interests include age trends in delinquency, the generality of deviance, and statistical models of criminal careers.

Gail L. Smith is a research assistant for residential research at Father Flanagan's Boys' Home. She was formerly research assistant for follow-up research there. Her research interests include statistical modeling and adolescent sexual abuse.

Stanley (1991) argues that both random measurement error in the pretest and treatment-effect interactions bias the estimate of the treatment effect when multiple regression is used to analyze the data from a regression-discontinuity design (RDD). Stanley also argues that these biases are so severe that they should cause researchers to consider using statistical procedures other than regression analysis. The authors of the present article disagree. Curvilinearity in the regression of the posttest on pretest scores can be difficult to model, can bias the regression analysis of data from the RDD if not modeled correctly, and therefore should cause researchers to consider alternatives to regression analysis. If the regression surfaces are linear, however, unbiased estimates can be obtained easily via regression analysis, whether or not either random measurement error in the pretest or treatment-effect interactions are present. Improving upon regression analysis is a worthy goal but requires understanding just what are and are not the weaknesses of the method. In addressing these issues, this article elucidates some of the general principles that underlie the use of multiple regression to analyze data from the RDD quasi-experiment.

REPORTS OF THE DEATH OF REGRESSION-DISCONTINUITY ANALYSIS ARE GREATLY EXAGGERATED

CHARLES S. REICHARDT
University of Denver
WILLIAM M. K. TROCHIM
Cornell University
JOSEPH C. CAPPELLERI
New England Medical Center

Multiple regression is the most commonly used procedure for analyzing data from the regression-discontinuity design (RDD). This article examines the use of multiple regression for this purpose and reaches a number of conclusions that differ from those reached by Stanley (1991).

AUTHORS' NOTE: *Preparation of this article was supported in part by Grant U01-AA08778 from the National Institute on Alcohol Abuse and Alcoholism, Grant H133B00011 from the U.S. Department of Education and the National Institute of Mental Health, Grant R01MH46712-01A1 from the National Institute of Mental Health, and Grant ROHS-07782 from the Agency for Health Care Policy and Research. We thank Harry Gollob, Will Shadish, and an anonymous reviewer for helpful comments. Requests for reprints should be addressed to Charles S. Reichardt, Department of Psychology, University of Denver, Denver, CO 80208.*

EVALUATION REVIEW, Vol. 19 No. 1, February 1995 39-63

Stanley (1991; Stanley and Robinson 1990a) argues that, in the RDD quasi-experiment, estimates of effects derived from regression analysis are biased both by random measurement error in the pretest and by interactions between the treatment and the pretest. Stanley (1991) further argues that these biases are so severe that researchers ought to consider abandoning regression analysis in the RDD in favor of alternative procedures. In addition, Stanley (1991) accuses Trochim, Cappelleri, and Reichardt (1991) of misunderstanding statistical estimation (p. 611), obfuscating the real issues (p. 620), and trying to deflect legitimate criticism of the regression-discontinuity design (pp. 614 and 619). The tone of Stanley's (1991) article is captured by the following remarks regarding Trochim et al. (1991):

> It is my thesis that [the article] is entirely misdirected. (p. 606)
>
> [It] serves no useful purpose; if taken seriously, it can only obstruct future progress in quasi-experimental evaluation. (p. 614)
>
> In the effort to defend RD from its own fallibility, Trochim et al. have made RD as useless at [sic] it is unassailable. (p. 615)

We see things differently. We believe that the Achilles heel of regression analysis, when used to analyze data from the RDD, is curvilinearity rather than either random measurement error in the pretest or treatment-effect interactions. That is, curvilinearity in the regression of posttest on pretest scores can bias the estimates of treatment effects in the regression analysis of data from a RDD, and the bias can be difficult to model (Cook and Campbell 1979; Reichardt 1979; Rubin 1977; Trochim 1984). In the absence of curvilinearity, however, estimates of treatment effects in the RDD are not biased by random measurement error in the pretest, and regression analysis can be adapted easily to provide unbiased estimates of treatment effects in the presence of treatment-effect interactions. Alternatives to regression analysis may well be useful in analyzing data from the RDD, and we strongly encourage their development and testing. To improve on regression analysis, however, it is important to understand what are and are not the true sources of its weaknesses.

This article examines the effects of random measurement error in the pretest, treatment-effect interactions, and curvilinearity in the regression analysis of data from the RDD and, in so doing, draws comparisons with the regression analysis of data from the randomized experiment (RE). In particular, three sets of conditions of increasing generality are considered. The first set of conditions restricts the regression of posttest scores on pretest scores to be linear and the treatment effect to be a constant, but it allows the pretest to contain random measurement error. The second set of conditions restricts the regression of posttest on pretest scores to be linear but allows both the effect of the treatment to interact with the pretest and the pretest to contain random measurement error. The third set of conditions allows the regression of posttest on pretest scores to be curvilinear, the effect of the treatment to interact with the pretest, and the pretest to contain random measurement error. While examining these three sets of conditions, we also try to point out a few of the ways in which Trochim et al. (1991) is not entirely misdirected.

DEFINITIONS AND NOTATION

This section provides the background definitions and notation that are needed to specify the three sets of conditions and their consequences. A population of N individuals is assumed to be available for study, where N is large. Each individual in the population can receive either one of two conditions, a treatment protocol or a comparison protocol. Each individual in the population also can be measured on both a pretest and a posttest. An individual's pretest score will be denoted by X_i, where the subscript i can take on values between 1 and N so as to denote the individuals either within the population or within a sample. Similarly, Y_i denotes an individual's posttest score.

Random measurement error in the pretest is allowed. More specifically, the observed pretest scores, the X_i, are assumed to be equal to $T_i + U_i$, where T_i denotes the true pretest score, U_i denotes random measurement error, and U_i is assumed to be uncorrelated with T_i in the population. Note that when we speak of the "pretest score" without using either "observed" or "true" as a modifier, we always mean the observed pretest score. If no measurement error is present, the Us are all zero.

We assume that an individual's posttest score does not depend on whether other individuals receive the treatment protocol or the comparison protocol. This assumption is called the stable-unit-treatment-value assumption (SUTVA) by Rubin (1980).

Let the distribution of posttest scores that would result if every individual in the population were to receive the comparison protocol be called Distribution 1. Let the distribution of posttest scores that would result if every individual in the population were to receive the treatment protocol be called Distribution 2. Any difference between these two distributions is an effect of the treatment protocol, as compared to the comparison protocol (Rubin 1974, 1978; Holland 1986). In particular, the "main effect of the treatment" is defined as the difference between the means of Distribution 2 and Distribution 1. This definition applies whether or not a treatment-effect interaction is present.

To define the size of the average treatment effect at a given observed pretest score, take all the individuals in the population who had the given observed pretest score and determine what their posttest scores would have been had they received the comparison protocol. Call the distribution of these posttest scores "Distribution 1 given the observed pretest score." Also take all the individuals who had the given observed pretest score and determine what their posttest scores would have been had they received the treatment protocol. Call the distribution of these posttest scores "Distribution 2 given the observed pretest score." Then the average treatment effect at a given observed pretest score is defined as the difference between the mean of Distribution 2 given the observed pretest score and the mean of Distribution 1 given the observed pretest score. (The average treatment effect at a given true pretest score is defined identically, except that the observed pretest is replaced by the true pretest.) An interaction of the treatment effect with the pretest is present when the average treatment effect at a given pretest score varies with the pretest scores. In the absence of a treatment-effect interaction with the pretest, the average treatment effect at a given pretest score is equal to the main effect of the treatment.

To estimate the effects of the treatment, the researcher randomly samples *n* individuals from the population of *N* individuals, where *n* is assumed to be a small proportion of *N* (otherwise a finite population correction might be needed), although *n* need not be small in absolute size. Each individual in the sample is measured on the pretest and then assigned to one of the two treatment conditions. Therefore, an individual's pretest score remains the same regardless of the treatment condition.

The assignment to treatment condition might occur in any number of ways. We consider only two. First, the researcher could assign individuals to treatment conditions at random. This type of assignment produces a randomized experiment (RE). Second, the researcher could assign all those individuals in the sample who have observed pretest scores on one side of a given cutoff value, C, to one condition and all those individuals with observed pretest scores on the other side of the cutoff value, C, to the other condition. This type of treatment assignment produces a regression-discontinuity design (RDD).

Following the assignment of individuals in the sample to the treatment conditions, the researcher administers the treatment and comparison protocols and also takes the posttest measure on each individual in the sample. Then knowledge of the assignment mechanism and of the observed pretest and posttest scores for the individuals in the sample is used to estimate the effects of the treatment.

CASE 1: LINEAR RELATIONSHIP AND CONSTANT TREATMENT EFFECT

In this section, we assume that the regression of the posttest scores on pretest scores would be linear in the population if all *N* individuals received the treatment protocol or if all *N* individuals received the comparison protocol. We also assume that the treatment effect is equal to a constant, K, for each individual in the population.

A Simple Regression Model

Consider the following model

$$Y_i = \alpha + \beta_1 Z_i + \beta_2 (X_i - X^*) + \varepsilon_i \qquad (1)$$

where ε_i is a residual, X* is an arbitrary constant that is chosen by the researcher based on the purpose of the analysis (as explained in later sections), and Z_i is an assignment variable that equals 1 if the individual is assigned to the treatment condition and 0 if the individual is assigned to the comparison condition.[1] Equation 1 is a regression model. It is also an analysis of covariance (ANCOVA) model. If the term involving the pretest ($\beta_2[X_i - X^*]$) were omitted, Equation 1 would be an analysis of variance (ANOVA) model. Using ordinary least squares (OLS) regression, this model could be fit to the data on the pretest (X), posttest (Y), and assignment variable (Z) from the sample of *n* individuals so as to produce estimates of the values of α, β_1, β_2, and their standard errors.

Estimating the Effect of the Treatment

Under the given conditions, the value of β_1 that would be produced by fitting Equation 1 using OLS regression would be an unbiased estimate of the main effect of the treatment, K, if the design were either an RE or an RDD. This holds regardless of the value chosen for X* in Equation 1. In both the RE and the RDD, the choice of the value of X* in Equation 1 influences the estimate of α and its standard error but does not alter the estimate of the main effect of the treatment (i.e., the estimate of β_1), the estimate of the regression slope (i.e., the estimate of β_2), or the standard errors of these estimates.

If, in the population, the variance of the posttest were the same at all levels of the pretest in both treatment groups, the estimate of the main effect of the treatment (i.e., the estimate of β_1) would be the best linear unbiased (BLU) estimate that is possible given the available data, in both the RE and the RDD (Johnston 1972, 126; Theil 1971, section 3.4). This means that no alternative estimate could be created by taking a linear combination of the posttest scores so that the estimate would both be unbiased and have a smaller standard error than the OLS estimate. If the variance of the posttest were not the same at all levels of the pretest in the two treatment groups, a generalized least squares (GLS) regression procedure, rather than the OLS procedure, would make β_1 a BLU estimate of the main effect of the treatment, in both the RE and the RDD (Johnston 1972, 210).[2] In essence, this means that it might be quite difficult to come up with an alternative statistical procedure that would produce a more accurate estimate of the main effect of the treatment than would regression analysis, under the given conditions.

Random Measurement Error in the Pretest

In both the RE and the RDD, the pretest can be any measure. The pretest could be conceptually identical to the posttest, or it could be conceptually unrelated to the posttest. If everything else is the same, however, the higher the correlation between the pretest and posttest in the population, the more precise will be the estimate (and the more powerful will be the statistical significance test) of the main effect of the treatment, in both the RE and the RDD. As random measurement error is added to the pretest (or to the posttest for that matter), the correlation between the pretest and posttest decreases. As a result, the precision of the estimate of the main effect of the treatment is diminished, in both the RE and the RDD.

Adding random measurement error to the pretest also attenuates the estimate of the regression slope (i.e., the estimate of β_2) in both the RE and the RDD (Cochran 1968). As more random measurement error is added to the pretest, the estimate of β_2 becomes more biased as an estimate of what the regression slope would be if there were no measurement error in the pretest. If the variance of the random measurement error in the pretest were the same at all values of the true pretest, and if ρ were an unbiased estimate of the reliability of the pretest, the estimate of β_2 divided by ρ would be an asymptotically unbiased estimate of what the regression slope would be if the pretest contained no random measurement error.

The attenuation of the regression slope due to random measurement error in the pretest, however, does *not* bias the estimate of the main effect of the treatment in either the RE or the RDD (Goldberger 1972; Judd and Kenny 1981; Mohr 1988; Reichardt 1979; Trochim 1984; Trochim and Cappelleri 1992). That is, under the present background assumptions (including the assumption that the regression of the posttest on pretest scores is linear), the estimate of β_1 remains an unbiased estimate of the main effect of the treatment in both the RE and the RDD, regardless of the amount of random measurement error in the pretest. Therefore, there is no need to make an adjustment for random measurement error in the pretest when estimating the main effect of the treatment. In addition, all the statements in the preceding section (including the statements about the estimate of β_1 being BLU and the difficulty of improving upon regression analysis when estimating the treatment main effect) hold regardless of the amount of random measurement error in the pretest.

Stanley and Robinson (1990b) mistakenly claimed that the attenuation in the regression slope that is produced by random measurement error in the pretest introduces a bias in the estimate of the main effect of the treatment in the RDD. This mistake was explicitly corrected in Cappelleri et al. (1991).

Dissimilarities Between the Randomized Experiment and the Regression-Discontinuity Design

Although Equation 1 would provide an unbiased estimate of the main effect of the treatment for both the RE and the RDD under the given conditions (regardless of the presence or absence of random measurement error in the pretest), there is at least one very important difference between the estimates of the main effect of the treatment that would be obtained in the RE and in the RDD. The estimate of the main effect of the treatment obtained by fitting Equation 1 with regression analysis would be more precise (and the test of its statistical significance would be more powerful) in the RE than in the RDD. In particular, even under ideal conditions, more than twice as many individuals would be needed in the RDD for the precision of the estimate of the main effect of the treatment (or for the power of the test of statistical significance) to be the same as in the RE (Goldberger 1972; Cappelleri, Darlington, and Trochim 1994). This difference in both precision and power between the RE and the RDD arises because of differences in multicollinearity. The difference in multicollinearity arises because the expected value of the correlation between the pretest (X) and the assignment variable (Z) is zero in the RE but far from zero in the RDD.

Stanley and Robinson (1990a) reach a conclusion that is in conflict with the literature (and with our conclusion above) about the relative power of the statistical significance test of the main effect of the treatment in the RDD and the RE:

> Looking at the best cases . . . , the statistical power of RD [our RDD] {3.6%, 32%, 80% and 100%} is slightly better than the corresponding values of TE [our RE] {2.2%, 23%, 68.6% and 99.8%}. Even when the R^2 of the RD model is reduced to 0.6, RD is only slightly less powerful than TE. (p. 11)

Stanley and Robinson (1990a) reached these discordant conclusions because they calculated power using different statistical models in the two designs. Specifically, they computed power for the RDD using Equation 1, which included the pretest as a covariate, but computed power for the RE using the ANOVA model, which is identical to Equation 1 except that it excludes the pretest from the model (i.e., it excludes the $\beta_2[X_i - X^*]$ term). As a result, Stanley and Robinson (1990a) compared apples with oranges. If a pretest is available for one design, an appropriate comparison of precision and power requires having a pretest available for the other design, using the same statistical model.

The omission of the pretest in the analysis of data from the RE when drawing comparisons with the analysis of data from the RDD leads Stanley and Robinson (1990a, 17) to the incorrect conclusion that "in ideal conditions, these quasi-experiments [meaning the RDD among other designs] may be as good as the 'real thing' [meaning the RE]." The truth is that even under ideal conditions, the RE enjoys a decided advantage over the RDD in terms of precision and power.

CASE 2: LINEAR RELATIONSHIP AND NONCONSTANT TREATMENT EFFECT

In this section, we retain the assumption that the regression of posttest on pretest scores would be linear in the population if all *N* individuals received the treatment condition or if all *N* individuals received the comparison condition. Instead of assuming that the treatment effect is a constant, however, we now allow for a treatment-effect interaction. Specifically, we allow the average treatment effect at a given pretest score to vary linearly with the value of the pretest score. In other words, we assume that the average effect of the treatment at a given pretest score is equal to X

$$K + L(X - \mu_x) \tag{2}$$

where K and L are constants and μ_x is the overall mean of the pretest scores in the population. Under these assumptions, K equals the main effect of the treatment and L represents the treatment-effect interaction, which equals the difference in the average treatment effect for individuals whose pretest scores differ by one unit.

Equation 1 does not allow for a treatment-effect interaction. Nevertheless, the estimate of β_1 obtained from Equation 1 would still be an asymptotically unbiased estimate of the main effect of the treatment (K) in the RE, although the standard error of the estimate would be biased (Johnston 1972). In contrast, using Equation 1 in the RDD would, under the present assumptions, generally produce asymptotically biased estimates of the main effect of the treatment. The bias, however, is easily removed. All that is needed is the slightly more complex model given below, which differs from Equation 1 because a term to represent a linear treatment-effect interaction has been added:

$$Y_i = \alpha + \beta_1 Z_i + \beta_2(X_i - X^*) + \beta_3 Z_i(X_i - X^*) + \varepsilon_i. \tag{3}$$

In this equation, X* is an arbitrary constant chosen by the researcher, as described below.

Estimating the Effect of the Treatment at a Given Observed Pretest Score

The estimate of β_1 derived from Equation 3 depends on the value chosen for X*. Setting X* equal to some arbitrary pretest value, W, makes β_1 an unbiased estimate of the average treatment effect for an observed pretest score equal to W, in both the RDD and the RE. In particular, setting X* equal to C makes β_1 an unbiased estimate of the average treatment effect at the assignment-cutoff point in the RDD.

In addition, if X* is set equal to the mean of the pretest scores in the sample, $\overline{X}$, the estimate of β_1 is an asymptotically unbiased estimate of the main effect of the treatment in both the RDD and the RE. In the RDD, however, it is usually recommended that researchers pay more attention to the estimate of the average treatment effect at the assignment-cutoff score than to the estimate of the main effect of the treatment (Campbell 1969; Reichardt 1979; Trochim 1984). This is because less extrapolation is generally involved in estimating the treatment effect at the cutoff score than in estimating the main effect of the treatment, and therefore the estimate of the treatment effect at the cutoff score tends to be more credible than the estimate of the main effect of the treatment.[3]

If the variance of the posttest scores is the same across levels of the pretest under both the treatment and the comparison conditions, the OLS estimate of the treatment effect (β_1) at the assignment-cutoff point (or at any other given pretest score) is BLU (Johnston 1972; Theil 1971). If the variances of the posttest scores are not constant, a GLS procedure would provide estimates that are BLU.

The above results hold regardless of the amount of random measurement error in the pretest. In other words, when the regression of posttest on pretest scores is linear, neither treatment-effect interactions nor random measurement error bias the estimate (β_1) of the treatment effect at a given observed pretest score if the model in Equation 3 is fit to the data using regression analysis, in either the RDD or the RE. In addition, it might be quite difficult to get more accurate estimates of the treatment effect than obtained by fitting Equation 3 using regression analysis, in either the RDD or the RE. These conclusions appear to be at odds with Stanley's (1991) conclusions.

Estimating the Treatment Effect Interaction

If Equation 3 were fit repeatedly with different values of X*, the estimates of β_1 would vary linearly with the value of X*, in both the RE and the RDD. That is, for every one unit change in X*, the estimate of β_1 would change by a constant. This constant would be equal to the estimate of β_3 (which would be the same regardless of the value chosen for X*) and would be an unbiased estimate of the treatment-effect interaction, L, in Equation 2 above in both the RE and RDD. These estimates of the treatment-effect interaction are BLU under the same conditions as specified above for the BLUness of the estimate of the effect of the treatment at a given pretest score. In addition, these results hold regardless of the amount of random measurement error in the pretest.

In practice, if a researcher is in doubt about the presence of a treatment-effect interaction, there is some benefit to including an interaction term in the analysis as in Equation 3. On the other hand, one potential disadvantage, especially in the RDD, is that if a treatment-effect interaction is not present or is very small, including an interaction term in the analysis may do more harm (by lowering the precision and power of the analysis because of multicollinearity) than good (by reducing bias).

Misunderstanding the Role Played by X*

Because the estimate of β_1 depends on the researcher-specified value of X* in Equation 3, as described above, researchers need to interpret the estimate of β_1 in light of the value that is chosen for X*. Stanley (1991)

misunderstands the role played by X* when he concludes, mistakenly, that β_1 in Equation 3 is biased:

> Trochim et al. simulate this precise combination of circumstances which they term, *Model 1* [Equation 3], in their first simulation study. Yet their results show quite clearly that *there will be a bias in the OLS estimate of the treatment effect* [β_1]. (p. 612, italics in original)
>
> Because the estimation model that Trochim (1990) advocates leads to estimates of [β_1] that systematically vary from the true [β_1] and because H_0: [β_1] = 0 is his recommended hypothesis test for a treatment effect, Trochim's RD evaluation method is biased. (p. 613)
>
> The point is that an evaluator who competently and correctly applies statistical techniques recommended by RD "methodologists" will likely obtain estimates that are systematically off and may find effects that are not actually there. (p. 613)

Stanley (1991) reaches these conclusions because he mistakenly believes that the estimate of β_1 in Equation 3 will be an estimate of the main effect of the treatment regardless of the value of X*. As explained above, the estimate of β_1 will be an estimate of the main effect of the treatment only if X* is set equal to the average pretest score, $\overline{X}$, in the sample. If, instead, X* is set equal to the value of the cutoff score (C), the value of β_1 will estimate the average effect of the treatment at the cutoff score, which can be quite different from the estimate of the main effect of the treatment. In Trochim et al.'s Model 1, which is the focus of Stanley's attention, the value of X* was set equal to the cutoff score C (as is clearly seen, for example, in Stanley 1991, 612). Therefore, it makes little sense for Stanley (1991) to expect the value of β_1 from this model to estimate the main effect of the treatment, but this is what Stanley does. As a result, Stanley's criticisms of "Trochim's RD evaluation method" are misguided. To "competently and correctly apply statistical techniques recommended by RD methodologists," one must understand the role played by X*. A misunderstanding of this role largely accounts for Stanley's (1991) mistaken belief that estimates of treatment effect in a regression analysis are biased by treatment-effect interactions.

A misunderstanding of the role played by X* also undergirds another error in Stanley (1991). In particular, Stanley (1991) mistakenly argues that Trochim et al. (1991) propose a model containing a "parameter," T_c, that cannot be estimated. Stanley (1991, 615) states his case:

> Conventional regression analysis (OLS or any other) cannot be used to estimate Trochim et al.'s model. Mathematically, it is simply not possible to solve for five unknown, independent parameters from four knowable estimates. Their model offers a true impasse

> for the evaluator. There is no a priori basis on which to infer T_c, and yet unknown values of T_c can mask the true program effects.

Stanley (1991, 619-20) raises the same incorrect criticism again four pages later:

> Trochim et al. add the heretofore unidentified, free parameter, T_c, to explain anomalies found in Monte Carlo simulations. By Popper's methodology, such an ad hoc auxiliary hypothesis may be added to a system only if its "introduction does not diminish the degree of falsifiability or testability of the system in question" (Popper 1959, 82-83). The role of T_c as employed by Trochim et al. is quite the opposite; its purpose is to protect RD from further Monte Carlo testing. This tactic must be rejected. Even if their explanation of these simulation anomalies were technically correct, their interpretation serves only to hinder future progress in understanding this quasi-experimental evaluation design.

These statements are incorrect because, contrary to Stanley's (1991) assertions, T_c is not a parameter and, in any case, it would serve no useful purpose were it to be estimated. This is because T_c in the simulation model in Trochim et al. (1991) plays the same role as X* in Equation 3 above. Rather than a parameter to be estimated, X* is a constant specified a priori to be equal to a given pretest score so as to estimate the average treatment effect at that pretest score. The bottom line is that regardless of the value chosen for T_c in the simulation model in Trochim et al. (1991), Equation 3 will provide unbiased estimates of these treatment effects. Therefore, it is not possible that "unknown values of T_c can mask the true program effects," as Stanley (1991) mistakenly claims. For the same reasons, it is also not possible that T_c protects "RD from further Monte Carlo testing" or that our interpretation of it "serves only to hinder future progress in understanding this quasi-experimental evaluation design."

To take account of the presumed (but, in fact, imaginary) difficulty that he believes is introduced by the presence of T_c, Stanley (1991, 617) proposes "an entirely different strategy for the statistical test of a treatment effect—one that does not depend on the unknown T_c." In fact, Stanley's procedure does not qualify as "entirely different" and only further illustrates his misunderstanding of the role played by T_c. This is because his "entirely different" procedure is identical to fitting Equation 3 with X* set equal to zero, and then performing a statistical significance test to determine if the values of β_1 and β_3 are simultaneously equal to zero (which reveals whether or not treatment effects exist but not their size). Researchers can use this statistical test if they like, but the procedures for estimating the size (rather than just the presence) of treatment effects, which we have described both above and in Trochim et al. (1991), will almost always be more informative.

Estimating the Effect of the Treatment at a Given True Pretest Score

In a preceding section, we explained how to assess the treatment-effect interaction in terms of the *observed* pretest scores. It is also possible to assess the treatment-effect interaction in terms of the *true* pretest scores. We believe the former is more common and usually more appropriate than the latter for practical purposes, but the latter is sometimes desired, is relevant to some of Stanley's (1991) comments, and so will be discussed in the present section.

As the preceding sections show, estimates of β_1 provide unbiased estimates of the average effect of the treatment for given pretest scores (given the appropriate specification of the value of X*) and the estimate of β_3 provides an unbiased estimate of the treatment-effect interaction, in terms of *observed* pretest scores. Because estimates of treatment effects for *observed* pretest scores are the information most often desired (and usually most appropriate to report), nothing needs to be done to take account of the presence of random measurement error in the pretest. If a researcher wishes to obtain estimates of treatment effects in terms of *true* pretest scores, however, corrections for the effects of random measurement error in the pretest must be made. The same correction is required in the RE as in the RDD.

Regardless of the degree of random measurement error in the pretest, the estimate of β_1 in Equation 3 remains an asymptotically unbiased estimate of the main effect of the treatment as long as X* is set equal to $\overline{X}$. If the variance of the measurement error in the pretest is constant across the values of the true pretest score and if ρ is an unbiased estimate of the reliability of the pretest, then β_2/ρ is an asymptotically unbiased estimate of what the regression slope would be in the comparison group if the pretest contained no measurement error. In addition, the value of β_3/ρ would be an asymptotically unbiased estimate of the value of L if the pretest contained no measurement error. Finally, the estimate of β_1 from Equation 3 would be an asymptotically unbiased estimate of the average treatment effect for individuals with *true* pretest score equal to an arbitrary pretest value of W if X* is set equal to $\overline{X} + (W - \overline{X})/\rho$. Note that these corrections have to be made in both the RE and the RDD to get asymptotically unbiased estimates in terms of true pretest scores, but note that these corrections are seldom made in practice in either the RDD or the RE because researchers usually prefer estimates of treatment effects in terms of observed scores.

Using simulations, Stanley and Robinson (1990a) discovered that the values of β_1 and β_3 can be biased as estimates of treatment effects in terms of *true* pretest score, but they did not realize that these biases have the very simple and correctable form that we just described. In addition, Stanley and Robinson (1990a) drew comparisons between the RE and the RDD and criticized the RDD because the estimates of β_1 and β_2 are biased as estimates of the treatment effect in terms of *true* pretest scores in the RDD, but they failed to realize that the same biases are present in the RE. There are many legitimate reasons to prefer the RE to the RDD, but this is not one of them.

If researchers are to reach appropriate conclusions about treatment effects in terms of true, rather than observed, pretest scores, adjustments for the effects of random measurement error in the pretest must be made not only when fitting Equation 3 but also when generating data in simulations. The necessary adjustments parallel those described above. For example, one of the ways to correct for the effects of measurement error when simulating data is to make an adjustment to the value of T_c in the simulation, just as a correction was made for the values of X* above. A formula for this correction is provided in Trochim et al. (1991). Stanley (1991, 611) criticizes this correction procedure:

> It is a great irony that Trochim et al. advocate correcting the true-score centering value, T_C, for the fallibility of the pretest. . . . To correct a population parameter for measurement error is a revealing conceptual misunderstanding of statistical estimation.

In fact, T_c is not a population parameter (as previously explained), correcting it is quite proper, and the source of the "conceptual misunderstanding" lies not in Trochim et al. (1991).

CASE 3: CURVILINEAR RELATIONSHIP

In this section, we allow the regression of posttest on pretest scores to be curvilinear in the population. We also allow for a treatment-effect interaction that can be curvilinear in the population. This means that, in the population, the regression of posttest on pretest scores may have a different curvilinear shape in the treatment condition than in the comparison condition.

Estimating Treatment Effects

Even though neither Equation 1 nor Equation 3 allows for a curvilinear regression surface, if X* is set equal to $\overline{X}$, the estimate of β_1 from either Equation 1 or Equation 3 would be an asymptotically unbiased estimate of the main effect of the treatment in the RE, although the standard error of the estimate would be biased. In contrast, the estimate of β_1 derived from either Equation 1 or Equation 3 is likely to be an asymptotically biased estimate of the main effect of the treatment in the RDD. In addition, both estimates of the average treatment effect at a given pretest score and estimates of the treatment- interaction effects are likely to be asymptotically biased, in both the RDD and the RE. To remove these biases, the researcher must properly take account of the curvilinearity in the regression surfaces. This will not always be an easy task. Therefore, it will not always be easy to obtain unbiased estimates of the treatment effects in the RDD when the regression surfaces are curvilinear.

One potential way to correct for biases resulting from curvilinearity is to try to transform the data so that the regression surface is linear in both the treatment and comparison conditions, and then apply either Equation 1 or Equation 3 to the transformed data (e.g., Hamilton 1992). The most difficult step in this approach is discovering a formula that will perform the proper transformation. Unfortunately, there is no mechanical procedure that will guarantee that the correct formula has been found, especially when curvilinearity is produced by sources such as floor and ceiling effects.

Another approach is to add polynomial terms to Equation 1 or Equation 3 so that the curvilinear regression surface in the untransformed data is modeled properly (Cappelleri and Trochim 1994). For example, suppose the regression surface in the comparison condition is quadratic. Further, suppose the effect of the treatment is quadratic in the pretest scores. That is, suppose the average effect of the treatment for individuals with pretest scores equal to X is

$$K + LX + MX^2 \qquad (4)$$

where K, L, and M are constants. Then K is the effect of the treatment when X = 0, and L and M are the linear and quadratic effects, respectively, of the interaction of the treatment with the pretest.[4] Then quadratic terms could be added to Equation 3 to produce the following equation:

$$Y_i = \alpha + \beta_1 Z_i + \beta_2(X_i - X^*) + \beta_3 Z_i(X_i - X^*) + \beta_4(X_i - X^*)^2 + \beta_5 Z_i(X_i - X^*)^2 + \epsilon_i . \qquad (5)$$

Under the present conditions, setting the value of X* equal to W would make the estimate of β_1 an unbiased estimate of the average treatment effect for a pretest score equal to W (Johnston 1972). In particular, setting the value of X* equal to the assignment-cutoff score, C, in a RDD would make the estimate of β_1 an unbiased estimate of the effect of the treatment at the cutoff score. In addition, regardless of the value chosen for X*, the estimates of β_3 and β_5 would be unbiased estimates of L and M, respectively, in both the RDD and the RE.[5] If the variance of the posttest scores is the same across levels of the pretest in both treatment groups, the estimate of β_1 derived from Equation 5 would be BLU, in both the RE and the RDD. Otherwise, a GLS procedure would be required for BLUness, in both the RE and the RDD. All these results hold regardless of the degree of random measurement error in the pretest.

If the regression surfaces were cubic rather than quadratic, Equation 5 would need to contain additional (cubic) terms if unbiased estimates of the effects of the treatment were to be produced. In theory, any curvilinear regression surface could be properly modeled if enough polynomial terms were added. In practice, polynomial regression can fail for two reasons. First, an infinite number of polynomial terms is required to perfectly fit some curvilinear shapes (such as might be caused by floor or ceiling effect, for example). Second, adding even a few polynomial terms can greatly increase multicollinearity, which can make the regression estimates unstable. In fact, just adding quadratic terms might greatly increase the sample size that is needed to obtain stable estimates in Equation 5, as compared to Equation 3.

Correctly Modeling Curvilinearity

How does the researcher know if the regression surfaces have been modeled correctly so that the estimates of the effects of the treatment are unbiased? This is a critical question, and it is especially difficult to answer in the RDD. In the RE, the available data cover the complete range of pretest scores in both the treatment and comparison conditions. In contrast, because of the way individuals are assigned to treatments in the RDD on the basis of the assignment-cutoff score, the available data cover only a limited range of pretest scores in each of the treatment and comparison conditions. This will generally make guessing, tentative modeling, and subsequent checking of the correct form for the regression surfaces much more difficult in the RDD than in the RE. In addition, it will be much more difficult to distinguish curvilinear regression surfaces from interactions in the RDD than in the RE. (Because bias can increase with increasing extrapolation, the difficulty of modeling curvilinearity provides another reason for preferring, in the RDD, the estimate of the effect of the treatment at the assignment-cutoff score to the estimate of the effect of the treatment at any other pretest score.)

To help in assessing and modeling curvilinearity in the regression surfaces, Trochim et al. (1991) suggested plotting the data using moving averages. Stanley (1991, 621) takes issue with this technique:

> The method of moving averages used by Trochim et al. (1991) to uncover nonlinearities is not likely to be useful in practice. First, it is a data "hog." They used 10,000 observations and moving averages of 100 data points to show how uniform true scores will distort the X-Y relationship. When most practical applications have 50 to 100 or so observations, it is unlikely that moving averages will be a very powerful tool in detecting any nonlinearity. Second, what is the appropriate statistical test? . . . Surely, Trochim et al. (1991) do not recommend that the evaluator use moving average data in the regression analysis of RD. To do so would induce yet another well-known regression misspecification, serial dependence (or autocorrelation).

Stanley's (1991) concerns are misplaced. First, Trochim et al. (1991) neither recommend nor imply that moving average data be included in the regression analysis to estimate treatment effects in the RDD. In fact, we see no benefit to doing so. Instead, moving averages are to be used only in drawing pictures so as to make the nature of any curvilinearity easier to discern. Second, a large number of data points is not required to use moving averages. Trochim et al. (1991) used 10,000 data points in an example only so that there would be no uncertainty about the shape of the population distribution. In practice, moving averages can be used with virtually any size data set. For example, we would not hesitate to apply the method with only 50 data points. Third, although formal statistical tests could be applied, simple observation and common sense will generally be more useful than a hypothesis test in assessing the nature of any curvilinearity that is present.

Trochim and colleagues (Trochim 1984; Trochim, Cappelleri, and Reichardt 1991; Cappelleri and Trochim 1994) provide other suggestions for discerning and modeling the correct shape of the regression surfaces, in both the RE and the RDD. Interested readers are directed to these references.

Sources of Curvilinearity

Curvilinearity in a regression surface can arise from many sources. Curvilinearity can arise either because the "true" relationship between two variables is curvilinear or because a relationship that is "truly" linear is made curvilinear by floor or ceiling effects. Curvilinearity in the regression of posttest on pretest scores may (or may not) also arise because of random measurement error in the pretest. For example, suppose that Y and T are both distributed normally and that the regression of Y on T is linear. In this case, adding normally distributed, random measurement error to T so as to create X scores will attenuate the regression of Y on X as compared to the regression of Y on T, but the regression surface will remain linear. If the distribution of the T scores and the distribution of the random measurement error that is added to the T scores to create the X scores are not from the same family (e.g., one is normally and the other is uniformly distributed), the regression of Y on X can become curvilinear (Cochran 1970).

In their simulations of the RDD, Stanley and Robinson (1990a) used a uniform distribution for the true pretest scores and a normal distribution for the random measurement error in the pretest. As a result, the regression of the posttest on pretest scores, which was linear in the true scores, became curvilinear in the observed scores. Stanley and Robinson (1990a) did not realize that their simulation had produced a curvilinear regression surface in the observed scores, fit a regression model like Equation 3 that did not allow for the presence of curvilinearity, found that the estimates of the treatment effects in the RDD were biased, and attributed the source of the bias to the presence of random measurement error per se rather than to the presence of curvilinearity.

A similar misattribution arose in a well-known article by Campbell (1969). In that article, Campbell reported the results of a simulation in which the true pretest scores were distributed uniformly, whereas the random measurement error was distributed normally. This produced curvilinearity in the regression surfaces, a bias in the estimate of the treatment effect because a regression model assuming linear regression surfaces was fit to the data, and the conclusion that something new (and as yet unknown) must be wrong with the OLS regression as applied to data from the RDD. This misattribution was corrected in subsequent reprinting (e.g., Campbell 1971, 1983), after Campbell (1984, 20) realized that these biases were simply "another example of the subtle effects of overlooking slight degrees of curvilinearity."

Like Campbell (1969), Stanley and Robinson (1990a) thought that they had discovered a new source of bias in the regression analysis of data from the RDD. In fact, they had merely demonstrated what had long been well known; namely, that fitting a linear model to curvilinear data can produce a bias in the RDD (Cook and Campbell 1979; Reichardt 1979; Rubin 1977; Trochim 1984). One of the explicit purposes of Trochim et al. (1991) was to demonstrate visually how curvilinearity can be caused by random measurement error in the pretest and to describe the potential biases that can thereby arise in the RDD, so as to remove the repeated misunderstandings of this issue that have occurred. Apparently, we were not completely successful, because in his reaction to Trochim et al. (1991), Stanley (1991) still appears not to understand the source of the bias. In addition, Stanley is wrong when he claims that:

> To protect RD from potential specification bias, Trochim et al. wish to prohibit the use of uniform true scores in Monte Carlo simulation. (p. 607)

> In fact, Trochim has no rational basis on which to restrict the simulator's choice of a true-score distribution to the normal or any other specific distribution. (p. 608)

Trochim et al. (1991) do not place restrictions on the use of uniform, or any other, distributions in simulations: We encourage simulators to use whatever distributions they believe are appropriate. All we were attempting to curtail were misunderstandings of the conditions under which random measurement error does and does not introduce bias.

Although it is important to recognize that curvilinearity can arise because of random measurement error, we suspect that curvilinearity produced by random measurement error is likely to be small relative to the curvilinearity that arises from other sources. For example, we suspect that far more curvilinearity arises from curvilinearity in the relationship between true scores or from floor or ceiling effects than from random measurement error.

ALTERNATIVES TO REGRESSION ANALYSIS

We wholeheartedly concur with Stanley's (1991, 619) conclusion that "progress only occurs when we uncover anomalies or problems in simulations (or elsewhere) and as a result construct new methods that remedy these difficulties." Two processes are implicit in this statement: uncovering problems and constructing new methods. Below we address additional claims that Stanley (1991) makes about regression analysis, with reference to each of these two steps.

First, to make progress you have to "uncover anomalies or problems." The bigger the problem, the greater is the potential for advance, but if you misattribute the source of a difficulty, you may be trying to solve a problem that does not exist. In this vein, Stanley (1991, 609) argues that OLS regression analysis is inappropriate if measurement error is present in an independent or explanatory variable (such as the pretest in RDD) because measurement error makes the variable stochastic:

> Classical regression analysis requires that the independent variable be nonstochastic, that is, fixed or known. . . . However, with measurement error, the explanatory variables are no longer fixed or known, and regression analysis is no longer appropriate.

This is simply not true. For simplicity, introductory chapters in econometric texts, for example, often begin by assuming that the explanatory variables in regression analysis are fixed. This assumption is quickly dropped in later chapters, where it is demonstrated that regression analysis applies equally well with stochastic explanatory variables (Johnston 1972, 29-32 and 274-7; Goldberger 1964, 266-72; Kmenta 1971, section 3.3). In other words, just because a regression model includes explanatory variables that are stochastic does not make regression analysis inappropriate either for the RE or the RDD. Indeed, the vast majority of regression analyses involve stochastic explanatory variables because the vast majority of regression analyses include independent variables that contain measurement error.

Stanley (1991, 620) also argues that regression analysis requires bivariate normal distributions:

> For example, the estimation of RD using OLS may well lead to biased program assessments because the real world is not necessarily normal. To protect the evaluator from the statistical artifact that RD contains in the presence of nonnormal true score, a caveat should be added to its application. That is, RD may be safely applied only if the pretest/posttest distribution may be adequately described by the bivariate normal.

In fact, the use of regression analysis in the RDD does *not* require that the pretest/posttest distribution be bivariate normal (Johnston 1972; Kmenta 1971; Theil 1971). The results on bias that have been presented in this article hold under the conditions stated herein without further restrictions on the shape of the distributions, and nowhere has this article specified that the distributions must be bivariate normal. Additional restrictions are required, however, if hypothesis tests and confidence intervals are to be exactly valid, although the procedures will be approximately valid under a much wider range of conditions. Even the exact validity of hypothesis tests and confidence intervals does not require bivariate normality. Neither the pretest nor the posttest scores need be distributed normally; all that is required for the exact validity of hypothesis tests and confidence intervals is that the residuals in the model be normally distributed. The central limit theorem furthermore implies that hypothesis tests and confidence intervals can be approximately valid even when the distributions of the residuals deviate from normality. The analysis of covariance, however, is more sensitive than the analysis of variance to violations of the assumption that the residuals are normally distributed (Atiqullah 1964; Elashoff 1969; Glass, Peckham, and Sanders 1972).

Second, to make progress you have to "construct new methods that remedy" the difficulties that are discovered. Stanley (1991) cites, with approval, a technique proposed by Robbins and Zhang (1988, 1989, 1990) and states (Stanley 1991, 618) that the Robbins and Zhang method not only "holds great promise" but "escapes all of the difficulties we have discussed."[6]

Although we strongly support further examination of the method proposed by Robbins and Zhang (as well as other methods), the Robbins and Zhang method is as yet largely untested, and we would be surprised if it too did not have limitations and weaknesses. Along this line, it could also be of benefit to consider solving some of the analytical problems of the RDD by elaborating the design or by combining the RDD with the RE. In this regard, Cappelleri and Trochim (1994), Trochim (1984, 1990), Trochim et al. (1991), and Trochim and Cappelleri (1992) offer a number of useful suggestions. Rubin (1977) also provides useful analytic results.

CONCLUSIONS

With data from an RE, asymptotically unbiased estimates of the effects of the treatment can be obtained with regression analysis even if the researcher cannot correctly model the shape of the regression of the posttest on the pretest. In contrast, with data from an RDD, asymptotically unbiased estimates of the effects of the treatment are unlikely to be obtained with regression analysis unless the researcher correctly models the shape of the regression of the posttest on the pretest. Unfortunately, it is not always easy to model correctly the shape of the regression surface. In addition, even when the regression surface is correctly modeled, treatment effects can be estimated with greater precision in the RE than in the RDD.

On the other hand, the RDD may be feasible in instances in which the RE is not because of either ethical or practical constraints. In addition, the RDD may be less susceptible than the RE to threats to validity such as differential attrition, resentful demoralization (e.g., Fetterman 1982), and compensatory rivalry (Cook and Campbell 1979). As a result, circumstances may arise in which the RDD is preferable to the RE, in spite of the relative weaknesses that have been discussed in this article.

Stanley (1991; Stanley and Robinson 1990a, 1990b) argues that random measurement error and treatment-effect interactions are sources of bias in the regression analysis of data from the RDD. In addition, Stanley (1991, 621) argues that these biases are so severe that "we might be better off to remove the 'regression' from regression-discontinuity design." In contrast, we (and others such as Cook and Campbell [1979] and Rubin [1977]) believe that curvilinearity is a more serious source of difficulty for the regression analysis of data from the RDD than either random measurement error in the pretest or treatment-effect interactions. In the absence of curvilinearity, unbiased estimates of the treatment effect are easily obtained via regression analysis, whether or not there is random measurement error in the pretest or treatment-effect interactions. In addition, in the absence of curvilinearity, it may be quite difficult to create estimates of treatment effects in the RDD that are more precise than the estimates derived from regression analysis. When curvilinearity is present, however, estimates of the treatment effects in the RDD may be biased because of difficulties in modeling the shape of the regression surface, as we have consistently maintained. With this in mind, the flaws in regression analysis that Stanley (1991) reports are greatly exaggerated.

NOTES

1. Alternatively, the Z variable could be coded so that 1 denotes the treatment group and –1 denotes the comparison group. Using 1 and –1 rather than 1 and 0 as in the text, however, would alter the estimate the treatment effect (i.e., the estimate of β_1) and its standard error by a factor of 0.5. This means that a researcher using 1 and –1 coding would have to multiply the estimate of β_1 and its standard error by 2 to obtain the same results as would be obtained if 1 and 0 coding were used, as is being assumed.

2. In addition, if the distribution of the posttest were normal at all levels of the pretest, either OLS or GLS regression would produce treatment effect estimates that had the minimum variance among the class of all possible—not just linear—unbiased estimators (Johnston 1972, 210).

3. In addition, the estimate of the main effect of the treatment would need to be interpreted in the light of a treatment-effect interaction.

4. One could make K equal to the main effect of the treatment by altering Equation 4 to be $K + L(X - \mu_x) + M(X^2 - \Gamma)$, where $\Gamma = \Sigma X^2/N$, which is the mean of the square of the pretest scores (*not* the square of the mean) in the population.

5. Estimating the main effect of the treatment would require the following, alternative formulation of Equation 5:

$$Y_i = \alpha + \beta_1 Z_i + \beta_2(X_i - X^*) + \beta_3 Z_i(X_i - X^*) + \beta_4(X_i^2 - X^{**}) + \beta_5 Z_i(X_i^2 - X^{**}) + \varepsilon_i.$$

Then setting X^* equal to $\overline{X}$, which is the mean of the pretest scores in the sample, and setting X^{**} equal to $\Sigma X^2/n$, which is the mean of the square of the pretest scores (*not* the square of the mean) in the sample, the estimate of β_1 would be an asymptotically unbiased estimate of the main effect of the treatment, in both the RDD and the RE.

6. It is ironic that the body of Stanley's (1991) article casts Trochim et al. (1991) as inhibiting rather than promoting the consideration of new methods, given that in a footnote Stanley (1991, 622) thanks "Professor Trochim for bringing [the Robbins and Zhang] references to my attention."

REFERENCES

Atiqullah, M. 1964. The robustness of the covariance analysis of a one-way classification. *Biometrika* 51:365-72.

Campbell, D. T. 1969. Reforms as experiments. *American Psychologist* 24:409-29.

———. 1971. Reforms as experiments. In *Readings in evaluation research*, edited by F. G. Caro, 233-61. New York: Russell Sage Foundation.

———. 1983. Reforms as experiments. In *Handbook of evaluation research: University edition*, edited by E. L. Streuning and M. B. Brewer. Newbury Park, CA: Sage.

———. 1984. Foreword. In *Research design for program evaluation: The regression-discontinuity approach*, by W.M.K. Trochim, 15-43. Newbury Park, CA: Sage.

Cappelleri, J. C., R. B. Darlington, and W.M.K. Trochim. 1994. Power analysis of cutoff-based randomized clinical trials. *Evaluation Review* 18:141-52.

Cappelleri, J. C., and W.M.K. Trochim. 1994. An illustrative statistical analysis of cutoff-based randomized clinical trials. *Journal of Clinical Epidemiology* 47:261-70.

Cappelleri, J. C., W.M.K. Trochim, T. D. Stanley, and C. S. Reichardt. 1991. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review* 15:395-419.

Cochran, W. G. 1968. Errors of measurement in statistics. *Technometrics* 10:637-66.

———. 1970. Some effects of errors of measurement on linear regression. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* 1:527-39.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Elashoff, J. D. 1969. Analysis of covariance: A delicate instrument. *American Educational Research Journal* 6:383-401.

Fetterman, D. M. 1982. Ibsen's baths: Reactivity and insensitivity. *Educational Evaluation and Policy Analysis* 4:261-79.

Glass, G. V., P. D. Peckham, and J. R. Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Education Research* 42:237-88.

Goldberger, A. S. 1964. *Econometric theory*. New York: Wiley.

———. 1972. *Selection bias in evaluating treatment effects: Some formal illustrations*. Discussion Paper 123-72. Madison: University of Wisconsin, Institute for Research on Poverty.

Hamilton, L. C. 1992. *Regression with graphics: A second course in applied statistics*. Pacific Grove, CA: Brooks/Cole.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81:945-70.

Johnston, J. 1972. *Econometric methods*. 2d ed. New York: McGraw-Hill.

Judd, C. M., and D. A. Kenny. 1981. *Estimating the effects of social interventions*. London: Cambridge University Press.

Kmenta, J. 1971. *Elements of econometrics*. New York: Macmillan.

Mohr, L. B. 1988. *Impact analysis for program evaluation*. Chicago: Dorsey.

Popper, K. R. 1959. *The logic of scientific discovery*. New York: Basic Books.

Reichardt, C. S. 1979. The statistical analysis of data from nonequivalent group designs. In *Quasi-experimentation: Design and analysis issues for field settings*, in T. D. Cook and D. T. Campbell, 147-205. Chicago: Rand McNally.

Robbins, H., and C.-H. Zhang. 1988. Estimating a treatment effect under biased sampling. *Proceedings of the National Academy of Sciences* 85:3670-2.

———. 1989. Estimating the superiority of a drug to a placebo when all and only those patients at risk are treated with the drug. *Proceedings of the National Academy of Sciences* 86:3003-5.

———. 1990. *Estimating a treatment effect under biased allocation.* Working paper, Rutgers University, Institute of Biostatistics and Department of Statistics.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688-701.

———. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2:1-26.

———. 1978. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 6:34-58.

———. 1980. Discussion of paper by D. Basu. *Journal of the American Statistical Association* 75:591-3.

Stanley, T. D. 1991. "Regression-discontinuity design" by any other name might be less problematic. *Evaluation Review* 15:605-24.

Stanley, T. D., and A. Robinson. 1990a. "Second best" evaluation design: Regression-discontinuity or abbreviated time series? Paper presented at the annual conference of the American Evaluation Association, October, Washington, DC.

———. 1990b. Sifting statistical significance from the artifact of regression-discontinuity design. *Evaluation Review* 14:166-81.

Theil, H. 1971. *Principles of econometrics.* New York: Wiley.

Trochim, W.M.K. 1984. *Research design for program evaluation: The regression-discontinuity approach.* Newbury Park, CA: Sage.

———. 1990. The regression-discontinuity design. In *Research methodology: Strengthening causal interpretations of nonexperimental data*, edited by L. Sechrest, P. Perrin, and J. Bunker. Washington, DC: U.S. Department of Health and Human Services.

Trochim, W.M.K., and J. C. Cappelleri. 1992. Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials* 13:190-212.

Trochim, W.M.K., J. C. Cappelleri, and C. S. Reichardt. 1991. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review* 15:571-604.

Charles S. Reichardt is a professor of psychology at the University of Denver. His research focuses on the logic and practice of causal inference. He is the editor of Qualitative and Quantitative Methods in Evaluation Research *(with Tom Cook) and of* Evaluation Studies Review Annual, *Vol. 12 (with Will Shadish). Currently, he is working (with Nick Braucht and Mick Kirby) on a 3-year study of homeless individuals with alcohol or other substance abuse problems.*

William M. K. Trochim is a professor in program evaluation and planning in the Department of Human Service Studies at Cornell University. He has written widely on quasi-experimental design and analysis and is the author of the only book-length treatment of the regression-discontinuity quasi-experimental design. He has also conducted research on multivariate techniques for conceptualization and pattern matching in research, and on the use of microsimulation for studying experimental and quasi-experimental designs.

Joseph C. Cappelleri is a member of the Division of Clinical Care Research at the New England Medical Center. He is primarily involved in meta-analysis research. He is an investigator in the Real-Time Meta-Analysis System and the AIDS clinical trials meta-analysis database project. In addition to developing new meta-analytic protocols, he contributes to statistical, methodological, and clinical advances in meta-analysis. His other work includes helping to pioneer the development of cutoff-based randomized clinical trials and writing about the Second National Incidence Study of Child Abuse and Neglect.