A recently published Evaluation Review *article (April 1990) claimed that because of random measurement error in the pretest (and the regression toward the mean that results) the estimate of the treatment effect of the regression-discontinuity (RD) design is biased. A conceptual approach and a set of computer simulations are presented to arrive at the opposite conclusion: random measurement error in the pretest does* not *bias the estimate of the treatment effect in the RD design. This article, the first of two dealing with measurement error in the RD design, concentrates specifically on the case of no interaction between pretest and treatment on posttest. The claim that the RD effect estimate is not biased due to measurement error is in full agreement with the conclusion reached by several authors who have examined the design over the last two decades.*

RANDOM MEASUREMENT ERROR DOES NOT BIAS THE TREATMENT EFFECT ESTIMATE IN THE REGRESSION-DISCONTINUITY DESIGN

I. The Case of No Interaction

JOSEPH C. CAPPELLERI
WILLIAM M. K. TROCHIM
Cornell University
T. D. STANLEY
Hendrix College
CHARLES S. REICHARDT
University of Denver

> There is no more rational procedure than the method of trial and error—of conjecture and refutation: of boldly proposing theories; of trying our best to show that these are erroneous; and of accepting them tentatively if our critical efforts are unsuccessful. (Popper 1963, 51)

> From the amoeba to Einstein, the growth of knowledge is always the same: we try to solve our problems and to obtain, by a process of elimination, something approaching adequacy in our tentative solution. (Popper 1972, 281)

EVALUATION REVIEW, Vol. 15 No. 4, August 1991 395-419

The regression-discontinuity (RD) design fills an important niche within the constellation of research designs used in program evaluation and applied social research (Trochim 1984). The RD design is distinguished from other pretest-posttest, treatment-control group designs by its unique method of assignment to treatment—persons are assigned to receive treatment solely on the basis of a cutoff value on the pretreatment (pretest) measure, with all persons scoring on one side of the cutoff assigned to one group and all scoring on the other side assigned to the other. Although the design is largely underused, there are several notable implementations that deserve recognition.

Seaver and Quarton (1976) used the RD design to examine how college students' grades in one quarter (the posttest measure) were affected by making the dean's list on the basis of their grades from the previous quarter (the pretest measure). All students in the study who had a first-term grade point average of 3.5 or above were placed on the dean's list and considered to be the experimental group, and those below 3.5 were considered the control group. The RD design was used extensively in the mid and late 1970s under the name "Model C" in the evaluation of compensatory education programs mandated by Title I of the Elementary and Secondary Education Act of 1965 (Tallmadge and Wood 1978; Trochim 1982). Since then, applications of the design outside of compensatory education have begun to surface. Lipsey, Cordray, and Berger (1981) applied the RD design as one component in an evaluation of a juvenile diversion program. Cutoff points on a disposition assignment continuum (DAC), a composite of 11 factors that played a major role in police officer's decisions regarding the disposition of individual cases, determined each case assignment into either the counsel-and-release program (lower values of DAC), the diversion program (near the middle of DAC), or the probation program (higher values of DAC). Recidivism percentages were the values of the outcome measure. Berk and Rauma (1983) conducted a large-scale criminal justice evaluation using the RD design to study whether ex-offenders who received unemployment benefits (experimental condition) have lower reincarceration rates than do ex-offenders who did not receive such benefits (the control condition). The number of recorded hours of prison work was the sole factor in determining whether or not unemployment benefits would be received, with the cutoff of program eligibility set at 652 hours.

Visser and de Leeuw (1984) used the RD design to evaluate a project designed to educate employees about their life-styles in relation to risk of heart disease. Their example used a prespecified value of the amount of serum cholesterol in the blood to partition subjects who were given advice about their life-styles, especially their eating and smoking habits, from subjects who were not. After a few years, serum cholesterol posttreatment values were recorded to see if the intervention was effective. A RAND study by Carter, Winkler, and Biddle (1987), prepared for the National Institutes of Health (NIH), used the RD method to evaluate the research productivity of the Research Career Development Award (RCDA) program of NIH. Priority scores determined whether or not an applicant would be funded as a participant in the program. Havassey et al. (1989) are involved with an in-progress investigation of the relative effectiveness of inpatient versus outpatient treatment for persons dependent on cocaine. They are using two cutoff points on a composite pretreatment assignment measure of severity of cocaine addiction to define a cutoff interval, where all persons scoring below the lower cutoff (less addicted patients) are automatically assigned to outpatient status, those scoring higher than the upper cutoff (more addicted patients) are assigned to inpatient status (the more intensive therapy), and those scoring between the cutoffs are randomly assigned to the two conditions. Robinson and Stanley (1989) and Robinson, Bradley, and Stanley (1990) have implemented an RD design to the evaluation of an accelerated mathematics program for gifted children at different grade levels. Achievement tests were used as preprogram (pretest) and postprogram (posttest) measures.

Although the bias of ordinary least squares (OLS) estimators in the presence of fallibly measured independent variables has been adequately documented for observational studies (Lord 1960, 1967; Cochran 1968, 1970; Griliches 1974, 1986; Fuller 1987; Chatterjee and Hadi 1988), some confusion apparently exists about the consequences of measurement error in the regression-discontinuity design. Stanley and Robinson (1990a) limited their discussion to an analysis of covariance (ANCOVA) model with a single covariate in arguing that random measurement error on the pretest assignment variable in the RD design results in a biased estimate of the treatment effect. If their conclusion were true, the RD design would be compromised and the methodological literature that states otherwise would be wrong. In this article, we use the term *treatment effect* to refer to the effect of the treatment at any given value along the pretest. And because we are assuming no interaction effect, the treatment effect is an additive constant across the pretest.

Contrary to the claim by Stanley and Robinson (1990a), the RD design does not yield a biased treatment effect estimate due to error in measurement. Subsequently, Cappelleri (1990a) convinced Stanley and Robinson of the error of their previously published statements about RD's bias, and they have since recanted in Stanley and Robinson (1990b). The discussion presented

here will use the term *measurement error* to mean *random* measurement error—specifically, random measurement error in the continuous pretest measure. There are different types of measurement error models (Cochran 1968; Fuller 1987). Just like Stanley and Robinson (1990a), we are concerned only with the simplest, although not unrealistic, variation. This type is based on the classical true score model (Carmines and Zeller 1979) and takes the form

$$X_i = T_i + \mu_i \qquad i = 1, 2, \ldots, n \qquad [1]$$

where X_i represents the observed (fallibly measured) score for the ith observation, T_i represents the true (perfectly measured) score for the ith observation, u_i represents the (random) measurement error in X_i, and n represents the total sample size.

Rubin (1977) and Goldberger (1972) provide rigorous statistical proofs that show why measurement error in the pretest does not bias the treatment (main) effect estimate in the RD design. Here we provide an intuitive explanation and Monte Carlo simulations to that end. We limit our focus to the ANCOVA model for ease of exposition, but the same conclusions apply to a general multiple linear regression model. A follow-up article by Trochim, Cappelleri, and Reichardt (forthcoming) deals specifically with showing that the treatment effect estimate remains unbiased even when an interaction effect exists. We assume throughout that the fitted ANCOVA model correctly specifies the true functional form between pretest (X) and posttest (Y), and that the relevant assumptions of multiple regression analysis are not seriously violated. It is also assumed that the binary treatment variable contains no misclassification error: All subjects are correctly classified into the treatment group in which they belong.

First, a brief review of the RD design is presented. Then, an intuitive rationale and computer simulations are presented to explain why the treatment effect estimate is unbiased in the RD design even when the pretest assignment measure contains measurement error.

THE REGRESSION-DISCONTINUITY DESIGN

The RD design is a pre-post, treatment-control strategy that is characterized by its method of assignment to treatment—persons are assigned to either the treatment or control group solely on the basis of a cutoff score on the pretreatment measure. In notational form, the simplest RD design can be depicted as

C O X O

C O O

where the C indicates that groups are assigned by a cutoff score on the pretest, an O stands for the administration of a measure to a group (either before or after treatment), an X depicts the implementation of a treatment, and each group is described on a single line (i.e., program group on top, control group on the bottom).

A hypothetical study of the effect of a new medical treatment for inpatients with a particular diagnosis can be used to illustrate how the RD design works. (Throughout this article *pretest* and *pretreatment* are taken as synonymous terms, as are *posttest* and *posttreatment*; moreover, we interchangeably use the terms *program* and *treatment* in referring to an intervention.) For ethical reasons the new treatment is given to the patients who are most ill. For each patient there is a continuous quantitative indicator of severity of illness that is a composite rating that can take values from 1 to 100, where higher scores indicate greater illness. A pretreatment cutoff score of 50 is (more or less arbitrarily) chosen as the assignment criterion so that all those scoring 50 or higher on the pretreatment indicator are to be given the new treatment, and those with scores lower than 50 are given the standard or control treatment. Figures 1 through 3 depict three possible results using simulated data.

If the treatment is not given (the null case) and patients in both groups are simply measured at two different times on the same measure, the bivariate pre-post distribution for the simulated data may be fit by the regression line shown in Figure 1. Each dot on the figure indicates a single person's pretest and posttest score. The dot labeled "a" shows an individual who had a high pretreatment and posttreatment score. This person was severely ill on the first measure and remained so on the second. The dot labeled "b" shows the pretest and posttest for an individual who was not severely ill on both occasions. The vertical line at the pretreatment score of 50 indicates the cutoff point (although for Figure 1 it is assumed that no treatment is given). Because the solid line through the bivariate distribution depicts the linear regression line, the distribution depicts a strong positive relationship between the pretest and posttest—in general, the more severely ill a person is at one point in time, the more ill he or she is at another time.

If the treatment is administered and has a positive effect, the result might look like Figure 2, where it is assumed that the treatment had a constant and beneficial effect that lowered each treated person's severity of illness by 5 points. Figure 2 is identical to Figure 1 except that all points to the right of the cutoff (i.e., the new treatment group) have been lowered by 5 points on the posttreatment measure, the assumed beneficial treatment effect. The dashed line in Figure 2 shows what one would expect the treated group's regression line to look like if the program had no effect (as was the case in Figure 1).

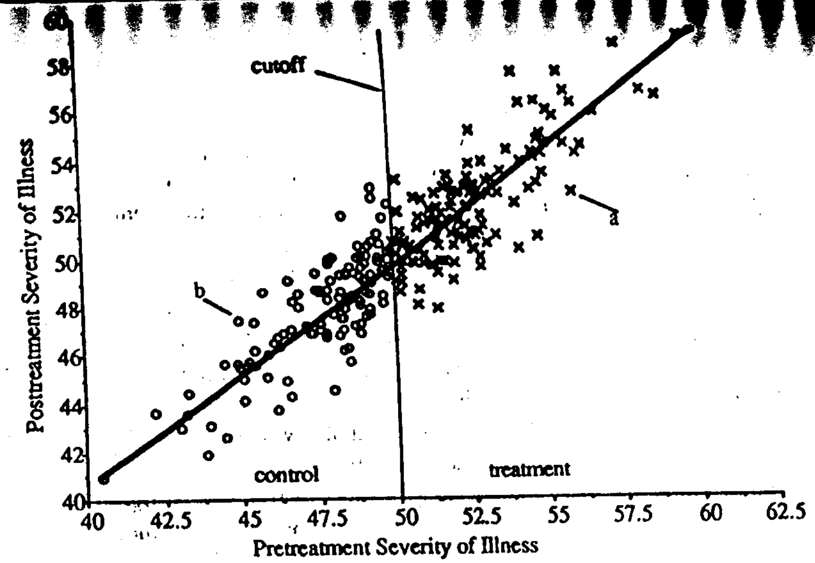

Figure 1: Hypothetical Regression Line for an RD Design With No Treatment Effect

Figure 2 shows how the RD design got its name—a treatment effect is implied when we observe a discontinuity in the regression lines at the cutoff point. This figure portrays a very simple version of the design with a uniform outcome. But it is possible to have a pretest-treatment interaction effect in addition to an additive treatment effect. Figure 3 is identical to Figure 2 except that a –0.5 interaction effect is added to a –5 treatment effect. Besides inducing a 5-point reduction in the posttreatment scores at the cutoff value of 50, the treatment further reduces posttreatment scores for treated patients with higher pretreatment scores. The "sicker" a patient is, the more effective the treatment, Again, the dashed line shows the expected regression line for the no-effect or null case. As in all RD designs, it is the discontinuity in the regression lines at the cutoff point that implies that the treatment has an additive treatment effect.

This introduction to the RD design describes only the simplest variation. More complex variations as well as important implementation and analysis

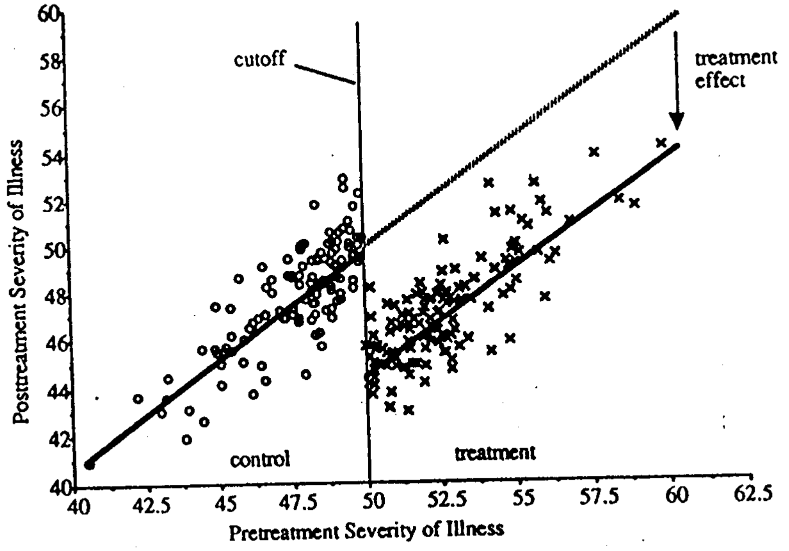

Figure 2: Hypothetical Regression Line for an RD Design With a Treatment Effect of –5 Points

issues are discussed in Sween (1971), Rubin (1977), Cook and Campbell (1979), Judd and Kenny (1981), Berk and Rauma (1983), Trochim (1984, 1990a, 1990b), and Mohr (1988), among other places.

The standard statistical model for the basic RD design is discussed in Trochim (1984, 1990a) and is similar to the approach recommended in Judd and Kenny (1981). Given a pretest assignment measure, X_i, and a postprogram measure, Y_i, the ANCOVA model assumes only a constant additive effect term. For an RD design with a linear X-Y relationship, this model can be stated as follows:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_X\tilde{X}_i + \hat{\beta}_Z Z_i + e_i \,, \qquad [2]$$

where

$\tilde{X}_i$ = preprogram measure for individual i minus the cutoff value, X_c (i.e., $\tilde{X}_i = X_i - X_c$)
Y_i = postprogram measure for individual i
Z_i = binary group variable for individual i (1 if program participant; 0 if comparison participant)

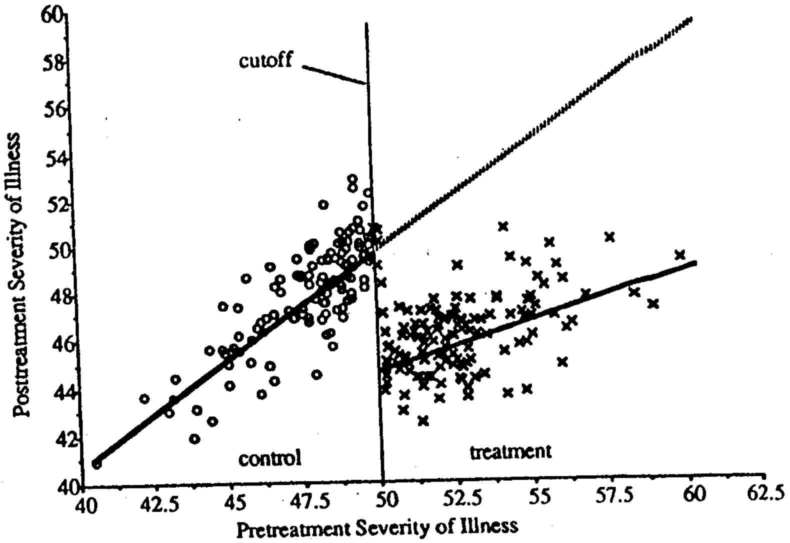

Figure 3: Hypothetical Regression Line for an RD Design With a Treatment Effect of –5 Points Plus an Interaction Effect of –.5 Points

$\hat{\beta}_0$ = estimate of comparison group intercept at cutoff
$\hat{\beta}_X$ = linear slope estimate
$\hat{\beta}_Z$ = program or treatment effect estimate
e_i = regression disturbance term for individual i.

The null hypothesis of interest tests the true program effect parameter (β_Z)

$$H_0\colon \beta_Z = 0$$

against the alternative hypothesis

$$H_1\colon \beta_Z \neq 0.$$

Traditionally, the model estimates the program (treatment) effect at the cutoff point (Trochim 1984). To accomplish this, the analyst subtracts the cutoff score from each pretest score. The term $\tilde{X}_i$ has a tilde (~) over it to indicate this transformation on the pretest X_i. Of course, with the above model, it does not matter where the program effect is estimated along the pretest range, because in the ANCOVA model the two regression lines are parallel. The relevant hypothesis is then a simple *t* test of the regression coefficient that estimates the treatment effect.

THE RD DESIGN IN RELATION TO OTHER DESIGNS

It is useful to contrast the RD design with other similar designs. The RD design falls into a broader class of pre-post, two-group designs that share the notation

$$\begin{matrix} O & X & O \\ O & & O \end{matrix}$$

The feature that distinguishes the different types of pre-post, two-group designs is the manner in which persons or units are assigned to treatment groups. The pre-post randomized experimental (RE) design uses random assignment to create the two groups. In the nonequivalent group design (NEGD), persons (or units) usually self-select into the treatment groups or, less frequently, are assigned in some unknown, uncontrolled way. This design is referred to in some quarters as an observational design when assignment is based on self-selection. In the NEGD, although one might hope that the two groups are equivalent prior to treatment, because treatment assignment is not controlled, there is no way to know whether or not this is the case. Thus the term *nonequivalent* is taken here as a reminder that group assignment is determined by the persons themselves (or in some other uncontrolled manner) and is therefore not known perfectly.

There is confusion in the literature regarding the definition of a "true" experiment. Cook and Campbell (1979) argued that a true experiment always uses random assignment to create comparisons from which treatment caused change is inferred. By their definition, the RD design is a quasi-experiment because it lacks the essential quality of random assignment. As such, it would be logically grouped with other quasi-experimental designs like the NEGD. In contrast, Mosteller (1990) defined a true experiment as a design in which the assignment to treatment is *controlled* by the investigator. Thus he claimed that the RD design should be considered an experiment: "By this author's definition—which is that in an experiment the investigator controls the application of treatments—the regression-discontinuity design is actually an experiment" (225). The distinction is more than just semantics. If the RD design is classified as a quasi-experiment, the implication is that it is subject to the same types of weaknesses common to other quasi-experiments (e.g., observational and NEGD). This article argues that it makes sense to classify the RD design as an experiment because it operates just like an RE does, *at least with respect to measurement error*. When implemented correctly, both the RE and RD designs yield unbiased estimates of the treatment effect even when measurement error is present. An intuitive explanation and a set of computer simulations are offered to help explain why the RD design functions like the RE design with respect to measurement error.

INTUITIVE EXPLANATION: TREATMENT ASSIGNMENT IS PERFECTLY KNOWN

An intuitive explanation is offered for why the treatment effect estimate remains unbiased in the simplest RD design, even though pretest scores have measurement error. This explanation is not limited to the ANCOVA model; it holds for the general linear model. Furthermore, the rationale provides a distribution-free argument; that is, it is "free" or independent of the type of pretest distribution, be it normal or otherwise.

The RD design yields an unbiased estimate of treatment effect because it incorporates a perfectly known assignment rule that is fully modeled and accounted for in the ANCOVA analysis. In the basic RD design, subjects who fall below a preestablished cutoff point on some fallible pretreatment indicator are placed in one group, and those who fall above this point are placed in the other group. If the treatment goes to subjects most in need of it, as evidenced by (say) higher pretest scores, the probability of being assigned to the experimental condition is 1 for subjects whose scores fall above the cutoff score, and the corresponding probability is 0 for subjects who fall below the cutoff score. Thus assignment occurs through an observed treatment assignment (i.e., pretest) variable that renders a completely known decision rule. It is this assignment variable, X—not some unobserved true score or true ability value—that completely and perfectly determines group assignment. Because assignment to experimental and control groups is entirely based on a known set of observed scores, the treatment effect is isolated by conditioning on or controlling for the observed pretest assignment variable in the ANCOVA analysis—resulting in an unbiased OLS estimate of treatment effect (Reichardt 1979).

Similarly, the RE design also incorporates complete knowledge of the assignment process. Here, however, each subject regardless of his or her pretest score has a fixed probability of assignment to the experimental group. Random assignment across the pretest range renders a known probability of assignment to either treatment condition, which probabilistically makes the dichotomous treatment group variable and pretest covariate uncorrelated. It is this expected lack of association between treatment and covariate, triggered by a known assignment rule, that keeps the treatment effect unbiased in the RE design when measurement error exists in the covariate. However, in NEGD (or observational) studies, the underlying determinants of selection are not known and hence not measured and controlled for in the analysis (Reichardt 1979). Generally, for NEGD studies, a correction needs to be made (like the useful correction formulas provided by Stanley and Robinson 1990a) in the ANCOVA model that adjusts the treatment effect estimate to account for the unreliability of pretest measurements and for the expected correlation between the fallibly measured pretest and the treatment group variable.

Consider a hypothetical study that relates an individual's posttreatment measure of systolic blood pressure (the dependent variable) to his or her pretreatment measure of systolic blood pressure (the assignment covariate) and whether or not the person exercises regularly (the binary treatment group variable). In an NEGD or an observational study, a person's decision to exercise regularly is determined by his or her true, unobserved level of motivation to exercise, not by motivation as measured by the pretreatment measure of systolic blood pressure or anything else. Because systolic blood pressure as a measure of motivation to exercise is a fallible indicator of the true level of motivation to exercise, and because it is this true (unknown) level that underlies the selection process to exercise, the inaccuracies of measurement found in the pretreatment measure of systolic blood pressure and its expected correlation to the binary exercise variable must be taken into account if the coefficient for the exercise variable is to be unbiased.

In contrast, RD design determines selection not by some unknown construct but by the pretest *as measured*. This design allocates individuals to either a regular exercise program or no regular exercise program solely on the basis of their systolic blood pressure readings. Because systolic blood pressure is the true cutoff variable, its values really have no measurement error for the purpose of creating and analyzing the RD design. The only way to have error in this cutoff variable is to have an assignment strategy that plans to assign people to treatment groups on the basis of their systolic blood pressure readings, *then* add measurement error to it by misclassifying some of them to the other condition, and then use this fallible measure as the independent variable in an analysis of data from the design (i.e., the fuzzy RD design; see Trochim 1984). Stanley and Robinson's (1990a) adjusted formula for the OLS estimator of treatment effect, as well as for the OLS estimator of the slope coefficient, is correct for NEGD or observational studies. These formulas are useful to know. But if such an adjustment is used in the RD design, as they recommend, it will actually induce a bias into the treatment effect when in fact there was initially no such bias.

Using this argument that X and nothing else completely determines Z, we can see algebraically why the OLS treatment effect estimator is not biased. The "corrected" estimators for the RD design provided by Stanley and Robinson (1990a) originated from the derivations and proof of the econometricians Griliches and Ringstad (1971, Appendix C), who examined the case when one independent variable, but not the other, is measured with random error. Griliches and Ringstad's set of adjusted estimators, those given by Stanley and Robinson (1990a), are preceded by and depend on an intermediate result that eventually goes into determining the corrected estimators. This intermediate formula is

$$E(\hat{\beta}_Z) = \beta_Z - \beta_X(\beta_{u,Z|X}), \qquad [3]$$

where β_Z is the true treatment effect parameter, β_X is the true slope parameter and $\beta_{u,Z|X}$ is the population partial regression coefficient obtained by regressing measurement error, u, on the perfectly measured treatment variable, Z, controlling for or holding constant the fallibly measured covariate, X.

If the partial population correlation coefficient between u and Z (controlling for X), denoted by $R_{uZ.X}$, is zero, then this implies that $\beta_{u,Z|X}$ is zero because of the direct link between correlation and regression. Examining the existence of a linear relationship between u and Z, controlling for X, is the same as examining whether there is a significant partial correlation between them.

In RD, Z is determined exactly by the random variable X, so Z would depend on the random variables T and u as X = T + u, where T and u are assumed to be independent random variables in the classical test theory model. But once X is fixed at some value (i.e., once X is controlled), Z becomes completely determined and fixed, hence independent of anything else, including u as well as T.

This means that Z is correlated with u, but its *partial* correlation with u is zero when X is controlled. Hence the partial correlation between u and Z, $R_{uZ.X}$, equals zero, which implies that the partial regression coefficient from the regression of u on Z, $\beta_{u,Z|X}$, is also zero in the RD design. Therefore $E(\hat{\beta}_Z) = \beta_Z - \beta_X(0) = \beta_Z$, confirming algebraically that the treatment estimate in the RD design is unbiased when the observed covariate is measured with error.

Once treatment assignment is perfectly known, the coefficient of the pretest covariate completely absorbs the bias due to measurement error in the pretest. It can be shown (e.g., in Gujarati 1988, 417) that random measurement error in a covariate causes that covariate (X) and the stochastic regression disturbance term to be correlated. This violates one of the crucial assumptions in the classical linear regression model that no such correlation is present. When this assumption is violated, the OLS estimator of X is both biased and inconsistent (i.e., it remains biased even if the sample size increases indefinitely).

Yet in the RD design once measurement error becomes identified and controlled for by X or, more specifically, by the (slope) coefficient of X, measurement error then becomes unconfounded with the coefficient of the dichotomous treatment variable (Z) because this variable is merely two subdivisions of an already controlled, fallibly measured covariate. In the randomized experimental design the same bias attributed to measurement error is fully "absorbed" or "captured" (in a sense) by the coefficient of X, but this bias does not contaminate the treatment effect estimate because Z is probabilistically or theoretically uncorrelated with X. In both RD and RE designs, the slope coefficient of the pretest is attenuated by an amount equal to the reliability coefficient. That is, if β_X is the true population slope parameter and $\hat{\beta}_X$ is the sample slope estimate, then $\hat{\beta}_X = \beta_X(\rho)$ where ρ is the reliability coefficient. Therefore, dividing $\hat{\beta}_X$ by ρ will give an unbiased estimate of the slope coefficient parameter.

An NEGD (observational) study, on the other hand, typically contains both a significant correlation between Z and X *and* lacks knowledge of the assignment rule that completely determines Z from X. This not only makes the slope coefficient biased but also makes the treatment effect coefficient biased. In essence, the correlation between pretest and the stochastic disturbance term "transmits" a bias to the treatment effect coefficient.

The above rationale can be shown pictorially, without loss of generality, for the case of no real treatment effect. Also assume, to simplify matters, that no extraneous factors (like measurement error in the posttest) can account for the differences between a given person's pre and post scores. Figure 4 shows what happens for NEGD and observational designs. In Figure 4, the observations are uniformly scattered in one ellipse for the treatment group and in the other ellipse for the control group. Line AB in Figure 4 shows the true null regression line for NEGD and observational designs when there is no measurement error in the pretest. When measurement error in the pretest is introduced (along with the regression toward the mean that results), however, an apparent treatment effect arises as evidenced by the fact that the fitted regression lines CD and EF do not intersect.

Figure 5 shows what happens for RE and RD designs. First, consider the RE design. In Figure 5, whose observations are uniformly scattered through the ellipse, line AB depicts the true regression line in the absence of measurement error. Although an attenuated regression slope inevitably surfaces in the RE design with measurement error—as shown by line CD—it is

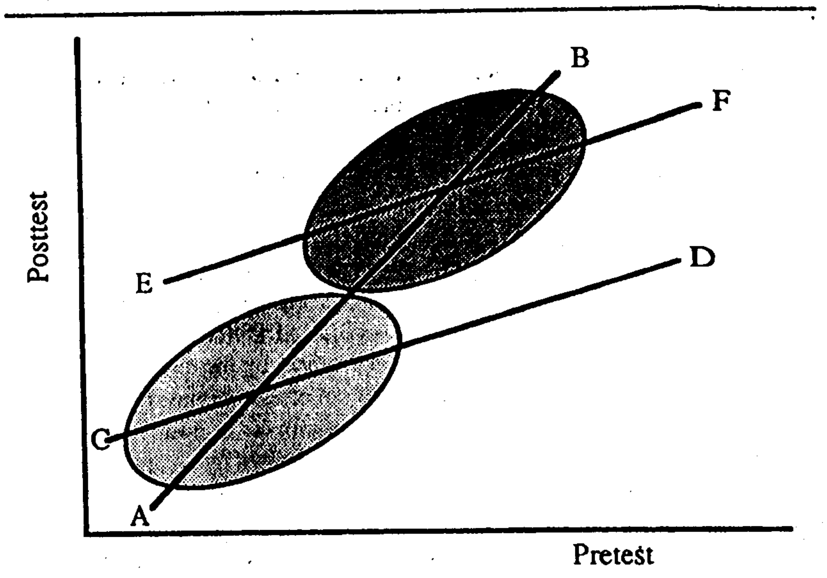

Figure 4: Effects of Measurement Error In NEGD and Observational Designs (null case)

generally known that measurement error will not bias the treatment effect estimate in the RE design, because randomization spreads the measurement error evenly over both groups, making the two groups share the same pretest population mean (Cochran 1968).

Consider next the RD design, also shown in Figure 5. It is in principle identical to the RE design except that half of the data in each group is missing (i.e., half of the data above and below the cutoff score), thereby making the RD design reach the same conclusion as the RE design. In the null case measurement, error in both RD and RE designs is adequately captured by the single continuous regression line itself, implying no adverse effect on the null treatment effect estimate (Trochim 1990a).

COMPUTER SIMULATIONS

CONSTRUCTING FOUR ANCOVA DESIGNS

Monte Carlo computer simulations are conducted with 1,000 repetitions of 100 cases per repetition for four different ANCOVA design strategies: the

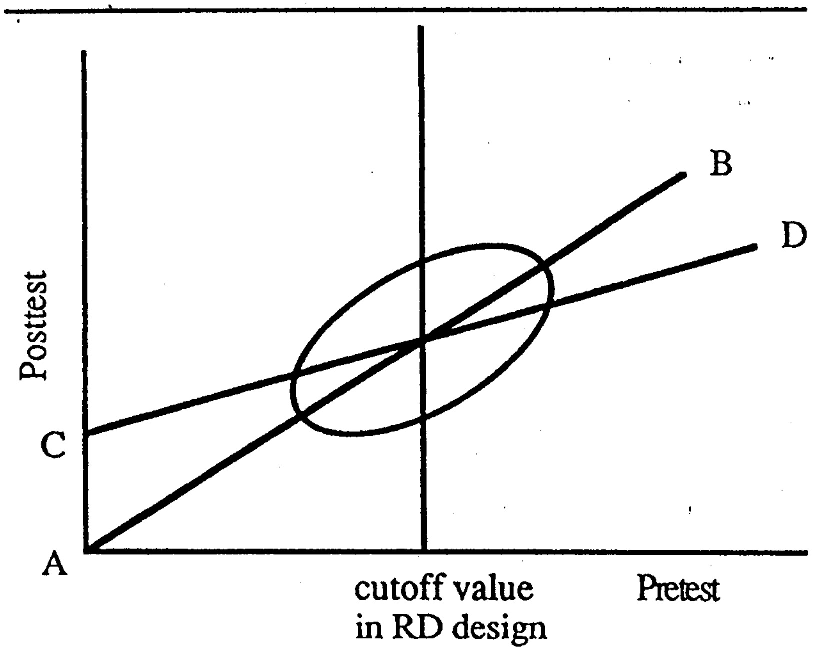

Figure 5: Effects of Measurement in RE and RD Designs (null case)

RD design with assignment into treatment groups based on true scores, the RD design with assignment based on observed scores, the RE design, and the NEGD. All four designs share about an equal number of cases in both groups. This set of 1,000 repetitions is more extensive (and therefore more closely mimicks the expected results) than does the work of Trochim (1984) and Trochim and Davis (1986a, 1986b), who have investigated a similar set of simulation models but with only 20 repetitions and 50 repetitions. Furthermore, unlike these previous simulation studies, this current simulation study demonstrates the bias associated with the intercept and pretest regression coefficients in RD and RE models and presents a couple of formulas to correct for the bias.

In what follows, the notation "~ N (mean, variance)" refers to a normally distributed random variable with a known mean and variance. Moreover, V denotes the variance operator and E denotes the expectation operator.

The setup of the simulation study takes the following arrangement:

- number of cases = 100
- number of runs = 1,000
- true pretest score, $T \sim N(50, 25)$

- random measurement error, $u \sim N(0, 9)$
- observed pretest score, $X = T + u$
- known reliability coefficient, $\rho = \frac{Var(T)}{Var(X)} = \frac{Var(T)}{Var(T) + Var(u)} = \frac{25}{25 + 9} = .735$
- true treatment effect = –8 points
- binary treatment variable, Z; Z = 1 if in treatment group, Z = 0 if in control group
- regression disturbance term, $e \sim N(0, 4)$
- simulated model for posttest is $Y = T + -8(Z) + e$.

A description of the simulation specifications for group assignment for the four design strategies follows.

Design RD—no error refers to the basic RD design that incorporates the unrealistic selection procedure that allocates subjects to the two treatment conditions on the basis of their true scores rather than their observed scores. In practice, of course, this design is impossible because we never know the true scores. All subjects with true scores at or above 50 (the average value of true scores) are placed into the treated group (Z = 1), and all subjects with true scores below 50 are placed into the control group (Z = 0). Although not a practical illustration, because measurement error is prevalent in all continuous variables, this method of assignment is included here to simulate the true situation in which there is no measurement error in the pretest (in essence, X = T) and to serve as a baseline for comparison with the other three designs, which contain measurement error in the pretest measure.

Design RD refers, as usual, to the basic RD design that assigns subjects into treatment groups strictly on the basis of their fallibly measured observed scores. The cutoff value of 50 is chosen on the observed measure: Everyone at or above X = 50 (the mean of X) is assigned a Z value of 1; everyone below 50 is assigned a Z value of 0. This design is the main design of interest and the one that Stanley and Robinson (1990a) contended should be biased because of measurement error.

Design RE refers to the RE design that assigns subjects randomly across the entire pretest range. The probability of assignment to either group is .50, accomplished by generating an independently and identically distributed standard normal variable. Cases with values at or above zero on the standard normal variable are automatically placed into the treatment group; otherwise, cases are placed into the control group.

Designs RD—no error, RD, and RE have one very important feature in common: They are modeled, by the nature of their designs, with a perfectly known treatment assignment function.

Design NEGD, on the other hand, represents the nonequivalent group design, whose unknown selection process cannot be perfectly modeled into the analysis. An NEGD is created in the simulation by assigning cases with *true scores* at or above 50 to treatment and below 50 to control, and by using the *observed* covariate (X) instead of the true covariate (T) as the pretest regressor. For the purpose of simulating this particular design, the data analyst knows only Y, X, and Z. He or she is unaware that a subject's (unknown) true score solely determines the selection process. This assignment strategy can also be viewed as an RD design with assignment based on true scores but with its regression analysis based on observed scores, known as the fuzzy RD design (Trochim 1984). Conclusions drawn from this special type of NEGD can be generalized to NEGDs broadly. Trochim and Davis (1986a, 1986b) provide a more general way to simulate NEGDs.

All four designs are analyzed in the regression model via the estimated ANCOVA regression function

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_X X + \hat{\beta}_Z Z \qquad [4]$$

where $\hat{Y}$ is the predicted value of Y from the OLS regression of Y on X and Z, and the coefficient estimates (the $\hat{\beta}$s) are not necessarily the same across designs. Although all four analysis use this same analytic model, they differ in their assignment strategies, as discussed earlier.

SIMULATION RESULTS

By "bias" in the treatment effect estimate we mean it in the usual sense, that is, $E(\hat{\beta}_Z) \neq \beta_Z$. In Monte Carlo simulations, it is common to consider an estimate unbiased $[E(\hat{\beta}_Z) = \beta_Z]$ if the average estimate (in our case averaged over 1,000 samples, each containing a different set of 100 observations) lies within 1.96 standard errors of the true population value it is attempting to estimate. The sample value of interest is the average estimate, and we wish to investigate how it deviates from the fixed parameter it is attempting to estimate. The standard deviation of the average estimate is, of course, the standard error of the mean, which in our simulation is the sample standard deviation of 1,000 individual beta coefficient estimates divided by the square root of 1,000. Here, if the actual parameter value falls within a 95% confidence interval of the average estimate plus or minus 1.96 times the standard error of the mean, then the average estimate is declared unbiased.

Table 1 shows the simulation results for the four ANCOVA designs. For this table, we denote the standard error (SE) of the mean intercept estimate $\hat{\beta}_0$ as SE($\hat{\beta}_0$), of the mean slope estimate $\hat{\beta}_X$ as SE($\hat{\beta}_X$), and of the mean treatment estimate $\hat{\beta}_Z$ as SE($\hat{\beta}_Z$). The results from the tables illustrate at least five important methodological principles. First, designs RD—no error, RD, and RE—designs that model a perfectly known assignment rule—all give an unbiased estimate of the true population treatment effect of –8 points. In particular, design RD, the main design of interest, clearly yields an unbiased (average) $\hat{\beta}_Z$ value of –8.005 as evidenced by mere inspection or by the true value of –8 falling within the 95% confidence interval –8.005 ± 1.96(.034) = (–7.938, –8.071). Using any correction factor, therefore, would introduce a bias.

TABLE 1: Simulation Results for All Four Designs (average results across 1,000 repetitions, 100 cases per run)

Mean Estimates and Standard Errors of Mean Estimates	Design RD—No Error	Design RD	Design RE	Design NEGD
$\hat{\beta}_0$	.025	13.161	13.182	22.998
SE($\hat{\beta}_0$)	.098	.136	.090	.102
$\hat{\beta}_X$	.999	.736	.736	.500
SE($\hat{\beta}_X$)	.002	.002	.001	.002
$\hat{\beta}_Z$	–7.983	–8.005	–8.000	–4.004
SE($\hat{\beta}_Z$)	.021	.034	.020	.025

Second, design RD—no error, which deals with the impractical event of no measurement error, in addition elicits unbiased OLS estimates for the intercept term [$\hat{\beta}_0$ = .025, SE($\hat{\beta}_0$) = .098] and the slope coefficient [$\hat{\beta}_X$ = .999, SE($\hat{\beta}_X$) = .002]. With the simulated posttest model being Y = T + –8(Z) + e, the population intercept term is 0 and the population coefficient for the pretest is 1. Thus, when there is no pretest measurement error, all OLS estimates are, as expected, unbiased.

Third, sample intercept terms and slope coefficients for the other three designs are in fact biased. However, we know from Goldberger (1972) and Cappelleri (1990a) how to correct these biased OLS estimators in randomized and RD designs. Thus, when pretest measurement error exists, intercept and slope coefficients in RD and RE designs are, as expected, biased, although the treatment effect estimates are not.

Let the superscript "c" on an estimator represent an estimator that has been corrected for bias attributed to measurement error. For designs RD and RE the corrected or adjusted intercept estimator is

$$\hat{\beta}_0^c = \hat{\beta}_0 - \mu_X(1 - \rho) \qquad [5]$$

in which $\mu_X = \mu_T$ represents the population mean of X or, equivalently, of T, under the assumption that X and u are normally distributed. For design RE

$$\hat{\beta}_0^c = 13.182 - 50(1 - .736) = -.018$$

and for design RD

$$\hat{\beta}_0^c = 13.161 - 50(1 - .736) = -.039,$$

which are not biased when corrected in this manner.

With both μ_X and ρ known constants, the corrected or adjusted standard error for $\hat{\beta}_0$ is the same as the standard error for $\hat{\beta}_0$ of .136 given for design RD and of .090 given for design RE. Therefore, the sample intercept terms in designs RD and RE are now unbiased as their respective confidence intervals include 0.

Designs RD and RE should have an estimated slope coefficient near 1 in the absence of measurement error. In both these designs the estimated slope coefficient is .736, differing from 1 by more than chance. The reliability coefficient of .736 accounts fully for the slope attenuation. For designs RD and RE the corrected or adjusted slope estimator is simply

$$\hat{\beta}_X^c = \frac{\hat{\beta}_X}{\rho}. \qquad [6]$$

Both designs RD and RE have a corrected slope estimate of $\hat{\beta}_X^c = .736/.736 = 1$, which in this simulation is the value it should equal in the absence of slope attenuation. Because this revised estimate exactly equals its corresponding parameter value, we now clearly see no bias in the pretest coefficient. With ρ taken as a known population constant, the corrected or adjusted standard error for $\hat{\beta}_X$ for both designs is

$$\begin{aligned} SE(\hat{\beta}_X^c) &= SE\left[\frac{\hat{\beta}_X}{\rho}\right] \\ &= \frac{SE(\hat{\beta}_X)}{\rho}. \qquad [7] \end{aligned}$$

For design RD, $SE(\hat{\beta}_X^c) = .002/.736 = .002$; for design RE, $SE(\hat{\beta}_X^c) = .001/.736 = .001$. Thus slope coefficients are unbiased when corrected.

Notice that the simulated design NEGD clearly yields not only a biased estimate of the intercept term and slope coefficient but also a biased estimate of the treatment effect. Although not done so here, revised, unbiased OLS estimates for design NEGD can be obtained from the adjustment formulas provided by Stanley and Robinson (1990a), which are correct for the NEGD design.

Fourth, although the RE and RD designs give virtually identical OLS estimates, the standard errors of those estimates are lower in design RE than in design RD. For instance, the average standard error of the 1,000 individual treatment estimates in design RD is higher than that in design RE by a factor of about 1.7:

$$\begin{aligned} \frac{SE(\hat{\beta}_Z)_{RD}}{SE(\hat{\beta}_Z)_{RE}} &= \frac{.034(\sqrt{1,000})}{.020(\sqrt{1,000})} \\ &= \frac{1.075}{.632} \\ &= 1.7. \end{aligned}$$

This corroborates the work done by Goldberger (1972) and Cappelleri (1990b) on the relative efficiency and statistical power advantages of RE over RD designs. Thus a drawback of the RD design is that it is less likely to show a significant program effect, when one exists, than is the RE design.

Finally, although the treatment effect point estimate in designs RD—no error, RD, and RE is not susceptible to measurement error, a fallibly measured covariate in the ANCOVA model implies more variability in the treatment effect estimate. This is expected and can be seen in the simulation by noting that the average standard error of the 1,000 treatment effect estimates in design RD (1.075) is noticeably higher than that in design RD—no error ($.664 = .021\sqrt{1,000}$).

A CAVEAT ABOUT THIS SIMULATION STUDY

Caution should be undertaken when doing RD simulation studies of this type. The fitted regression model should take the same functional form as the simulated (generated) theoretical model. Suppose, for instance, that true scores followed a nonnormal distribution (like a uniform distribution) and measurement error in observed scores continued to follow a normal distribution, with a mean of zero and a given variance. Also assume that the simulated model is $Y = T + \beta_Z Z + e$, where β_Z is the program effect size. Measurement error in X, which equals T + u, will then induce a nonlinear trend in the observed X-Y functional form as the addition of a nonnormal T and normal u makes Y and X have a nonlinear relationship (Cochran 1970). Trochim, Cappelleri, and Reichardt (forthcoming) provide a detailed discussion on what happens when the fitted RD model has a different form than the simulated RD model.

Consequently, if an RD design were simulated in this manner, a linear regression of Y on X and Z may result in a bias treatment estimate in the RD design. *But this is not due to measurement error in the pretest per se.* It is due to incorrectly specifying a linear regression function between X and Y when, in fact, a nonlinear relationship exists between them. (See the discussion in Cook and Campbell 1979, 140-41, for further insight.) This is not to say that the pre-post distribution needs to be bivariate normal for the RD design to yield an unbiased treatment estimate, because even if X is normally distributed by virtue of T and u being normally distributed, Y is a nonnormal variable when β_Z does not equal zero.

It is imperative when doing RD simulations of this type to create X by adding a normally distributed u variable to normally distributed T variable, because in this case both the fitted and simulated models assume a linear pretest-posttest functional relationship. As a means toward correctly modeling the X-Y relationship in practice, whether or not the assignment variable or the measurement error variable is normally distributed, Rubin (1977) suggested using strong a priori information about the functional form, Boruch (1978) suggested using a "dry run" approach, Mohr (1988) suggested using double pretests, and Trochim (1984) suggested coupling the RD design with the RE design. Perhaps, if feasible, implementing each of these suggestions for a single study would be the most favorable strategy in carefully specifying the RD model. In practice, an appropriate transformation on the pretest or posttest or both should be considered if a linear fit is deemed inadequate.

Because the RE design propitiously has its treatment group regression lines spread over the entire pretest continuum, the RE design is not as sensitive to model misspecification as its RD counterpart. Although the standard error of its treatment estimate may become inflated, the RE design may still yield an unbiased estimate of treatment effect when a straight line is fit to a nonlinear X-Y relationship.

In the simulation of design RD, the mean of the normally distributed pretest is chosen as the cutoff value, as well as the point at which to estimate the treatment effect. No caveat is warranted here, however. The cutoff value can be chosen at a point other than at the mean of the normally distributed pretest variable for the simulation to shown an unbiased treatment effect. The situation is more complex when an interaction term is included (Trochim, Cappelleri, and Reichardt forthcoming).

CONCLUSION

When classified as an NEGD (Cook and Campbell 1979), the RD design may at first glance be taken as just another NEGD and hence susceptible to measurement error problems that face NEGDs. But NEGDs can be broadly classified by whether or not the selection process is controlled. This distinction is important for determining the influence that random measurement error in the pretest has on the treatment effect estimate in the linear model in general and in the analysis of covariance model in particular. Although the RD design does deliberately create two nonequivalent groups on the pretest measure, it is not a typical NEGD because assignment is perfectly controlled. Because its assignment mechanism to experimental and control groups is entirely based on a *known set of observed scores* and nothing else, its treatment effect estimate is unbiased and isolated from the contamination caused by measurement error in the pretest covariate. And the same conclusions hold for randomized experiments. The unique and beneficial design structure of RD designs is not employed by NEGDs when assignment is not controlled. Their treatment effect estimates, along with their slope and intercept estimates, are prone to bias.

The basic point is that for the process being measured—the assignment process—there is no measurement error of any kind in a properly implemented RD design. The pretest (containing whatever type of measurement error for measuring some underlying construct) is not really fallibly measured for the purpose of creating and analyzing the RD design. The fact that there is random measurement error for the underlying construct, which is contained in the observed pretest, is besides the point for the assignment process. Assignment is literally done by the pretest score, not by the underlying construct (whatever that is). So, in principle, the treatment effect is estimated in an unbiased fashion. Any possible bias that does arise from measurement error is completely absorbed by the regression coefficient of the covariate itself. As Trochim, Cappelleri, and Reichardt (forthcoming) demonstrate, this same line reasoning holds when a pretest-treatment interaction is included.

The premise taken in this article is in full agreement with those authors who have contributed methodological advances to the RD design. These authors include, but are not restricted to, Barnow (1972), Goldberger (1972), Cain (1975), Rubin (1977), Cook and Campbell (1979), Barnow, Cain, and Goldberger (1980), Judd and Kenny (1981), Berk and Rauma (1983), Trochim (1984, 1990a, 1990b), Mohr (1988), and now Stanley and Robinson (1990b).

REFERENCES

Barnow, B. S. 1972. *Conditions for the presence or absence of a bias in treatment effect: Some statistical models for head start evaluation.* Discussion paper #122, Institute for Research on Poverty, University of Wisconsin, Madison.

Barnow, B. S., G. G. Cain, and A. S. Goldberger. 1980. Issues in the analysis of selectivity bias. In *Evaluation studies review annual*, vol. 5, edited by E. W. Stromsdorfer and G. Farkas. Beverly Hills, CA: Sage.

Berk, R. A., and D. Rauma. 1983. Capitalizing on nonrandom assignment to treatment: A RD of a crime control program. *Journal of the American Statistical Association* 78:21-28.

Boruch, R. F. 1978. Double pretests for checking certain threats to the validity of some conventional evaluation designs, or stalking the null hypothesis. Northwestern University. Typescript.

Cain, G. G. 1975. Regression and selection models improve nonexperimental comparisons. In *Evaluation and experiment: Some critical issues in assessing social programs*, edited by C. A. Bennett and A. A. Lumdaine. New York: Academic Press.

Cappelleri, J. C. 1990a. Random measurement error in regression-discontinuity designs. Paper presented at the annual conference of the American Evaluation Association, Washington, DC, October.

———. 1990b. Power analysis of regression-discontinuity designs. Paper presented at the annual conference of the American Evaluation Association, Washington, DC, October.

Carmines, E. G., and R. A. Zeller. 1979. *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Carter, G. M., J. D. Winkler, and A. K. Biddle. 1987. *An evaluation of the NIH research career development award.* Santa Monica, CA: RAND.

Chatterjee, S., and A. Hadi. 1988. *Sensitivity analysis in linear regression.* New York: Wiley.

Cochran, W. G. 1968. Error 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.

Fuller, W. 1987. *Measurement error models.* New York: Wiley.

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

Griliches, Z. 1974. Errors in variables and other unobservables. *Econometrica* 42(6): 971-98.

———. 1986. Economic data issues. In *Handbook of econometrics*, vol. 3, edited by Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Griliches, Z., and V. Ringstad. 1971. *Economies of scale and the form of the production function.* Amsterdam: North-Holland.

Gujarati, D. N. 1988. *Basic econometrics.* New York: McGraw-Hill.

Havassy, B. E., D. R. Wesson, J. M. Tschann, S. M. Hall, and C. J. Henke. 1989. *Efficacy of cocaine treatment: A collaborative study.* Grant proposal submitted to the National Institute on Drug Abuse (NIDA #DA05582).

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

Lipsey, M. W., D. S. Cordray, and D. E. Berger. 1981. Evaluation of a juvenile diversion program. *Evaluation Review* 5:283-306.

Lord, F. M. 1960. Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association* 55:307-21.

———. 1967. A paradox in the interpretation of group comparisons. *Psychological Bulletin* 68:304-5.

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

Mosteller, F. 1990. Improving research methodology: An overview. In *Research methodology: Strengthening causal interpretations of nonexperimental data*, edited by L. Sechrest, E. Perrin, and J. Bunker. Washington, DC: U.S. Public Health Service.

Popper, K. R. 1963. *Conjectures and refutations: The growth of scientific knowledge.* New York: Basic Books.

———. 1972. *Objective knowledge: An evolutionary approach.* Oxford: Oxford University Press.

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

Robinson, A., R. D. Bradley, and T. D. Stanley. 1990. Opportunity to achieve: Identifying and serving mathematically talented black students. *Contemporary Educational Psychology* 15 (1): 1-12.

Robinson, A., and T. D. Stanley. 1989. Teaching to talent: Evaluating an enriched and accelerated mathematics program. *Journal of the Education of the Gifted* 12:253-67.

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

Seaver, W. B., and R. J. Quarton. 1976. Regression-discontinuity analysis of dean's list effects. *Journal of Educational Psychology* 66:459-65.

Stanley, T. D., and A. Robinson. 1990a. Sifting statistical significance from the artifact of RD design. *Evaluation Review* 14:166-81.

———. 1990b. "Second best" evaluation design: Regression discontinuity or abbreviated time series? Paper presented at the Annual Conference of the American Evaluation Association, Washington, DC, October.

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

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.M.K. 1982. Methodologically based discrepancies in compensatory education evaluations. *Evaluation Review* 6:443-80.

———. 1984. *Research design for program evaluation: The regression-discontinuity approach.* Beverly Hills, CA: Sage.

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

———. 1990b. Cutoff assignment strategies for enhancing randomized clinical trials (RCTs). Paper presented at the Annual Conference of the American Evaluation Association, Washington, DC, October.

Trochim, W.M.K., J. C. Cappelleri, and C. S. Reichardt. Forthcoming. Random measurement error doesn't bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review.*

Trochim, W.M.K., and J. E. Davis. 1986a. Computer simulation for program evaluation. *Evaluation Review* 10:609-34.

———. 1986b. Computer simulation of human service program evaluation. *Computers in Human Services* 1 (4): 17-38.

Visser, R. A., and J. de Leeuw. 1984. Maximum likelihood analysis for a generalized regression-discontinuity design. *Journal of Educational Statistics* 9 (1): 45-60.

Joseph C. Cappelleri earned his Ph.D. in research methodology and evaluation from the Department of Human Service Studies at Cornell University and is currently pursuing an M.P.H. degree in quantitative methods at the Harvard School of Public Health. He was recently awarded a cancer epidemiology trainee fellowship on behalf of the National Cancer Institute. His interests in research methodology include cutoff-based experimental designs, and his substantive interests include cancer epidemiology and the epidemiology of child abuse and neglect.

William M. K. Trochim is Associate Professor in Program Evaluation and Planning in the Department of Human Services 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.

T. D. Stanley is Associate Professor of Economics and Business at Hendrix College. He has published in Economica, Journal of Forecasting, Journal of Economic Surveys, *and the* Industrial and Labor Relations Review, *among others, on topics involving econometrics, statistics, and the methodology of economics. He is the methodology editor of the* Journal of Socio-Economics.

Charles S. Reichardt is Associate 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, *Volume 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.*