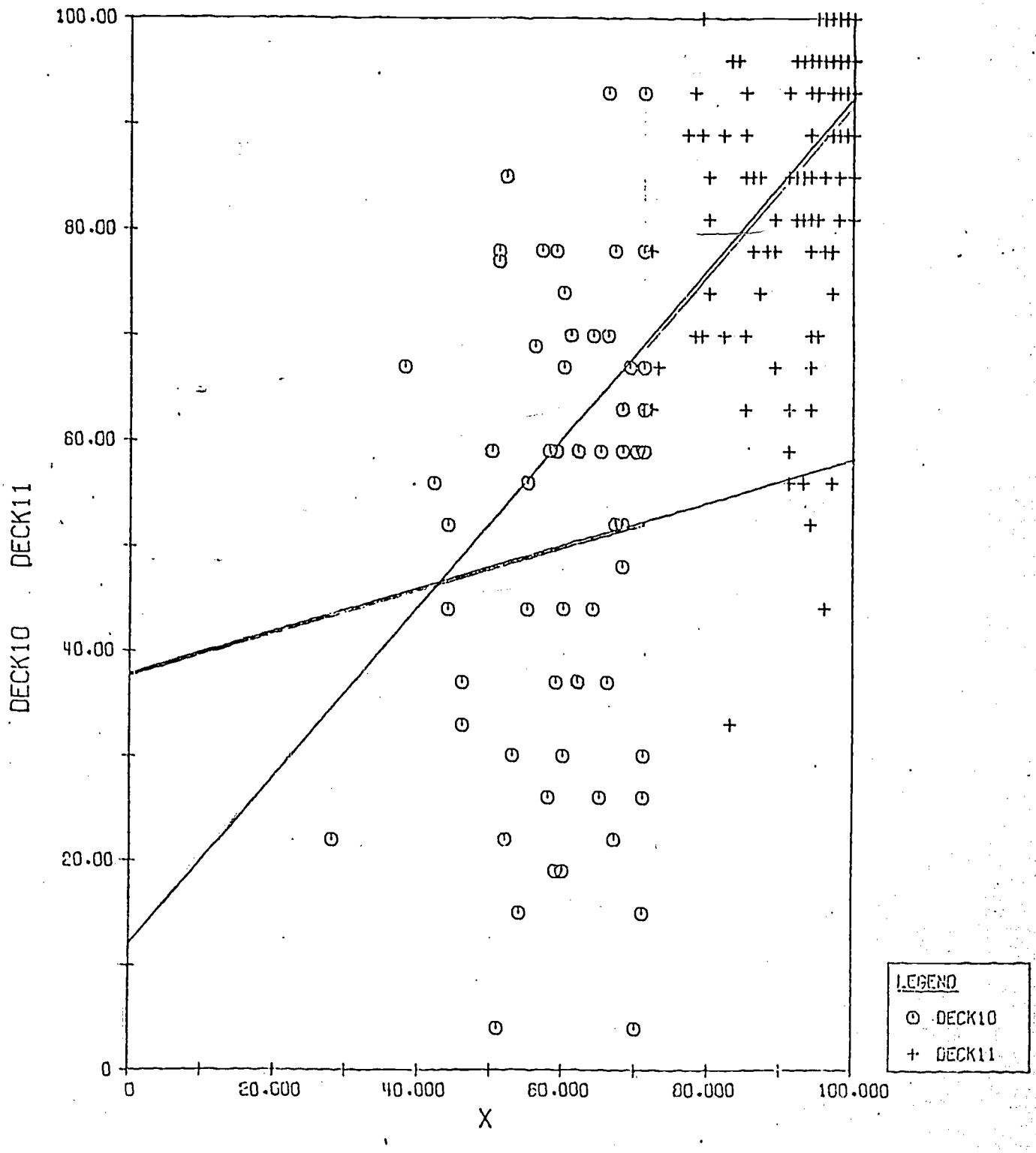


Figure 2. Posttest regression on pretest for First-grade students, Eligible recipients (0), and Nonrecipients (+), with nominal cutting point of X = 71.

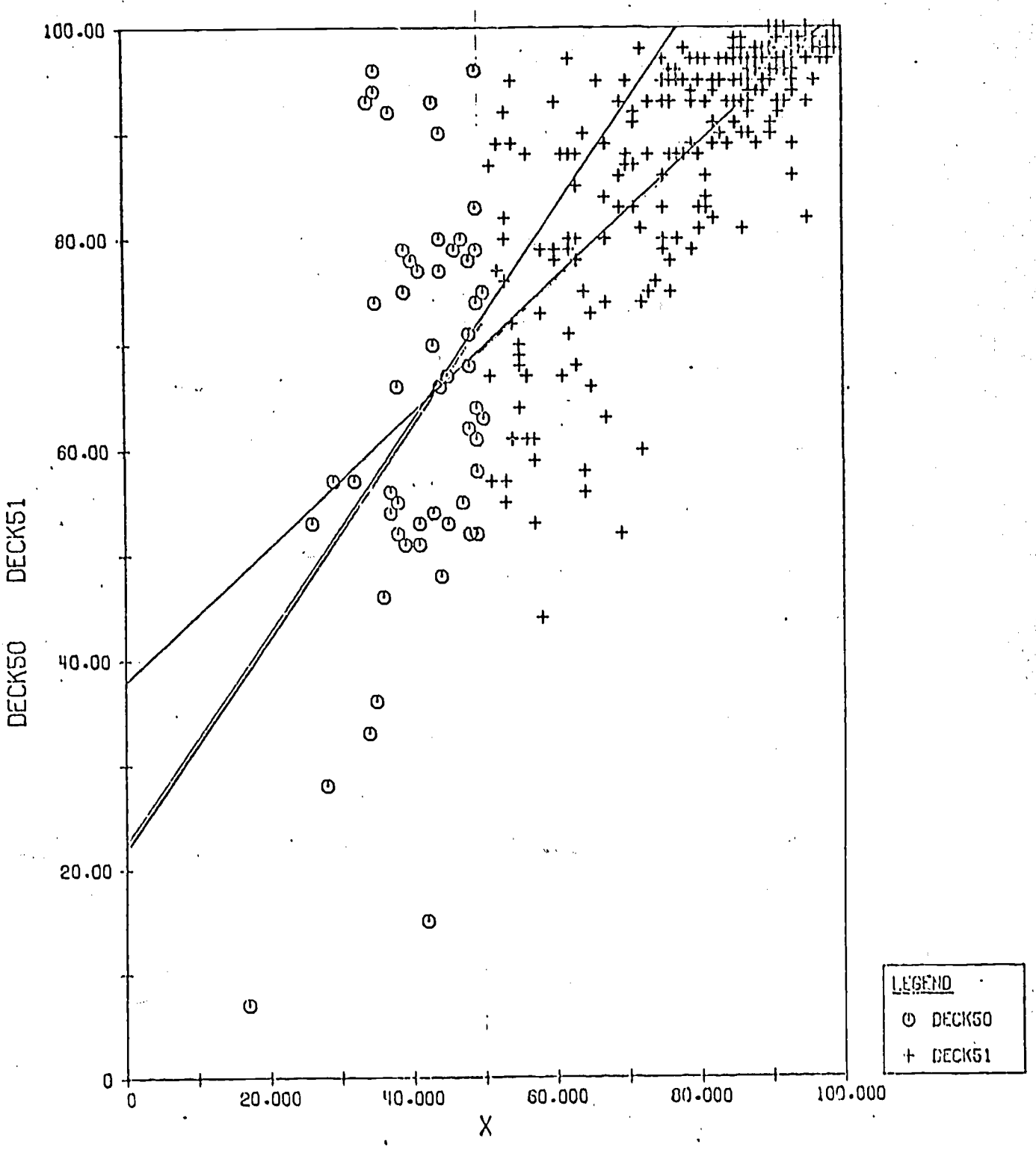


Figure 4. Posttest regression on pretest for Fifth-grade students; eligible recipients (0), and nonrecipients (+), nominal cutting point at X = 50.

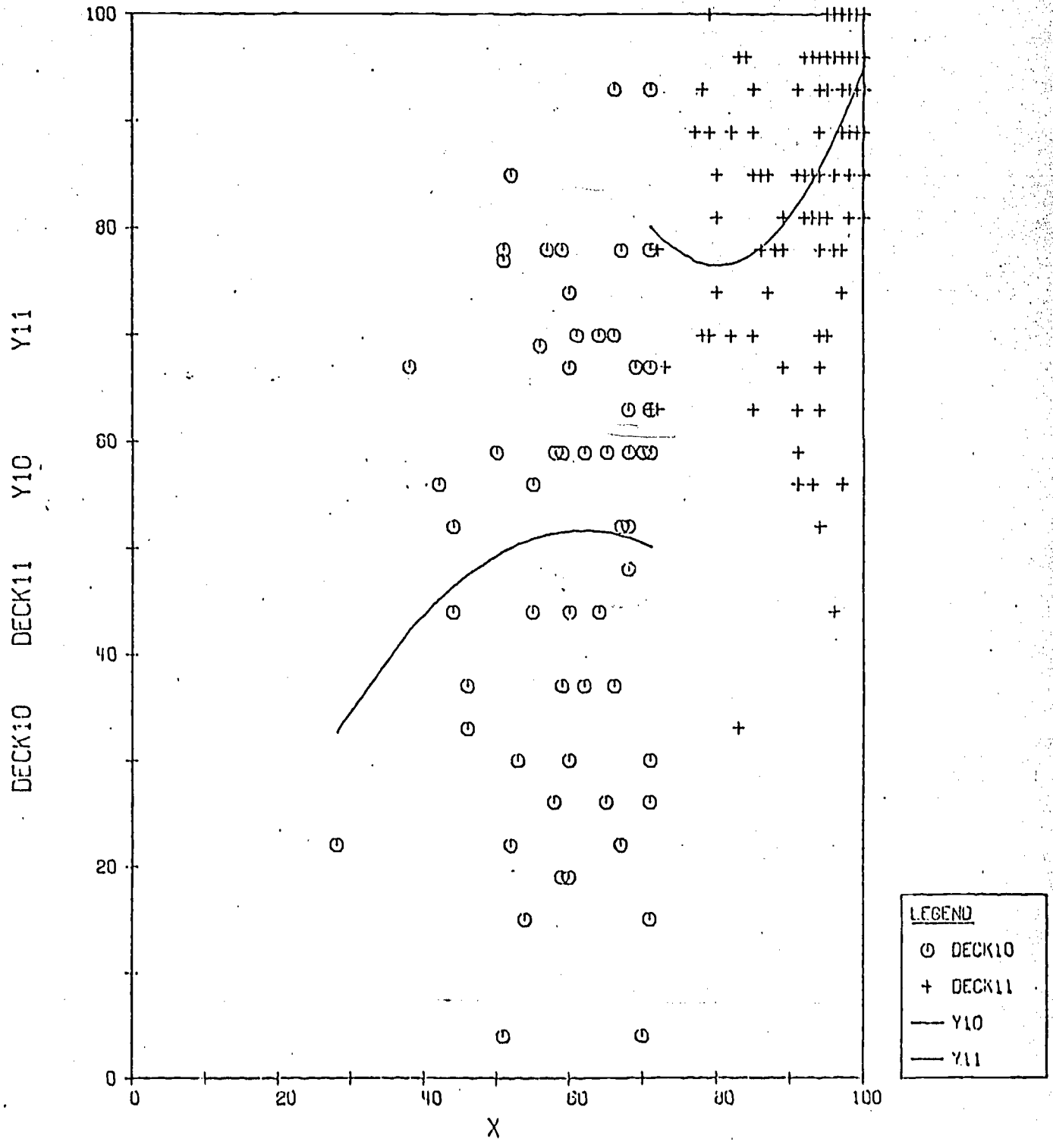


Figure 5. Quadratic regressions of posttest on pretest for First-graders.

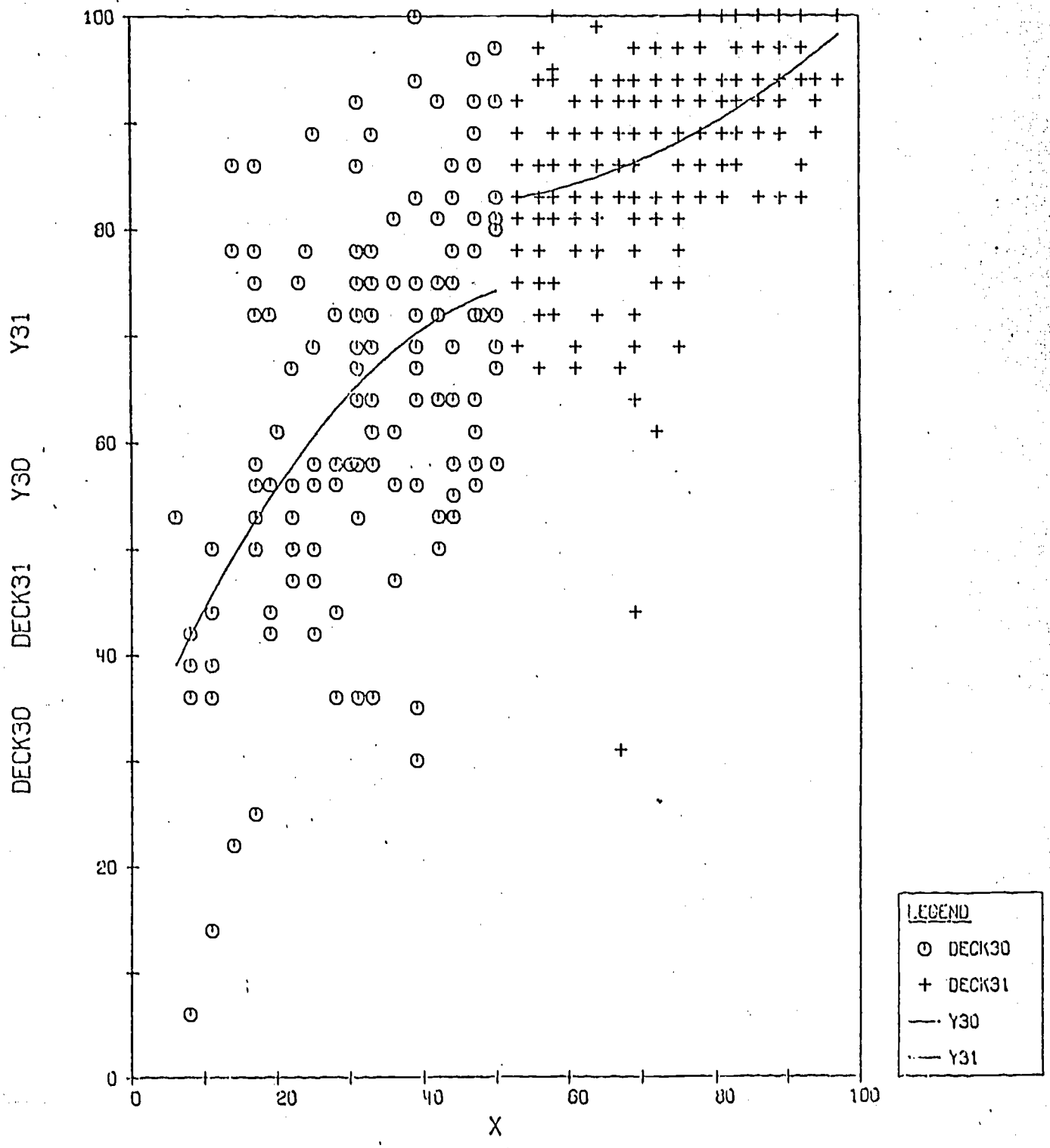


Figure 6. Quadratic regression of posttest on pretest for Third-graders.

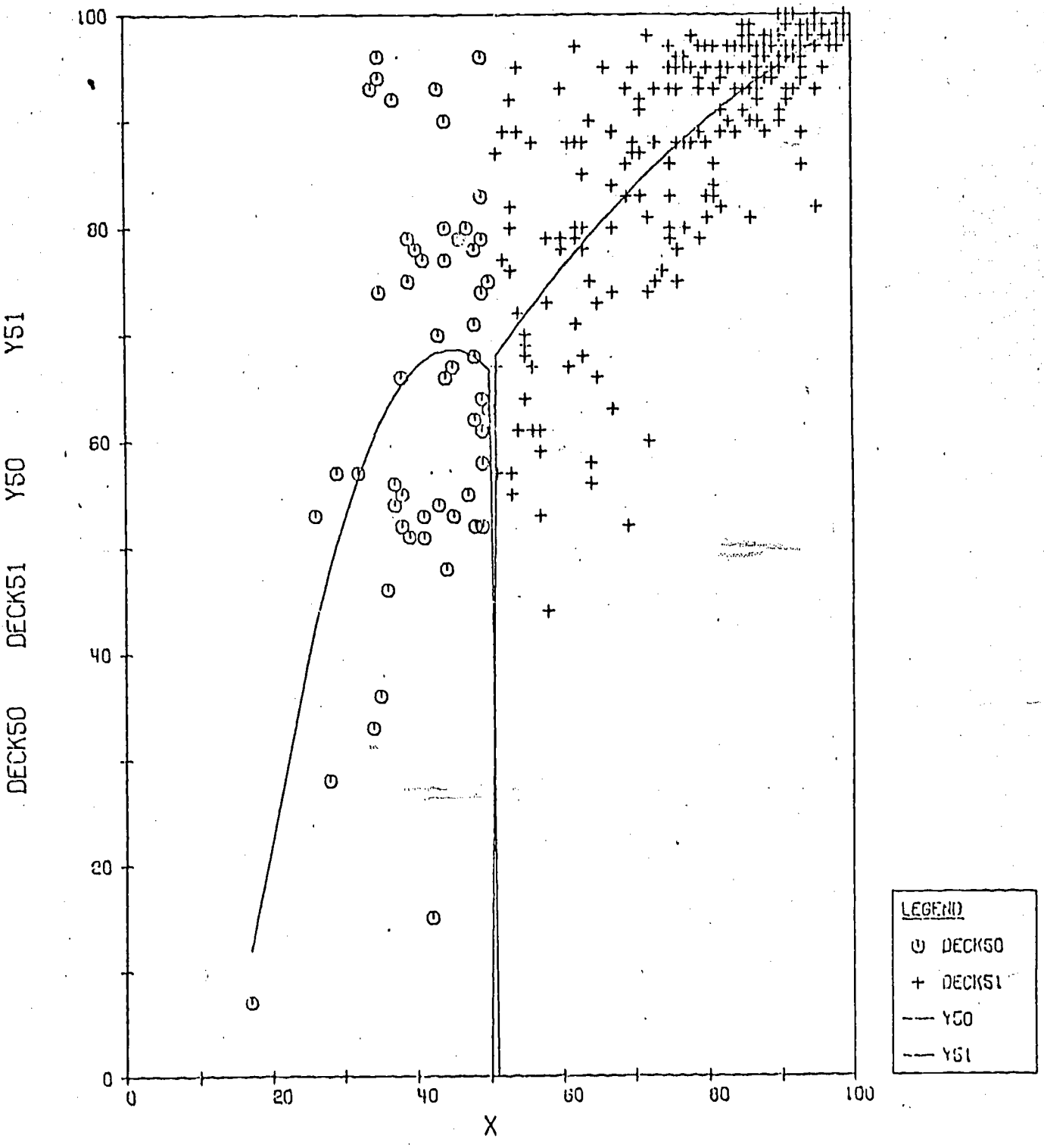


Figure 7. Quadratic regression of posttest on pretest for Fifth-graders.

Table 1

Descriptive Statistics for Scores of First, Third, and Fifth Grade Children on Reading Tests

First Grade		N	Mean	Variance	Skewness	Kurtosis
Eligible Recipients	x	60	60.12	91.77	-.98	.79
"	y	60	50.05	511.20	-.19	-.88
Nonrecipient	x	157	93.63	48.31	-1.45	1.32
"	y	157	87.43	174.86	-1.58	2.28
Third Grade						
Eligible Recipients	x	156	33.36	144.54	-.43	-.88
"	y	156	65.07	291.56	-.55	.43
Nonrecipient	x	278	74.51	134.03	-.13	-.94
"	y	278	88.66	77.03	-2.05	8.3
Fifth Grade						
Eligible Recipients	x	55	41.51	50.11	-1.02	1.14
"	y	55	64.26	369.23	-.59	.62
Nonrecipient	x	232	78.13	185.32	-.46	-.97
"	y	232	87.91	141.67	-1.40	1.25

Table 2

Descriptive Statistics for Ineligible Recipients

		N	Mean	Variance	Skewness	Kurtosis
First Grade	X	130	85.4	72.52	.08	-1.32
	Y	130	70.52	349.93	-.70	.21
Third Grade	X	51	61.94	78.42	1.23	.80
	Y	51	81.26	157.15	-.92	1.11
Fifth Grade	X	70	70.14	151.75	.20	-1.15
	Y	70	82.04	159.49	-.97	.67

Table 3

Estimates of Parameters in Linear Model ($Y = \alpha + \beta X + e$, $e \sim I(0,\sigma^2)$ Fitted to First, Third, and Fifth-Grade Students

	$\hat{\alpha}$	$\hat{\beta}$	R	$SE(\hat{\alpha})$	$SE(\hat{\beta})$
First Grade					
Recipients	37.80	.20	.09	18.79	.31
Nonrecipients	12.05	.81	.42	12.99	.14
Third Grade					
Recipients	40.85	.73	.51	3.49	.10
Nonrecipients	63.77	.33	.44	3.09	.04
Fifth Grade					
Recipients	22.44	1.00	.37	14.58	.35
Nonrecipients	37.90	.64	.73	3.11	.04

Table 4

Estimates of (Linear) Regression Parameters for

within-School Analyses

	N	$\hat{\alpha}$	$\hat{\beta}$	$\sqrt{\hat{\sigma}^2_{Y \cdot X}}$
50-1	9	18.83	.99	16.17
51-1	46	58.99	.78	7.61
52-1	12	-----	---	-----
50-2	3	133.67	-1.45	17.41
51-2	63	58.77	.42	6.34
52-2	7			
50-5	3	36.15	.55	7.48
51-5	43	18.13	.87	9.24
52-5	1	-----	---	-----
50-6	7	-32.03	2.22	17.47
51-6	11	18.91	.85	6.60
52-6	1	-----	----	-----
50-8	9	8.57	1.46	18.69
51-8	27	22.83	.82	10.33
52-8	2	-----	----	-----
50-10	13	39.33	.62	16.34
51-10	12	55.40	.51	12.88
52-10	8	-----	---	-----

Table 5

Linear Functions Fitted to Data on Ineligible Program Recipients

	$\hat{\alpha}$	$\hat{\beta}$	R	SE($\hat{\alpha}$)	SE($\hat{\beta}$)
First Grade	11.84	.69	.31	15.83	.18
Third Grade	50.80	.49	.35	11.86	.19
Fifth Grade	39.85	.61	.59	7.17	.10

Table 6

Linear Functions Fitted to Data on Combined Eligible Program Recipients and Nonrecipients

	$\hat{\alpha}$	$\hat{\beta}$	R	$\hat{\sigma}_{Y \cdot X}^2$	$\bar{X}$	σ_X^2
First Grade	-12.98	1.03	.50	244.44	89.9	75.9
Third Grade	59.70	.38	.47	75.01	72.6	145.9
Fifth Grade	34.21	.69	.74	78.88	76.0	207.0

Appendix: Effect of Ceiling on Observation of Y

in RD Analysis: Graphical Interpretation

1. In the simplest case, Y above a certain level will be unmeasureable, and because all units are then assigned the maximum score, we have

$$Y = \alpha + \beta X \qquad \text{for } Y < Y_0$$

$$Y = Y_0 \qquad \text{for } Y \geq Y_0$$

as in Figure 1. What we observe is a discontinuous regression line, actually two lines.

2. If, in the case just described, one tried to fit a single line to the observations, the slope and intercept estimator would appear as in Figure 2. The fitted line

$$Y = \alpha' + \beta' X$$

would be such that

$$0 < \beta' < \beta \quad \text{and} \quad \alpha < \alpha' < Y_0$$

3. Now suppose further that there is some point on the X axis which for theoretical or design reasons is thought to define two sections of the data to which different regressions must be fitted. For example, the cutting point in Figure 3 might indicate the separate sets of points to which regressions must be fitted in a regression discontinuity analysis. Again, if one ignores or does not recognize the ceiling effects, fitting the regressions to the left side will yield

$$Y = \alpha' + \beta' X$$

where $0 < \beta' < \beta$ and $0 < \alpha' < Y_0$, and for the right hand side,

$$Y = \alpha'' + (0)X = Y_0$$

In general, there will be a gap between the end of the line $Y = \alpha' + \beta'X$ and the line $Y = \alpha'' = Y_0$. The greater the latter distance between cutting point and point of natural unrecognized discontinuity, the greater the gap.

4. Figure 4 illustrates the consequence of phasing the cutting point below rather than above the point of natural inflection in the line, i.e. below the discontinuity produced by the ceiling. In this instance, below the cutting point we have fitted

$$Y = \alpha' + \beta'X$$

where $\alpha' = \alpha$ and $\beta' = \beta$, and above the cutting point we have fitted

$$Y = \alpha'' + \beta''X$$

where $0 < \beta'' < \beta$ and $0 < \alpha'' < Y_0$. Again, there is a gap between lines at the cutting point, this time showing the right-hand curve at a higher level.

5. The point of all this is that a ceiling effect, if unrecognized, will produce biases in estimates of slope and intercept parameters.

The consequence of this in a regression discontinuity analysis can be dangerous. Suppose, for example, that all the diagrams really reflect only null conditions, and curves are fitted as in Figures 3 and 4. The inference one would draw from fitted lines in Figure 3 (if the left-hand side represents program recipients) is that the program harmed its recipients since (a) average elevation of the left-hand side is depressed at the average X and at the margin relative to the right-hand line, (b) the slope

increased as a consequence of treatment C when in fact no such change occurred).

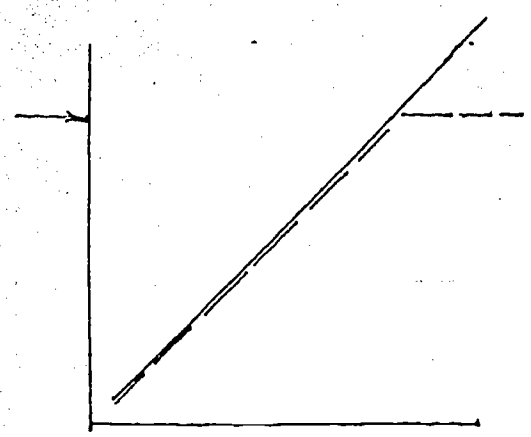


Figure 1

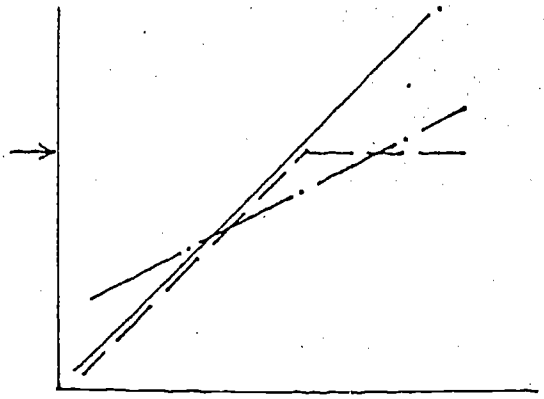


Figure 2

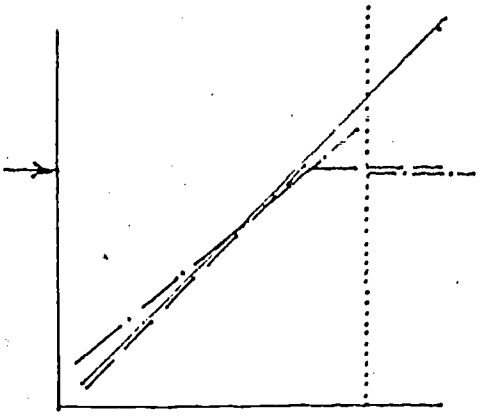


Figure 3

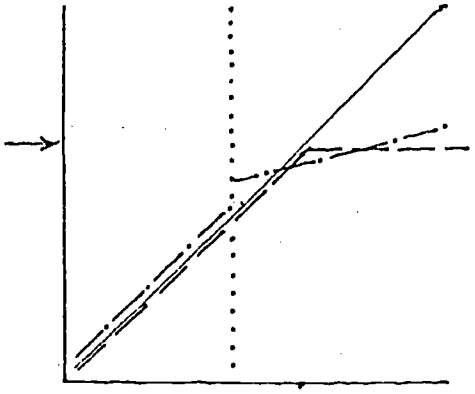


Figure 4

LEGEND

——— True Relation Between Y & X

——·—— Observed Relation Between Y & X (with ceiling)

——— Fitted Regression of Y on X

····· Cutting Point on X Axis, Regressions Fitted Above and Below Cutting Point

→ Ceiling on Y, Y Axis

FIGURES FOR APPENDIX

References
Articles Relevant to
Regression Discontinuity Designs

Battacharyya, G. K. and Johnson, R. A. Nonparametric test for shift at an unknown time point. Annals of Mathematical Statistics, 1968, 39,

Borich, G. D. Interactions among groups regressions: Testing homogeneity of group regressions and plotting regions of significance. Educational and Psychological Measurements, 1971, 31, 251-253.

Boruch, R. F. Headstart evaluation: Use of regression-discontinuity and randomized experiments. Evaluation Research Report NIE-05a/memo, Evanston, Illinois: Psychology Department, Northwestern University, 1974.

Boruch, R. F. Regression-discontinuity designs revisited. Evaluation Research Report NIE-05b/memo, Evanston, Illinois: Psychology Department, Northwestern University, 1974. (Presented at the Annual Meetings of the American Educational Research Association, Chicago, May 1974)

Farley, J. V. and Hinich, M. J. A test for a shifting slope coefficient in a linear model. Journal of the American Statistical Association, 1970, 65, 1320-1329.

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

Gulliksen, H. and Wilks, S. S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

Hellmuth J., Ed. Compensatory Education: A National Debate, 3. Disadvantaged Child. New York: Brunner/Mazel, 1970.

Horst, Donald P., Tallmadge, G. Kasten, and Wood, Christine, T. A Practical Guide to Measuring Project Impact on Student Achievement. U.S. Government Printing Office. Washington. 1975.

Pearson, E. S. Appendix to paper by B. H. Wilsdon, Journal of the Royal Statistical Society, Supplement 1, 1934, 178-192.

Quandt, R. E. Tests of the hypothesis that a linear regression system obeys two separate regimes. Journal of the American Statistical Association, 1971, 55, 324-330.

Quandt, R. E. New approach to estimating switching regressions. Journal of the American Statistical Association, 1972, 67, 306-310.

Rattner, D. D. The effect of early childhood education programs on school readiness. Ph.D. Dissertation. Northwestern University. Evanston, Illinois, 1975.

Smith, H. Fairfield. Effect of spacing and time of sowing on yield and yield and yield components of wheat varieties. Journal of the Royal Statistical Society, Supplement 4: 1937, 177-178.

Smith, H. Fairfield. Relationships among characters observed in area clearance tests. Biometrics, 1951, 7, 185-199.

Snedecor, G. W. and Cochran, W. G. Statistical methods. Ames: Iowa State, 1967, p. 432-436.

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

Welch, B. L. Some problems in the analysis of regression of k samples of two variables. Biometrika, 1935, 27, 145-160.

Wilson, J. W. and Carry, L. R. Homogeneity of regression--its rationale, computation, and use. American Educational Research Journal, 1969, 6, 80-89.