Meta-analysis of several hundred evaluations of Title I compensatory education programs shows that two distinct research designs consistently yield different results. The norm-referenced model portrays programs as positively effective while the regression-discontinuity design shows them to be ineffective or even slightly harmful. Three potential biasing factors are discussed for each design—residual regression artifacts; attrition and time-of-testing problems in the norm-referenced design; and assignment, measurement, and data preparation problems in the regression-discontinuity design. In lieu of more definitive research the tentative conclusion is that in practice the norm-referenced design overestimates the program effect while the regression-discontinuity design underestimates it.

METHODOLOGICALLY BASED DISCREPANCIES IN COMPENSATORY EDUCATION EVALUATIONS

WILLIAM M.K. TROCHIM
Cornell University

The complex nature of social phenomena and the inherent limitations of available research methodologies suggest that definitive conclusions about the effectiveness of social programs can best be approached through multiple replications of evaluative studies. However, even when multiple replications exist, discrepancies in their results can act to prevent generalizations across studies. This article discusses a conflict in the results obtained from a large number of individual evaluations which appears to be best explained by methodological factors. The conflict arises in compensatory education evaluation and, specifically, in the evaluation of programs funded under Title I

AUTHOR'S NOTE: *This work was supported by NIE Grant G-81-0086 (William Trochim, Principal Investigator), NSF Grant DAR 7820374 (Robert Boruch, Principal Investigator), NIE Grant G-79-0128 (Robert Boruch, Principal Investigator), and NSF Grant BNS 792-5571 (Donald T. Campbell, Principal Investigator).*

EVALUATION REVIEW, Vol. 6 No. 4, August 1982 443-480

of the Elementary and Secondary Education Act of 1965. When individual estimates of program effect are averaged by research design across studies of similar Title 1 compensatory education programs, the results for one commonly used research design portray the programs as positively effective while the results obtained using another design show the programs to have a zero or slightly negative effect.

This pattern of results raises several important issues. Perhaps most obvious is the question of why the discrepancy occurs at all. Certainly the results cast doubts on any notion that the two research methods in question can be considered "equivalent" or can be expected to yield "equivalent" results. Assuming that the programs studied are fairly similar, it seems reasonable that there might be definable methodological factors associated with the use of each design which help explain why they yield different results.

The identification of design-related bias is a meta-evaluative issue worthy of consideration. Ideally, it would be best to vary designs and uses of the designs systematically across studies so that design components which lead to discrepancies in results could be more definitively identified. This has not been done in this context, however, and given typical financial and temporal restraints, is not likely to be practically feasible in many other settings. A more reasonable approach to this meta-analysis problem involves the post hoc examination of studies in the hopes of identifying major design factors related to conflicting results. In addition to this meta-evaluative problem are issues involving the interaction of substantive and methodological theory. One such issue concerns which of the two designs, if either, should be given more importance in the interpretation of the effectiveness of the programs in question. Another problem arises because design-related bias can act as a source of conflict between the policy-making and evaluation communities. Certainly the credibility of evaluation to some extent rests on the ability of evaluators and methodologists to isolate design factors which lead to conflicting results.

This article discusses several of the most important likely methodological reasons for the discrepancy in results observed in this research context. Some of these explanations are related to general characteristics of the designs themselves while others are more germane to the use of the designs within the specific context of compensatory education. The remainder of this article involves a brief description of the research context and the designs involved, the presentation of evidence for the discrepancy in the results, and a discussion of several of the more plausible potential explanations for this discrepancy.

THE RESEARCH CONTEXT: TITLE I OF ESEA

The results discussed here are set in the context of the largest single federal compensatory education program—Title I of the Elementary and Secondary Education Act (ESEA) of 1965. In 1979, over $5.5 billion was authorized by Congress for implementation of Title I requirements, and nearly $3.4 billion was appropriated. Approximately 9 million low-income children were served and between 5 and 6 million of these were of elementary school age. Nearly 87% of all public school districts in the United States received Title I funds (National Center for Education Statistics [NCES], 1979) which support between 3 and 4% of all national elementary and secondary education expenses (U.S. Office of Education, 1979). Title I programs are usually (but not necessarily) confined to basic skill areas such as mathematics, reading, and language arts.

Title I is the first major social legislation which specifically required routine evaluation of its programs. The evaluation system which was constructed is characterized by a common measurement metric and by three alternative research designs, one of which a school district may choose for evaluation purposes. The metric, termed the Normal Curve Equivalent (NCE) Score is a standard score with a mean of 50 and a standard deviation of 21.06. District-level estimates of program effect which are generated by the chosen design are reported on this scale to enable aggregation of gains at the state and national level.

The three research designs have been termed Model A: the Norm-Referenced Model; Model B: the Control Group Model; and Model C: the Special Regression Model. A recent national survey of school districts (NCES, 1979) estimates that of the school districts which responded, 87% had Title I programs in the 1978-1979 school year and 63% had used a definable research model. Of those which had used a model, 86% used Model A, about 2% used Model B, about 2% used Model C, and 10% made use of an alternative or locally developed model. In a less formal attempt to contact users of Model B and C, however, Trochim (1980) found approximately 40 school districts using Model C and fewer than five which used Model B. Because so few instances of Model B could be located, this research is concerned only with a comparison of results for Models A and C.[1] For purposes of exposition, Model A is termed here the NR (Norm-Referenced) Design while Model C is labelled the RD (Regression-Discontinuity) Design because this latter nomenclature is more widely recognized outside of Title I compensatory education than the "Special Regression" label. These two designs can be described briefly as follows:

> *The Norm-Referenced (NR) Design.* Under this design students are administered a test which is used for assignment to the Title I program. Assignment might be made on the basis of some cutoff score on this test, with all students scoring below a certain selected value being assigned to the program, or test scores may be used in an advisory way with qualitative judgements as the deciding factors. For example, Tallmadge (1978) states: "If selection test results conflict with teacher opinions, then these opinions may be used to change the assignment of individual children[p. 6]." Once program students have been selected they are given a pretest, and after the program are given a posttest. The analysis involves computing the mean pretest and posttest scale scores for the program students who took both tests, converting these to the equivalent national norming group percentiles using tables supplied by the test producer, and finally, converting these percentiles into average NCE scores. The estimate of program effect is simply the difference between the posttest and pretest average NCE values for the program students. The crucial assumption in this design is that in the absence of any program the average percentile rank of the group (and hence, the average NCE value) would be the same on the pretest and the posttest. That is, in the null case the program group should on both tests maintain the same relative position within the population of the norming sample. In effect, the implied or pseudo-comparison group is a hypothetical subgroup of the norming population which achieved the same average pretest percentile score as the program group. Gains in average percentile rank or average NCE score theoretically reflect gains of the program students relative to this pseudo-comparison group.
>
> *The Regression-Discontinuity (RD) Design.* With this design students are assigned to program or comparison groups by means of a cutoff score on the pretest. For example, all students scoring below the 30th percentile on an achievement pretest might be assigned to the Title I program while all others would act as the comparison group. After the program, students in both groups are administered a posttest. Although the design is usually termed the "Special Regression Model" within Title I circles, it is in essence the regression-discontinuity design as described in Cook and Campbell (1979) and Campbell and Stanley (1966). Hypothetical data from such a design are illustrated in Figure 1. Each point on the figure represents a single pretest and posttest score for an individual student. Program group student scores are indicated by an "X" while comparison group student scores are indicated with an "0". The cutoff value in this simulated data is a pretest score of 0. Figure 2 indicates the regression lines for these data. The dashed line in the figure is the extension of the comparison group regresion line into the program group pretest range. This dashed line represents the program group posttest performance which would be expected if the program had no effect. In this example the actual program group regression line is displaced above this projection for any pretest value. One might conclude that on the average the program resulted in an increase in posttest performance over what would be expected normally. In Title I evaluation the statistical analysis of data from the RD design involves computing separate linear regression lines for the program and comparison groups and estimating the difference between the projection of the comparison group line and the program group line at both the cuttoff value and the program gropu pretest mean.

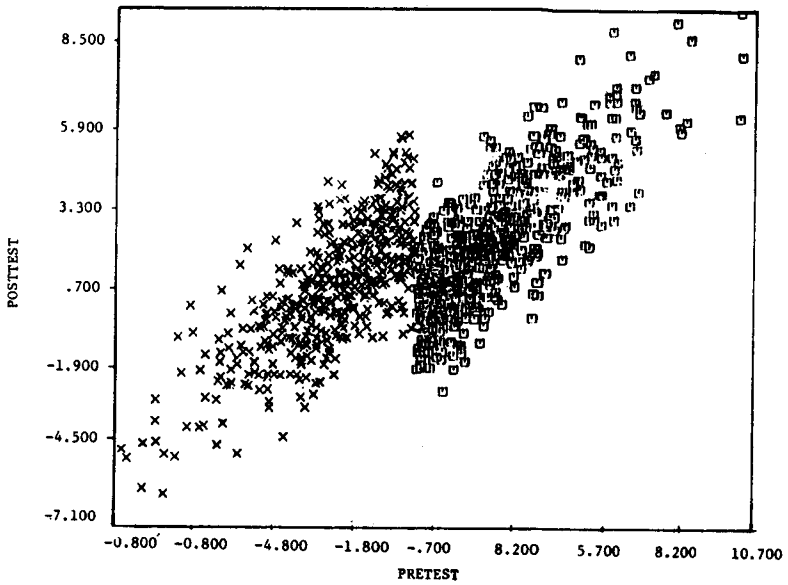

Figure 1: Hypothetical Data for Regression-Discontinuity in Compensatory Education

These two designs represent distinct methodological traditions in applied social research. NR-type designs have been used or implied in much of the educational research which relies on normative information from standardized tests for an estimate of the growth which would be expected in the absence of special training (Tallmadge, 1980). The RD design represents a tradition in quasi-experimental pretest-posttest control group designs as illustrated in the work of Campbell and Stanley (1966) and Cook and Campbell (1979). Conflicts in the results yielded by these two designs within the context of compensatory education are

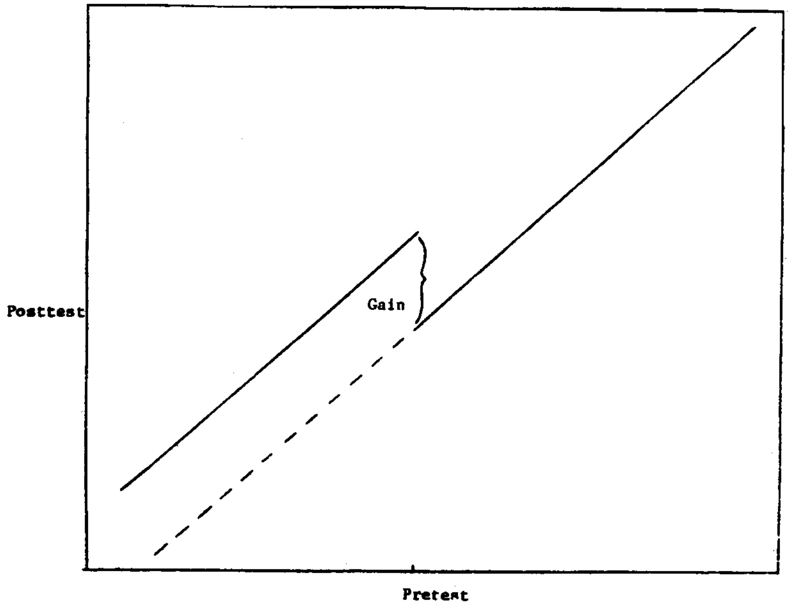

Figure 2: Regression Lines for Hypothetical Regression-Discontinuity Data in Figure 1

therefore likely to have implications for instances of these two traditions in other applied social research arenas.

THE PATTERN OF RESULTS

Three major sources of information provide evidence for a discrepancy in the average gains obtained when using the NR and RD designs—interviews of Title I evaluators at the local and regional levels, review of relevant Title I literature, and the distributions of gain estimates obtained from Title I programs.

Trochim (1980) reports the results of interviews conducted with at least one representative of each of the ten Title I regional Technical Assistance Centers (TAC) and many local Title I evaluators. Interviewees who were aware of instances of both designs indicated virtually unanimously that the designs appeared to yield different results on the average. It was also generally agreed that the average results from the NR design tended to be higher than those from the RD design. Furthermore, persons who were most familiar with the results from many evaluations corroborated the notion that in general the NR design yielded positive gains while the RD design yielded gains which were near zero or even negative.

This discrepancy has been acknowledged in the Title I literature. Hardy (1978) and Echternacht (1978, 1980) cite results obtained in Florida where sufficient instances of both designs permitted the determination of a pattern of results. Others who have attempted to compare the two designs directly on the same program (Murray, 1978; House, 1979; Long et al., 1979) obtained results which are not inconsistent with the general pattern cited here, although these studies are based on too few instances to permit confident generalizations.

The most convincing evidence for the pattern of gains for the two designs comes from the State of Florida, largely because there are sufficient instances of the application of both designs to permit meaningful comparisons. All estimates of program effect in Florida for the NR design for the 1978-1979 school year (n= 614) and all estimates for the RD design for the 1977-1978 and the 1978-1979 school years (n= 273) have been obtained. The average gains were 6.595 NCE units (SE= .302) for the NR design, -.799 NCE units (SE= .398) for the RD estimate at the program group pretest mean and -2.371 NCE units (SE= .377) for the RD estimate at the cutoff. The distributions of gains for these three estimates are depicted in Figure 3. Clearly, the evidence indicates that on the average the NR design yields significantly positive estimates, while the RD design appears to yield a zero or perhaps slightly negative gain.

Other studies of compensatory education provide little guidance concerning which of these two designs yields estimates which are closer to the truth. On the basis of what is known about the effects of compensatory education in general, it is difficult to say what the "expected" gain might be. Many previous studies have been criticized on methodological, measurement, or analytic grounds (Wick, 1978; Campbell and Erlebacher, 1970; Campbell and Boruch, 1975). Even granting that 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 (Magidson, 1977; Bentler and Woodward, 1978). In spite of this, significantly harmful effects have so far primarily been explainable as mistaken methodology. Thus, in order to determine the likely source of the discrepancy in results

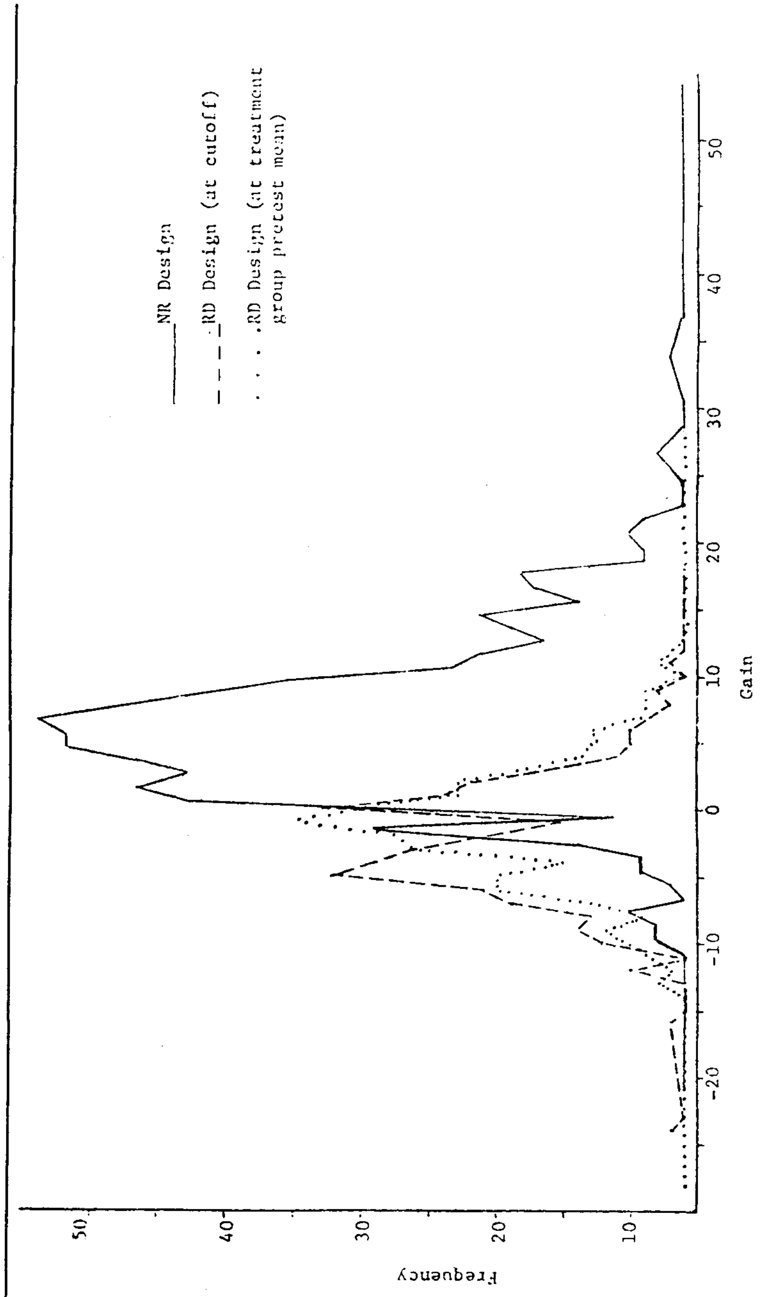

Figure 3: Distribution of Gains

reported here it is necessary to examine the designs in question within the context of Title I evaluation.

SOME LIKELY SOURCES OF THE DISCREPANCY IN RESULTS

It is possible, although hardly plausible, that both designs could be yielding accurate estimates of effect even though they disagree. For example, it may be that since the NR design tends to require less cost and effort than the RD design, districts which use it have more time, money, or energy to devote to programmatic efforts. The discrepancy in gains might then be attributable to differential implementation of the programs rather than to the designs themselves. However, explanations of this type are not likely and it is reasonable to hypothesize that one or both of the designs yields biased estimates of effect.

It is useful to begin an investigation of bias by considering how the methodological community views the strengths and weaknesses of each of the designs. Judgements about the relative strengths of research designs are often made in the methodological literature and, in general, the RD design is usually depicted as "theoretically" stronger than the NR model (Tallmadge and Wood, 1978; Murray et al., 1979; Echternacht, 1979). Typical of such distinctions is a statement by Linn (1979) where "Model A" is the NR design and "Model C" is RD:

> If viewed as research designs, the three RMC models are ranked easily in terms of their relative internal validity. In its idealized form Model B is a classic experimental design and ranks highest in terms of internal validity. Model A ranks third, with Model C somewhere in between. This ranking agrees with the stated order of preference provided by developers of the models [p. 25].

The quality of the NR design has been questioned in several key areas (Hansen, 1978)—the appropriateness of using the norming sample for comparison, especially when many norm students also receive compensatory education; the viability of the equipercentile (or, more properly, equi-NCE) assumption which holds that in the null case the program group pre- and post-average NCEs should be equal; the use of out-of-level testing; and, the testing of students at different times during the academic year than the norm group was tested.

While the RD design is generally perceived as methodologically stronger than NR, it is also usually seen as more difficult to implement.

This is at least in part due to the requirement of strict adherence to the cutoff value in assignment, to the need to compile data for both program and comparison students, and to the relatively more complex statistical analysis which must be conducted.

The remainder of this article addresses three potential biasing factors for each of the designs. These are by no means the only such factors nor are they necessarily the most important ones. Each of these factors is considered likely to be present within Title I evaluation contexts although it is dificult to estimate the degree and, in some instances, the direction of bias which might be expected. Three problems relevant to the NR design will be discussed first, followed by consideration of three major problems with the RD design.

THREE POTENTIALLY BIASING FACTORS IN THE NR DESIGN

RESIDUAL REGRESSION ARTIFACTS IN THE NR DESIGN

A distinguishing characteristic of the NR design is the requirement of a selection measure which is separate from the pretest. This was included in an attempt to avoid the commonly recognized phenomenon of regression to the mean. Before examining whether the separate selection measure in fact eliminates the regression phenomenon, it is useful to review briefly the traditional presentation of the regression artifact.

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 appear to "regress" toward the overall mean of this other distribution. If the selection measure is a pretest and the other measure a posttest, students will appear on the average to change even in the absence of a program. The amount of regression to the mean which occurs between any two measures, x and y, can be specified. A group may be chosen from the lower end of the distribution of variable x as shown in Figure 4. When the standard deviations of the two distributions are equal (e.g., they are in standard score form) the correlation between the two measures 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, 100(1-r_{xy}) gives the percentage of regression to the mean for standardized variables. As illustrated in Figure 4, if there is a perfect correlation between x and y there is no regression to the mean (i.e., 100(1-1) = 0% regression). Conversely, if there is no correlation (i.e., r_{ry} = 0) there is maximum regression to the mean (i.e., 100(1-0)= 100% regression). If r_{xy} = .6 the mean of the selected group on y will be 40% closer to the overall mean of y than their mean on x is to the overall x mean (i.e., 100(1-.6)= 40% regression). If x is a pretest and y a posttest, except when they are perfectly correlated, the typical compensatory education program group will appear to have improved even in the absence of any program simply because of the regression artifact which

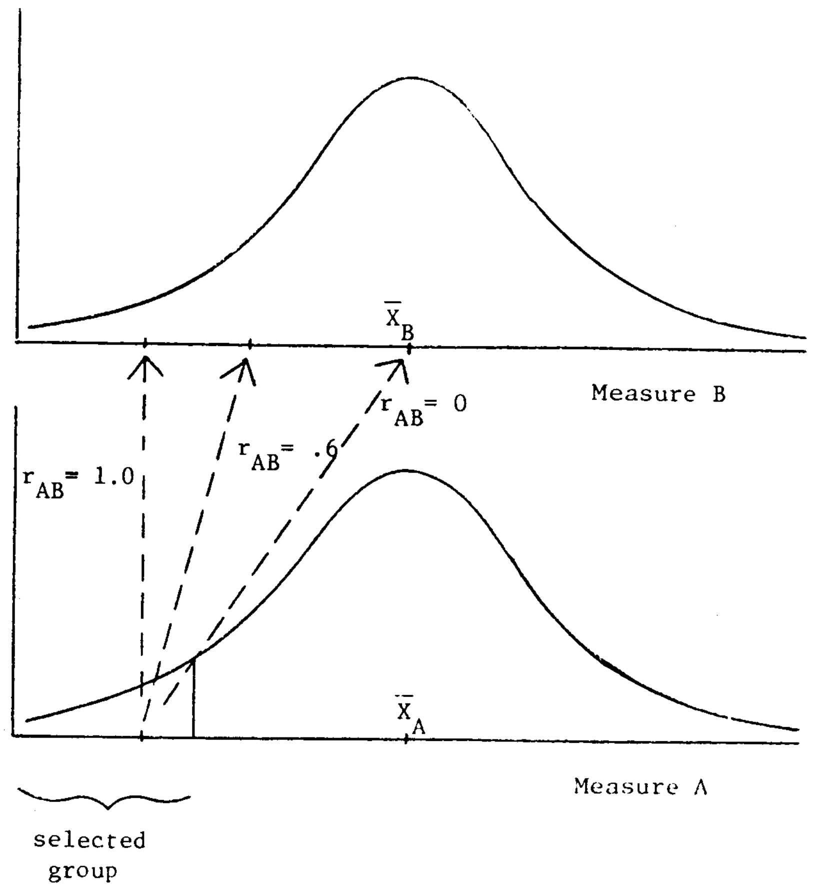

Figure 4: Regression Artifact with Two Variables

results from the less than perfect pretest-posttest correlation and selection from the lower extreme of the pretest distribution.

The NR design was constructed to avoid the regression artifact by including a separate selection test. It was thought that assignment would then be independent of the potential regression artifact. The fact is that this separate selection measure is likely to remove only part of the regression to the mean.[2] The regression artifact argument can be extended to the NR design by considering an assignment variable, z, a pretest, x, and a posttest, y, all in standard score form and hence having equal standard deviation units. In this example, the assumption is that the program group is selected from the lower end of the distribution of variable z as is commonly done in Title I (i.e., if z is a measure of achievement those "most needy" would be chosen). If the correlation between the assignment measure and the pretest is $r_{zx} = .8$, (Figure 5) there would be regression to the mean from z to x of 20%. Further, if the correlation between the assignment measure and the posttest, y, is $r_{zy} = .5$, regression to the mean from z to y equalling 50% of the distance toward the overall mean of y would exist. By subtracting these two regression artifacts the amount of regression between the pretest and posttest, in this case 30%, is obtained. This is termed here the "residual regression artifact." The use of a separate selection measure in this example has reduced but not removed the regression artifact and one would again find that students improve from pretest to posttest even if no program is ever given.

The size of the residual regression artifact in the NR design depends entirely on two correlations, r_{zx} and r_{zy} where these are correlations between standardized variables. If $r_{zx} > r_{zy}$ some regression to the mean will be unaccounted for by the separate selection measure. If $r_{zx} = r_{zy}$ there is no residual regression to the mean. If $r_{zx} < r_{zy}$ there is actually regression away from the posttest mean and the program students appear to lose ground from pretest to posttest. In order to judge whether a separate selection measure removes the regression artifact one needs to determine which of the three patterns of correlations, if any, is typically obtained.

In general, it is reasonable to assume that correlations are higher the closer in time two measurements are taken. Thus, over time, repeated measures of the same variable tend to show progressively smaller correlations with the first measurement. The size of the correlations r_{zx} and r_{zy} therefore depends on two factors—the time between the measurement of x, y, and z and the rate at which the correlations erode over time. Figure 6 shows a hypothetical erosion pattern and two measurement scenarios. In the first case (left side of Figure 6) the pretest

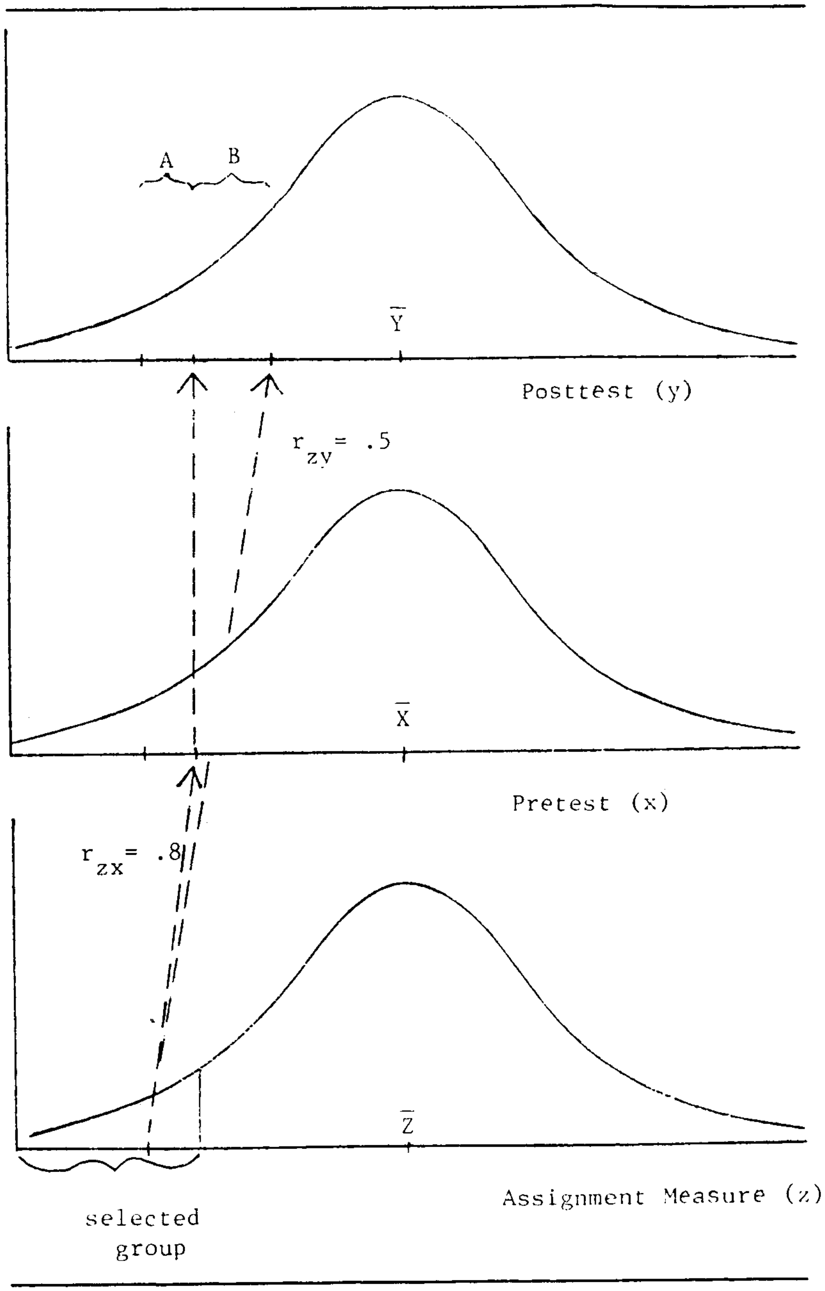

Figure 5: "Residual Regression Artifact" in Model A

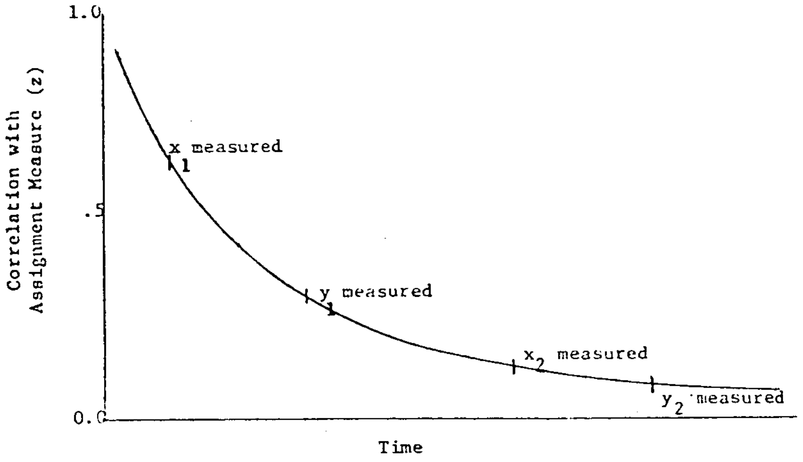

Figure 6: Hypothetical Correlation Erosion Pattern in Relation to Time of Pretest (x) and Posttest (y) Measurement

and posttest (x_1 and y_1) are measured while their correlations with the assignment variable are eroding at a rapid rate. In the second case (right side of the graph) both measures (x_2 and y_2) are taken after the greatest erosion has occurred. In both scenarios the same time elapses between the pretest and posttest. While a residual regression artifact occurs in both cases, it is much smaller in the second case than the first.

Theoretically, if one knows the correlations r_{zx} and r_{zy} in the absence of a program effect, it should be possible to adjust estimates of gain to account for the residual regression artifact. However, corrections of this type are hazardous for two reasons. First, they require knowledge of the temporal erosion pattern and this is likely to vary for different traits, and even perhaps for different tests of the same attribute (e.g., different levels or forms of achievement tests). Second, even if the overall rate of erosion is reasonably known, accurate estimates of the correlations in the absence of a program effect are necessary. This would necessitate use of some type of comparison group or, for the NR design, use of published correlations over time. In either case, the quality of such corrections would be doubtful and only serve to increase the assumptive character of the design.

TABLE 1
Average NCE Gain for Projects Grouped by Pretest Mean NCE

Pretest Mean NCE[a]	*N*	*NCE Gain*
0-5	7	18.30
5-10	10	10.67
10-15	13	15.29
15-20	37	9.46
20-25	116	8.23
25-30	161	6.40
30-35	126	5.57
35-40	80	5.05
40-45	40	4.00
45-50	18	3.58

a. Projects with Pretest mean NCEs which fell on interval boundaries were assigned to the lower value interval. For example, the first interval is actually 0-4.999 . . . but was rounded for clarity of presentation.

The degree to which the discrepancy in results reported earlier may be due to the residual regression artifact is difficult to estimate for the same reason that it is hard to devise corrections for the residual regression bias. However, it is possible to get some empirical confirmation that a bias is occurring by examining the pattern of results obtained for the NR design in Florida. This can be done by classifying the studies into NCE intervals on the basis of group pretest means. If a residual regression artifact is operating, gains for groups who scored low on the pretest should be greater on the average than gains of the higher pretest scoring groups. Table 1 shows average gain for projects which are grouped into ten intervals of five NCE units each, covering a range of 0 to 50 NCEs. Six projects were excluded from the table (therefore, n = 608) because their average pretest score exceeded the mean of 50 NCE units. Clearly, the results do nothing to repudiate the hypothesis that a positive bias due to a residual regression artifact occurs with the NR design. Although a similar pattern would be expected if there were an interaction effect between the program and the pretest, this is considered a less plausible explanation than the residual regression artifact one.

ATTRITION BIAS IN THE NR DESIGN

Although the NR design requires that a selection measure be administered to all potential participants, only program students are given the pretest and posttest. The required analysis for estimating gain is based on the pretest and posttest averages for only those students who took both tests. Two attrition-related problems tend to arise when dropouts are nonrandom. First, if only matched cases are used (as required) one can at least expect the positive bias due to the residual regression artifact for the matched (i.e., nonattrited) students. This may be greater or less than the bias expected for the entire original sample depending on whether attrition is disproportionately greater for higher- or lower-scoring program students. In addition, attrition of this sort obviously calls into question the use of the norming sample as a comparison standard.

The second attrition-related problem occurs if the requirement that only matched cases be analyzed is violated. This will be termed the unmatched attrition case. Here, if a program student takes one test and not the other, the available test score is still included in the group averages. While this is a clear violation of the Title I requirements, there is some reason to believe that the difficulty of matching pretest and posttest scores (Trochim, 1980) leads school districts into this practice.

The major purpose of this discussion is to determine the likely effects of unmatched attrition on estimates of gain, and ultimately, the discrepancy in results. The effects of such attrition between the pretest and the posttest will be considered first. Subsequently, the combined effects of attrition between the selection measure and pretest and the pretest and posttest will be discussed.

The typical Title I program group is composed of students whose pretest scores will in most cases be below the population average of 50 NCE units. If unmatched attrition is proportionately greater among students who score below the prestest mean of the program group, a positive bias will result. This is due to two factors. First, the remaining group will consist of the higher-scoring program students. In the absence of any program an apparent gain will occur from the observed pretest mean (which includes attrition cases) to the observed postest mean. Second, in addition to this gain there will also tend to be a positive regression artifact bias from the remaining students' pretest mean to their posttest mean. Thus, the positive regression artifact will augment the positive bias due to the high-scoring remaining group.

The direction of bias is not easily specified when there is proportionately greater unmatched attrition from the high-pretest-scoring program students. The remaining group would have a lower pretest average, indicating a potential negative bias. However, a positive regression artifact bias is also expected for this group. Thus, when attrition occurs primarily among high pretest program students, a positive regression artifact bias competes with a likely negative bias resulting from the low scoring remaining group.

Consider a hypothetical example of how a positive bias can result when attrition occurs among the higher pretest scorers. It is assumed that the entire program group had a pretest mean of 20 NCEs and that after attrition of the higher scorers the remaining group would have a pretest mean of 15 NCEs. The observed pretest mean is therefore 20 NCEs but the expected posttest mean in the absence of the regression artifact would be 15 NCEs for an apparent negative bias of -5 NCEs. However, if in this example the standardized $r_{xy} = .8$ there would be a positive regression to the population posttest mean of 20% (i.e., 100(1-.8) = 20% of the distance between the expected posttest mean of 15 NCEs and the population mean of 50 NCEs. Thus, there would be a positive regression artifact bias of 7 NCE units, that is, .2(50-15)= 7, and there would be an *overall positive bias* of 2 NCE units.

Several conclusions are reasonable at this point. First, if unmatched attrition occurs primarily in the lower pretest scorers there will be a positive bias. Second, slightly greater rates of attrition among higher pretest scorers are also likely to result in a positive bias (although this would be less than for attrition of lower scorers). Finally, the attrition bias will be negative in direction only when there is a disproportionately great enough attrition rate among higher scorers so that the resulting loss due to lower-scoring retainees exceeds their gain due to regression to the mean.

Because attrition rates for various pretest levels are not routinely reported in the Title I literature it is difficult to say what pattern of attrition is most common. With no knowledge of the distribution of attrition rates it is reasonable to conclude that attrition between the pretest and posttest will in general be more likely to result in a positive bias. One can, however, obtain a rough idea of the likely attrition pattern by examining the major sources of attrition. Kaskowitz and Friendly (1980) report several likely sources:

—students entering and leaving a school or district;

—students entering and leaving a project;

—students being held back or double promoted in grade progression;

—absence on test dates;

—invalid test administration;

—loss of data in processing and editing;

—deliberate omission of data.

For most of these factors it is plausible to argue that attrition would be more likely to occur at a greater rate among the lower-scoring students because these students would be more likely to miscode answer sheets, have greater absenteeism, be held back a grade, be discouraged and leave the program, and so on. If this assessment is correct, greater confidence can be placed in the likelihood of a positive bias due to unmatched attrition.

The more realistic case of attrition between both the selection and pretest and the pretest and posttest is more complex but leads to similar conclusions. Attrition bias may either inflate the pretest program group average or, less often, result in a lower pretest mean for the reasons discussed above. Because lower or higher pretest means will result in different amounts of residual regression artifact bias, attrition between the selection measure and pretest may affect the amount of bias resulting from attrition between the pretest and posttest. However, even in this three-variable case, the direction of bias will still be positive except when there is enough attrition among higher scorers to enable the negative bias of the low-scoring remaining group to exceed the residual regression artifact bias. On this basis it is reasonable to conclude that the discrepancy in results yielded by the NR and RD designs may in part be attributable to a positive unmatched attrition bias in the NR design.

TIME-OF-TESTING BIAS IN THE NR DESIGN

The NR design relies on a comparison between the program group and what has been termed here a pseudo-comparison group which is a hypothetical subsample of the norming group which is similar to the program students. It is important, therefore, to examine how test norms are developed in order to determine the reasonableness of such a comparison.

Typically, test publishers developed norms for a test on the basis of an annual test administration. Thus, samples of students might be tested in the fall of one year and the fall of the next or in the spring of one year and the spring of the next. In the typical NR design scenario, the selection test consists of an annual district-wide achievement test in the spring; the pretest is based on fall administration of the test to the program students and the posttest is comprised of the annual test given in the following spring (which then becomes the selection test for the subsequent year). If a test were normed on the basis of annual test samples, either the pretest or posttest in the NR design would have to be compared with interpolated norms. Thus, if the test had been normed based on fall-to-fall testing the spring norm would be an interpolation, while if the test had been normed spring-to-spring the fall norm would be an interpolation. As Linn (1979) explains:

> Normatively derived scores for other testing dates were usually obtained by linear interpolation with the three summer months treated as a single month. That is, it was assumed that growth was linear for the nine month school year and that one additional month's gain was made during the summer.

Thus, when the test was normed based on annual administrations, either the obtained pretest or posttest in the NR design is typically compared with an interpolated norm. Specifically, the question is whether the difference between an obtained and interpolated norm is the same as the difference which would be found if an actual testing were substituted for the latter.

Several attempts have been made to answer this question using data based on fall and spring norm testings. The results of these studies must be interpreted cautiously because they are often based, at least in part, on cross-sectional rather than longitudinal data. However, the general pattern of results indicates that larger gains occur between fall and spring testings (Beck, 1975; David and Pelavin, 1977; Linn, 1979), than between spring and fall testings. This is typically attributed to a lower rate of growth over the summer months.

A hypothetical graph of changes across testing times which is similar to those reported in the literature (Linn, 1979) is depicted in Figure 7. Such a pattern might be obtained if the same group of norm students were measured in the fall and spring for two successive years assuming that there are no clear floor or ceiling effects at any testing. It is important to recognize that the solid line depicts the growth pattern expected in norm scores even with the typical "summer growth" correction. The dashed lines between the two fall tests and two spring tests indicate hypothetical linear interpolations which might be used to obtain estimates of norm group performance for points in time between the norm testing administrations. In general, the figure shows that when the test has been normed with fall-to-fall tests the spring norm will be underestimated while with spring-to-spring norming the fall norm will be overestimated.

Assuming that this pattern is accurate, it is relatively easy to determine the bias which may result from using tests which are normed on the basis of such annual testings. If the test which is used in a

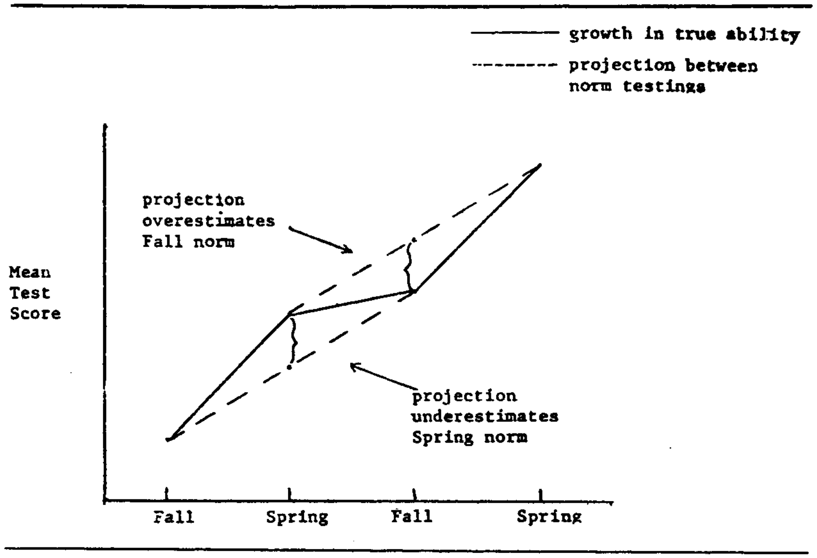

Figure 7: Time of Testing Bias

particular NR design was normed with fall-to-fall testings, the spring norm will be an underestimate of the true norm value. In the absence of a program effect, the program group will appear to improve from fall to spring simply because the norm posttest value is underestimated. Similarly, if the test which is used was normed using spring-to-spring testings, the fall norm will be an overestimate of the true norm value. Thus, in the absence of any program effect, the Title I program group would appear to be lower than the estimated pretest norm and therefore would show pretest-posttest improvement relative to the norm sample simply because the pretest norm is an overestimate. In this hypothetical scenario it is clear that whenever the test which is used was normed on the basis of annual testings the estimate of program effect is likely to be positively biased.

In practice, the situation is likely to be more complex. Tests may not be administered near the norm or interpolated norm testing dates and it may be necessary for a school district to attempt to construct local interpolations or extrapolations appropriate to local test administration times. In addition, in order to avoid floor and ceiling effects it is sometimes necessary to use different levels of a test for pretest and posttest. In this case, norms are dependent on the standardized scales developed by the test producers to "vertically equate" scores from different levels of a test. In any event, it is clear that the use of normative test data in the NR design relies on assumptions about change in the norming sample. In many cases these assumptions are unverified or, as in this case, actually suspect.

To some extent this time-of-testing problem can be reduced if tests are normed on the basis of fall-to-spring-to-fall longitudinal norm samples. There is some indication that test producers appreciate this fact and have modified or are considering modifying their norming procedures. Nevertheless, this is a relatively recent trend and it is a fair assumption that the majority of the 614 NR design projects which are aggregated in this article relied on tests which were normed on the basis of annual testings. Because of this, it is reasonable to conclude that the discrepancy in the results generated using the NR and RD designs can be attributed, at least in part, to a positive bias which results from time-of-testing problems in the NR design.

THREE POTENTIALLY BIASING FACTORS IN THE RD DESIGN

The RD design is based on the assumption that the true pretest-posttest relationship is known or can be estimated from the data at hand. Typically, a regression model is fit to the data which describes this relationship. Estimates of program effect are based on this model and thus, factors which lead to biased effect estimates must impinge upon the regression-modeling process in some way. Because of this, it is important to examine briefly the major issues in the statistical analysis of the RD design before turning to an elaboration of several likely biasing factors and the way in which they affect the analysis.

A general polynomial regression model which is often appropriate for data from the RD design is given in Trochim (1980):[3]

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 z_i^* + \beta_3 x_i^* z_i + \ldots$$

$$\beta_{n-1} x_i^{*s} + \beta_n x_i^{*s} z_i + e_i$$

where:

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

y_i = postprogram 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 = normally and independently distributed random error

The major hypothesis of interest is

$$H_0: \beta_2 = 0$$

tested against the alternative

$$H_1: \beta_2 \neq 0$$

Several important assumptions of this general model must be recognized. First, it is only appropriate if the assignment strategy is correctly implemented, that is, there is no misassignment relative to the pretest cutoff score. 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 estmates of program effect may be biased. 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 pre-post relationship is logarithmic, one can use the logarithm of y instead of y in the model. Third, the model allows for changes (or interaction terms) in slope or function between the program and comparison groups. Thus, one can fit the same linear, quadratic, and cubic functions across both groups as well as separate ones within each group. Fourth, the program effect is estimated only at the pretest cutoff point. This is accomplished by subtracting the cutoff value, x_0, from each pretest score, thus setting the cutoff equal to the intercept (i.e., $x_0 = 0$). This differs from the typical Title I analysis where estimates are calculated at both the cutoff value and the program group pretest mean (Tallmadge and Horst, 1976; Tallmadge, 1976). This restiction is included because of the inherent instability of extrapolations of regression lines. Estimates of effect at the cutoff point involve no extrapolation whereas estimates at the program group pretest mean require extrapolation of the comparison group model into the region of the program group scores. The general model could, however, be easily modified to estimate program effect at the program group pretest mean by subtracting this mean instead of the cutoff from each pretest test score.

It is important to recognize that the recommended Title I analysis is simply one possible subset of the more general model offered here. Specifically, the Title I model fits straight lines to each group and allows these lines to have different slopes. In the notation of the general model, the Title I analysis is

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 z_i + \beta_3 x_i^* z_i + e_i$$

where x_i is the pretest score for individual i with the cutoff value subtracted when estimates are at the cutoff or the program group pretest mean subtracted when estimates are made at that point.

Ideally, one wishes to select that subset of variables from the general model described above which best describes the true pretest-posttest relationship. A number of standard procedures are available for such model specification (Hocking, 1976). The final model which is selected may be under-, over-, or exactly-specified. A model is exactly-specified when it includes only the variables which are in the true model. It is overspecified if it includes all variables in the true model as well as additional extraneous terms. It is underspecified if it does not include all variables in the true model (even if additional extraneous terms are included). Trochim (1980) has demonstrated that estimates of effect are biased if the model is underspecified. Both exact and overspecification result in unbiased estimates although precision is lost with overspecification.

This brief discussion of the statistical analysis of the RD design is included here to illustrate the restrictive assumptions of the Title I analytic strategy. All three of the problems discussed below tend to lead to nonlinear pretest-posttest distributions regardless of the true model which would have occurred if the problems had been absent. Thus, the Title I analysis, which assumes a linear true model, will tend to be underspecified and to yield biased estimates in the situations described below. In most cases it will be shown that the bias is negative in direction and that the problems could therefore be contributing to the discrepancy in gains between the NR and RD designs.

MISASSIGNMENT BIAS IN THE RD DESIGN

Most school districts which use the RD design employ some procedure which makes it possible to challenge the assignment of a student by the cutoff criterion. Usually a challenge is initiated by a teacher although the source may at times be a parent or school principal. A greater proportion of students tend to be challenged into the program group than out of it. In some cases the teacher's judgment is considered sufficient evidence to warrant a change of group status, but more often the student is retested and a cutoff score on the retest is used as the criterion. Challenges can be motivated by an honest belief in the fallibility of the test instrument, by political factors or favoritism, by a reluctance to deny potentially useful training to "borderline" students, and for a number of other reasons. Trochim (1980) points out that the practice of challenging assignment is widespread, that there is often no limit put on the number of times a student may be retested, and that challenges sometimes go unreported.

The central question here is whether misassignment relative to the cutoff score might be related to the pattern of gains described above. It is useful to construct a hypothetical example to help clarify what might occur. In this example it is assumed that all of the challenges in a district are those which shift students into the program. If the challenges are reasonable, these might be students who parents, teachers, or administrators feel scored artificially high on the pretest—their true ability should have placed them in the program group. Furthermore, one might expect that these lower-ability students would on the average perform more poorly on the posttest than others who received the same pretest scores. An extreme version of this hypothetical group is indicated by the darkened portion of the graph in Figure 8. It is important to recognize that the graph portrays the original bivariate distribution and would be the same whether the challenge is based on teacher judgments or retest scores.

There are a number of potential strategies for analyzing data when challenges have been allowed. One could, for example, act as though the challenges were never made. In fact, if the challenges go unreported the evaluator cannot be aware of them and will simply analyze the data using the assignment indicated by the pretest cutoff score. For the null case depicted in Figure 8 no bias would be expected. However, if the program is effective the challenged cases should evidence this effect. If the analysis assumes that these students are in the comparison group, that group's posttest scores would be increased near the cutoff point and the slope of the comparison group linear regression line would be attenuated somewhat resulting in an underestimate of the program effect. If the challenges are reported, one might be tempted to analyze the data on the basis of actual assignment (i.e., as challenged). Here, in both the null and effect cases the slope of the program group linear regression line would be attenuated somewhat (due to inclusion of the low posttest scoring challenge students) and a negative pseudo-effect would be expected. Another strategy would be to exclude the challenge cases in the analysis using their retest scores in place of the pretest. This would have the effect of increasing the number of the cases immediately below the cutoff value which also tend to fall below

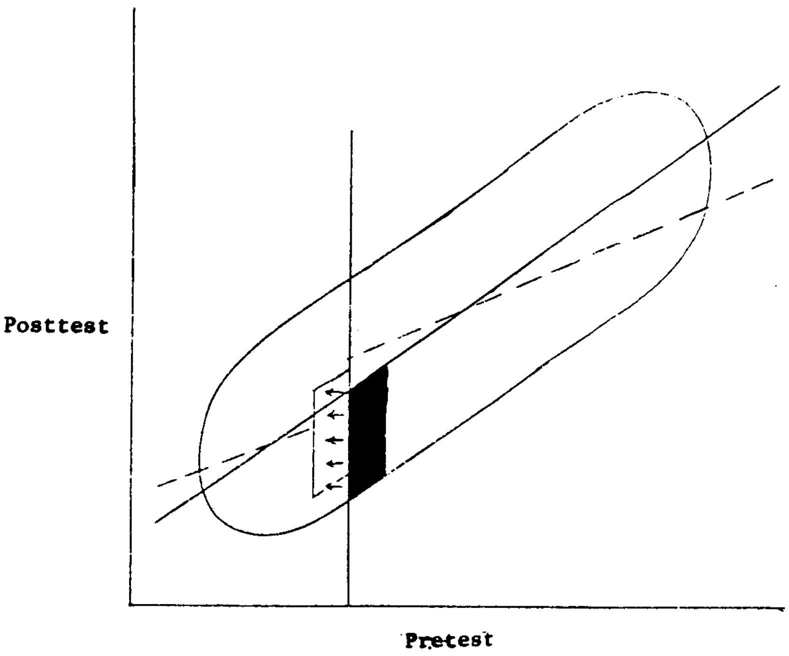

Figure 8: The Effect of Challenges on Within-Group Slopes

the linear regression line as shown in Figure 8. The addition of these cases in the program group and their removal from the comparison group would serve to attenuate the slopes of both linear regression lines and again one would expect a negative bias. In all of these analyses of this hypothetical "reasonable" challenge scenario one expects that the true program effect will be underestimated.

A similar scenario can be constructed for the case of challenges which move a student out of the program. If it is assumed that these cases are most likely students who scored just below the cutoff on the pretest and would be expected to do better than average on the posttest (because their pretest score underestimates true ability), one again expects to find negative pseudo-effects because their removal from the program group and inclusion in the comparison group will always act to attenuate the slopes of the linear regression lines in each group.

These intuitions about the likely direction of misassignment bias can be illustrated through some simple simulations. First, data are randomly generated for 1000 cases using the following models:[4]

$$x = T + e_x$$

$$TE = T + e_{TE}$$

$$RE = T + e_{RE}$$

where T is true ability, all e's are independent random error, x is the pretest, TE is the teacher rating of student ability, and RE is the score obtained for students who are retested. Thus it is assumed that the pretest, retest, and teacher rating are all imperfect but fairly reliable measures of true ability (e.g., achievement in reading or math). Second, assignment to program or comparison groups is constructed for three cases:

Sharp assignment:

$$\begin{aligned} z0 &= 1 \text{ if } x \leqslant 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

Teacher challenges:

$$\begin{aligned} z1 &= 1 \text{ if } x \leqslant 0 \text{ or } (x > 0 \text{ and } TE \leqslant 0) \\ &= 0 \text{ otherwise} \end{aligned}$$

Retest challenges:

$$\begin{aligned} z2 &= 1 \text{ if } x \leqslant 0 \text{ or } (x > 0 \text{ and } (TE \leqslant 0 \text{ and } RE \leqslant 0)) \\ &= 0 \text{ otherwise} \end{aligned}$$

The cutoff score of zero is arbitrary. The case of sharp assignment is included as a no-bias comparison. The program group consists of the lower pretest scorers and challenges are only in the direction of into the program. It is assumed that retests are given (i.e., can be a factor in assignment) only if a teacher recommends it (i.e., the teacher first judges that the student was misassigned). Thus, the retest challenge procedure is more restrictive in these simulations than the teacher challenge procedure. Third, posttest scores can be calculated under the general model:

$$y = T + gz + e_y$$

where y is the posttest, g is the program effect (either 0 or 3 units), and z is the dummy assignment variable (either z0, z1, or z2). Thus, the posttest is also a fallible measure of ability and is linearly related to the pretest through the common true score, T. Finally, the following analyses are applied to the simulated data:

(1) No challenges, analysis using actual assignment. This case is included as a no-bias comparison case.
(2) Teacher challenges, analysis using actual assignment. This analysis would only be feasible if teachers report the challenges and the actual challenged assignment variable (z1) is used.
(3) Teacher challenges, analysis using pretest assignment. Here it is assumed that the analyst is not aware of the challenges and that the original pretest assignment variable (z0) is used.
(4) Retest challenges, analysis using actual assignment. Again, this analysis is only feasible if the challenges are known to the analyst and the actual assignment variable (z2) is used.
(5) Retest challenges, analysis using pretest assignment. Again, it is assumed the analyst is unaware of the challenges and that the pretest assignment variable (z0) is used.
(6) Retest challenge, analysis using retest scores of challenged cases. Here the retest scores are substituted for the pretest scores of challenged cases.
(7) Teacher challenge, challenged cases excluded.
(8) Retest challenge, challenged cases excluded.

Each analysis is conducted using the following two-step procedure:

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

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

The first step fits a straight line with the same slope in the program and comparison groups while the second step allows the slope to differ between groups. Thus, step 1 represents the true simulated pretest-posttest relationship (in the absence of challenges) while step 2 is equivalent to the recommended Title I analytic approach. A total of twenty simulation runs were conducted for each combination of analysis (i.e., analyses 1 to 8) and gain (i.e., 0 or 3).

The average estimates of program effect (β_2) and standard errors across twenty runs are given for both steps in Table 2. Of course, when there is no effect one expects that use of the pretest assignment variable (z0) in the analysis will yield a zero effect estimate whether there were challenges or not. This is corroborated in Table 2 by the fact that estimates of effect under analyses 1, 3, and 5 are the same in the null case. With this exception in mind, it is clear that all analyses involving challenges tend to result in biased estimates of effect and in every case the bias is negative (i.e., the program effect is underestimated). Confidence intervals can be constructed using the standard errors provided. In every case, the upper limit of the .95 confidence interval falls below zero (i.e., $\beta_2 + 2SE(\beta_2) < 0$). Not surprisingly, teacher challenges tend to result in a greater bias than retest challenges in part because in these simulations the latter is more restrictive and results in fewer numbers of challenged cases. Results appear to be least biased when retest scores are substituted for the pretest scores of challenged cases. In practice, the results for this analysis may be more biased due to problems of equating the scales of pretests with retests given at a different time. In general, knowledge of challenges and use of actual assignment in the analysis yields less bias than use of the original pretest assignment.

These simulations are only intended to be illustrative. Certainly it might be useful to add more runs, include curvilinear pre-post relationships, systematically manipulate the true score and error variances, include other possible analyses and challenge scenarios, and so on. Nevertheless, these simulations do illustrate that even under relatively "ideal" conditions (e.g., normally distributed variables, unidirectional challenges, linear pre-post relationships, fairly reliable measurement), misassignment in the compensatory education case tends to lead to

TABLE 2

Estimates of Gain (β_2) and Standard Errors for Simulations of Several Challenge Procedures

	True Gain (g) = 0				*True Gain (g) = 0*			
	Step 1		*Step 2*		*Step 1*		*Step 2*	
Analysis	β_2	σ_{β_2}	β_2	σ_{β_2}	β_2	σ_{β_2}	β_2	σ_{β_2}
1. No challenges	.003	.030	.002	.030	3.031	.031	3.031	.032
2. Teacher challenge, analysis using actual assignment	–.500	.027	–.495	.030	2.511	.035	2.519	.036
3. Teacher challenge, analysis using pretest assignment	.003	.030	.002	.030	2.154	.042	2.163	.041
4. Retest challenge, analysis using actual assignment	–.375	.030	–.368	.032	2.622	.040	2.630	.040
5. Retest challenge, analysis using pretest assignment	.003	.030	.002	.030	2.551	.037	2.556	.037
6. Retest challenge, analysis using retest scores	–.127	.031	–.139	.032	2.881	.037	2.871	.037
7. Teacher challenge, challenge cases excluded	–.258	.031	–.278	.032	2.762	.038	2.744	.038
8. Retest challenge, challenge cases excluded	–.172	.031	–.179	.032	2.837	.037	2.831	.037

underestimates of the program effect using the RD design. Because of this, the misassignment problem must be considered a plausible explanation for at least part of the discrepancy in the results yielded by the NR and RD designs.

MEASUREMENT-RELATED BIAS IN THE RD DESIGN

Two separate measurement issues are discussed here, both of which are likely to have an effect on the pattern of gains obtained with the RD design. The first problem concerns the potential for floor and ceiling effects in the measures. These would 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 number of students will receive the lowest possible scores. Their scores will not be indicative of their true ability because the test does not measure that low. Floor or ceiling effects on the pretest would tend to result in a more positive pre-post slope in the vicinity of the floor or ceiling. Conversely, such effects on the posttest would tend to attenuate the slope in the vicinity of the floor or ceiling. The situation becomes especially complicated when considering that it is possible to have a floor or ceiling effect or both on either the pretest or posttest or both.

The second measurement issue of relevance is related to the chance level of the test. The concept of chance level can best be understood through a simple example. A hypothetical multiple choice test has 100 items, each having four possible answers. If a respondent guesses on all 100 items one would expect by random chance alone that the average test 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. Assuming that a portion of the students are guessing, cases with pretest scores near the chance level are likely to exhibit a lower pre-post correlation, and consequently a lower pre-post slope, than cases having higher pretest scores.

The direction of bias which would result from either of these two measurement problems depends in general on both the nature of the problem and the placement of the cutoff. For example, if there is a chance level or posttest floor effect, the pretest-posttest relationship might be best described by a line like the one shown in Figure 9. With a high pretest cutoff value the figure demonstrates that estimates of gain would be negatively biased. Conversely, with a low cutoff, estimates would tend to be positively biased.

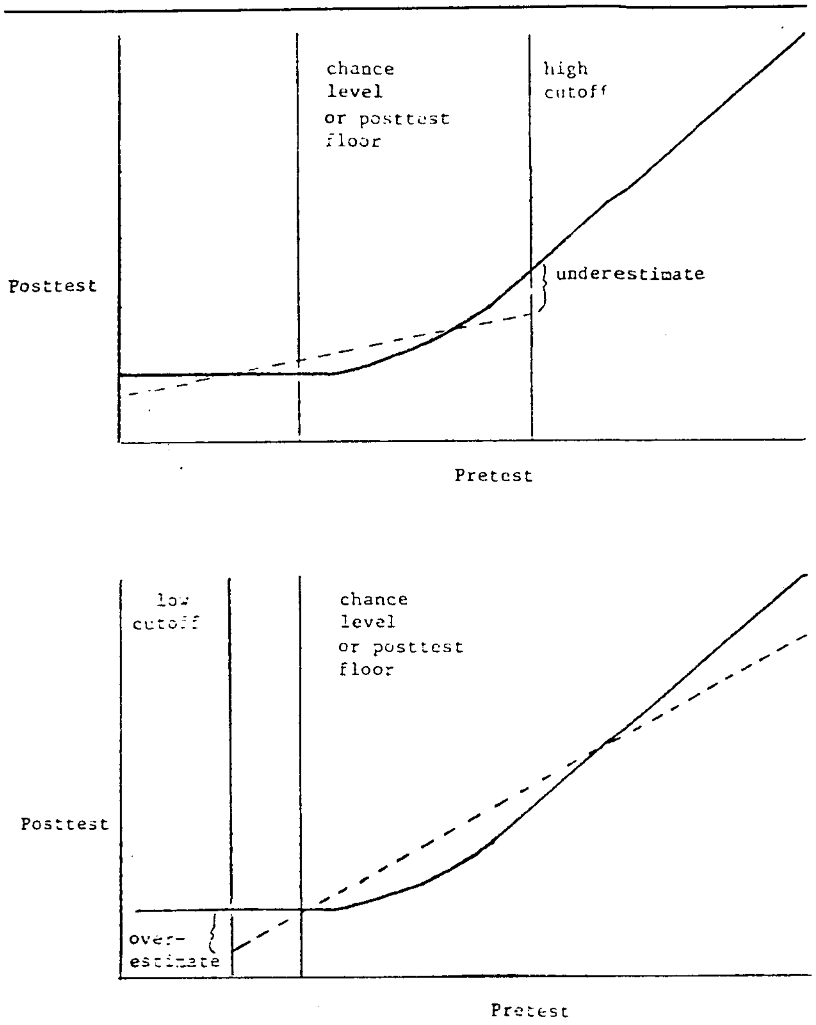

Figure 9: Effects of Cutoff Placement and Chance Levels or Floor Effects on Estimates of Effect

It is possible to get some indication of the likely direction of bias in practice by examining gains in relation to typical cutoff percentiles.

The median cutoff percentile for the 273 RD designs from the State of Florida described earlier is 28.6 with a range of 7 to 50. It is difficult to know whether this median value tends to be above or below typical chance level values or posttest floor ranges without examining the specific chance levels for the tests that were used. Nevertheless it is possible to get a rough idea of the effect of cutoff placement alone on the estimates of gain by looking at the average gain for all projects with cutoffs above and below the median cutoff percentile. When cutoff values were below the median the average gain was .2563 (SE = .720) for the estimate at the program group pretest mean and -2.3127 (SE = .737) for the estimate at the cutoff. When cutoff values were above the median, the average gain was -1.6527 (SE = .412) and -2.4181 (SE = .334) for estimates at the program group pretest mean and cutoff, respectively. Thus, for both estimates the average gain tended to be lower the higher the cutoff value. While these results must be interpreted cautiously (at least in part because of poor reporting of or adherence to cutoffs), they do not repudiate the notion that higher cutoffs may be associated with negative bias while lower ones may be linked to positive bias. In any event, the potential for bias due to the placement of the cutoff relative to the chance level of the test or floor and ceiling effects must also be considered a plausible contributing factor to the discrepancy between the results of the NR and RD design.

DATA PREPARATION PROBLEMS IN THE RD DESIGN

A large number of exclusions are routinely made in Title I evaluations in the process of preparing the data for statistical analysis. Cases are excluded from the analysis for lack of a pre-post match, because the student was "challenged" or misassigned, because the student moved either within the district or out of the district, and so on. Some exclusions are likely to have a consistent effect on estimates of gain and must be considered plausible sources of the discrepancy in gains.

This can be illustrated with the commonly made exclusion of grade repeaters, that is, those students who are held back a grade from one year to the next. It is certainly reasonable to expect that most of the students who repeat a grade are low achievement-test scorers who are eligible for Title I service. If there are a fair number of repeaters and if these cases are routinely excluded from the data analysis it is likely that the program group regression line, and subsequently the estimate of gain, will be distorted. Furthermore, it is not unlikely that these students come from even the lower portion of the Title I program group distribution and that they have more than the average share of disciplinary problems, learning disabilities, and so on. It is possible to conceive of circumstances in which excluding such students actually makes the Title I program look worse. A polar case is illustrated in Figure 10. It is assumed that the "hard-core" repeaters are low in true ability and, therefore, score low on both the pretest and posttest. Hypothetically, the majority of these cases should fall in the region indicated by the blackened portion of the figure. If such students are excluded, the slope of the program group regression line would be attenuated and an apparent negative gain would result. A similar scenario can be constructed for the exclusion of students who move either within or across districts.

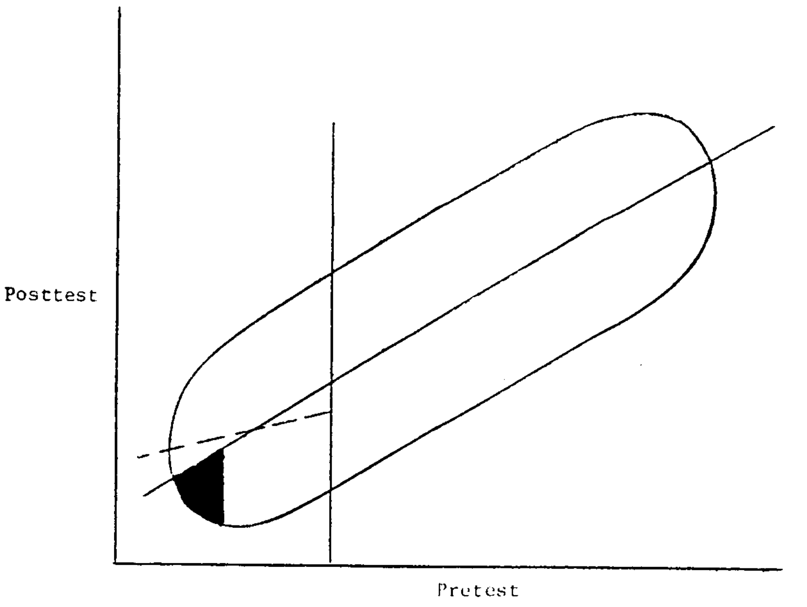

Figure 10: Effect of Excluding Title I Repeaters

The effects of excluding grade repeaters can also be illustrated through simulations. Here, a pretest is constructed using the model

$$x = T + e_x$$

where x, T, and e_x are constructed as described earlier. Next, a dummy assignment variable, z, is constructed. Again, a cutoff of zero is arbitrarily selected and the low pretest scorers receive the program. Finally, the posttest is constructed

$$y = T + gz + e_y$$

where the program effect (g) is either 0 or 3 units. It is assumed that grade repeaters on the average are the lowest in true ability, even within the low-scoring program group. In these simulations a hypothetical grade repeater group is excluded on the basis of their true scores. Specifically, all cases having a true score (T) less than −4.5 units are excluded from the analysis. As in the previous simulations there are n = 1000 cases in each run and twenty runs for each condition. The same two-step regression analysis is conducted where the first step fits the same linear function in both groups and the second step allows the slopes to differ between the groups (i.e., the Title I analytic strategy).

In the no exclusion case, as expected, results are unbiased for all analyses. With exclusions, the step 1 estimate of gain is −.093 (SE = .026) in the null case and 2.968 (SE = .036) when the true gain is equal to 3 units. When low true scorers are excluded from the step 2 (Title I) analysis, the estimate of gain is −.163 (SE = .025) in the null case and 2.90 (SE = .039) when true gain is 3 units. Obviously these simulations can only be considered illustrative. Nevertheless they do support the ideas that even under fairly optimal conditions (e.g., fairly reliable measures, normally distributed variables, linear pre-post relationships, less than 8% exclusions) the exclusion of grade repeaters can lead to biased estimates of program effect and that the bias is likely to be an underestimate of the true effect.

Another data preparation problem which can lead to bias is the occurrence of data coding errors. For example, matching of individual pretest and posttest scores is usually made using the name or ID number for the student. Both are subject to miscoding. In addition, because matching is typically done by computer it is usually essential that the coding of the name be identical on both tests. Mismatching can occur if the "long" name is coded on one test and the "short" name on the other (CATHERINE versus CATHY), if the middle initial is on one and not the other, because the middle initial is placed in the first name field, because of extension of the first name into the middle field (VINCEN E versus VINCEN T), because of abbreviated first names

(ROBT versus ROBERT), if the name is changed between testings (CLAY versus ALI), and for a variety of other reasons.

Problems of data preparation tend to result in bias if the characteristic on which a data exclusion is based is nonrandomly distributed across pretest scores. For example, if there is reason to believe that program students move or repeat grades more frequently or are more likely to make coding errors, exclusion of these student's test scores is more likely to distort the true function and bias estimates of effect. For reasons similar to the grade repeater case it may be reasonable to expect that the bias would be negative in direction. Thus, data preparation problems must be considered potential contributing factors to the discrepancy in the results of the NR and RD designs.

DISCUSSION

The discrepancy in results discussed here has implications for far more than just the arena of compensatory education evaluation. To the credit of the Title I evaluation system, the problem is noticeable primarily because there are a sufficient number of replications of each design to enable a comparison to be made. Certainly those who are engaged in the recently emerging area of meta-evaluation need to be cognizant of the potential for methodologically based meta-analytic factors. Analyses of multiple replications of evaluative studies should incorporate tests of differences between methodologies and, where discrepancies are detected, attempt to determine their causal factors.

Much of the previous discussion has to be classified as speculative in nature largely because many of the potential causal factors for the discrepancy are related to problems in implementing the research such as measurement difficulties, attrition, ill-advised data exclusions, time of testing problems, and the like. Data on the frequency and extent of such difficulties is not routinely collected in evaluation research studies. The obvious implication is that, wherever possible, it ought to be.

At this point, definitive conclusions about the effects of Title I compensatory education on achievement are not forthcoming. The more optimistic portrayal yielded by the NR design suggests a modest positive effect in the vicinity of 7 NCE units. However, the discussion presented earlier suggests that this design is likely to yield positively biased estimates of effect. On the other hand, the RD design, usually considered the methodologically stronger of the two, portrays the programs as ineffective or even slightly harmful, although the implication of the previous discussion is that this design in practice tends to yield results which underestimate the true effect. The conclusions may, of course, be dependent on a fortuitious selection of potential biasing factors. There are undoubtedly other factors involved in the discrepancy but it is not clear whether they would be prominent enough to change the patterns of bias expected on the basis of the factors presented here. It appears that the best estimate of the effect of Title I compensatory training lies somewhere between those yielded by the NR and RD designs. However, it seems prudent to withhold final judgment on the matter until more definitive information on the quality of the designs can be acquired and analyzed.

NOTES

1. The third design (Model B) is either a pretest-posttest true experiment or a nonequivalent group design depending on whether assignment to group is random or not.

2. The argument presented here is based on the Glass memo reported in Echternacht (1978).

3. See Trochim (1980) for a more detailed discussion of the statistical analysis of the RD design.

4. Subscripts indicating individual cases are omitted from the models for the sake of clarity.

REFERENCES

BECK, M. D. (1975) "Development of empirical 'growth expectancies' for the Metropolitan Achievement Tests." Presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

BENTLER, P. M. and J. A. WOODWARD (1978) "A Head Start reevaluation: positive effects are not yet demonstrable." Evaluation Q. 2: 493-510.

CAMPBELL, D. T. and R. F. BORUCH (1975) "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.

CAMPBELL, D. T. and A. ERLEBACHER (1970) "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.

CAMPBELL, D. T. and J. C. STANLEY (1966) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

COOK, T. D. and CAMPBELL, D. T. (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

DAVID, J. L. and S. H. PELAVIN (1977) "Research on the effectiveness of compensatory education programs: reanalysis of data." SRI Project Report URU-4425. Menlo Park, CA: SRI International.

ECHTERNACHT, G. (1980) "Model C is feasible for Title I evaluation." Presented at the annual meeting of the American Educational Research Association, Boston.

——— (1979) "The comparability of different methodologies for ESEA Title I Evaluation." Presented at the annual meeting of the American Psychological Association, New York.

——— (1978) A Summary of the Special Meeting on Model C Held in Atlanta in January 1978. Princeton, NJ: Educational Testing Service. (unpublished)

HANSEN, J. B. (1978) Report of the Committee to Examine Issues Related to the Use of the Norm Referenced Model for Title I Evaluation. Portland, OR: Northwest Regional Educational Laboratory.

HARDY, R. (1978) "Comparison of Model A and Model C in Florida." Atlanta: Educational Testing Service. (Memo)

HOCKING, R. R. (1976) "The analysis and selection of variables in linear regression." Biometrics 32 (March): 1-49.

HOUSE, G. D. (1979) "A comparison of Title I achievement results obtained under USOE Models A1, C1 and a mixed model." Presented at the annual meeting of the American Educational Research Association, San Francisco.

KASKOWITZ, D. H. and L. D. FRIENDLY (1980) "The effect of attrition on the Title I evaluation and reporting system." Presented at the annual meeting of the American Educational Research Association, Boston.

LINN, R. L. (1979) "Measurement of change." Presented at the second annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C.

LONG, J., HORWITZ, S., and A. PELLEGRINI (1979) "An empirical investigation of the ESEA Title I Evaluation system's no-treatment expectation for the special regression model." Presented at the annual meeting of the American Educational Research Association, San Francisco.

MAGIDSON, J. (1977) "Towards a causal model approach for adjusting for pre-existing differences in the non-equivalent control group situation: a general alternative to ANCOVA." Evaluation Q. 1: 399-420.

MURRAY, M. (1978) "Models A and C: theoretical and practical concerns." Presented at the Florida Educational Research Association Convention, Daytona Beach, Florida.

MURRAY, S., J. ARTER, and B. FADDIS (1979) "Title I technical issues as threats to internal validity of experimental and quasi-experimental designs." Presented at the annual meeting of the American Educational Research Association, San Francisco.

NCES [National Center for Education Statistics] (1979) "Quick Survey on Title I Evaluation Models." Washington, D.C.

TALLMADGE, G. K. (1980) "An empirical assessment of norm-referenced evaluation methodology." Mountain View, CA: RMC Research Corporation. (unpublished)

——— (1978) Selecting Students for Title I Projects. Mountain View, CA: RMC Research Corp.

——— (1976) "Cautions to evaluators," in M. J. Wargo and D. R. Green (eds.) Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. New York: McGraw-Hill.

——— and D. P. HORST (1976) "A procedural guide for validating achievement gains in educational projects." Monographs on Evaluation in Education, No. 2. Washington, DC: U.S. Department of Health, Education and Welfare.

TALLMADGE, G. K. and C. T. WOOD (1978) User's Guide: ESEA Title I Evaluation and Reporting System. Mountain View, CA: RMC Research Corporation.

TROCHIM, W. (1980) "The regression-discontinuity design in Title I evaluation: implementation, analysis and variations." Ph.D. dissertation, Northwestern University.

U.S. Office of Education (1979) ESEA Title I Annual Report. Washington, DC.

WICK, J. W. (1978) Title I Elementary and Secondary Education Act: Formation, Function and Purposes, 1965-1978. Chicago, IL: City of Chicago, Department of Education. (unpublished)

William M.K. Trochim is Assistant Professor of Human Service Studies at Cornell University. His research interests are primarily in the area of evaluation methodology, especially quasi-experimental design and analysis, and the study of research implementation.