CHAPTER 27

REGRESSION-DISCONTINUITY DESIGNS

Charles S. Reichardt and Gary T. Henry

Regression-discontinuity (RD) designs are used to estimate the effects of treatments, programs, or interventions. The distinguishing feature of an RD design is that participants are assigned to treatment conditions on the basis of their scores on a quantitative variable. Participants with scores below a specified cutoff value on the quantitative variable are assigned to one treatment condition, whereas participants with scores above the cutoff value are assigned to no treatment or an alternative treatment. Effects of the treatments are evidenced by a discontinuity (at the cutoff score) in the relationship between the quantitative assignment variable and an outcome variable.

RD designs tend to produce more credible estimates of effects than most other quasi-experimental, nonexperimental, or observational studies. RD designs tend to produce less credible estimates of treatment effects than randomized experiments, but RD designs can sometimes be implemented in situations in which randomized experiments are not acceptable. This chapter describes the logic of the RD design as well as the design's strengths and weaknesses compared with both randomized experiments and other quasi-experiments.

THE LOGIC OF THE RD DESIGN

The prototypical RD design compares the effects of two treatments that are assigned to different groups of individuals. The two treatments could be a novel intervention and a standard intervention. Or one of the treatments could consist of a *control* condition for which no treatment is provided above and beyond what the individuals seek out on their own. For simplicity, but without loss of generality, we shall refer to one of the treatments as the *experimental treatment* and the other treatment as the *comparison treatment*.

Which of the two treatments is assigned to an individual is determined by a cutoff score on a measured variable, called the *quantitative assignment variable* (QAV), on which each individual is assessed. Those individuals with a QAV score above the cutoff are assigned to one of the two treatment conditions, whereas individuals with a QAV score below the cutoff score are assigned to the other condition. Following the assignment of individuals to conditions, the two treatments are implemented and, after the treatments have had a chance to have their effects, the individuals in both groups are assessed on an outcome variable.

Any quantitative measure can be used as the QAV in an RD design. If the experimental treatment is intended to address a problem or deficit, the QAV could be a measure of the participants' need for the ameliorative intervention or their risk of negative outcomes in the absence of such an intervention, with the treatment given to those who reveal the greatest need on the QAV. For example, Trochim (1984) reported a series of RD designs in which tests of basic academic skills were used as the QAVs to assign low-performing children to compensatory education programs. Similarly, Jacob and Lefgren (2004) created an RD design based on the Chicago Public School's policy of assigning students to

DOI: 10.1037/13620-027
APA Handbook of Research Methods in Psychology: Vol. 2. Research Designs, H. Cooper (Editor-in-Chief)

summer remediation and retention in grade on the basis of low scores on an end-of-year test of academic ability. Buddelmeyer and Skoufias (2004) reported an RD design in Mexico used to assess the effects of a contingent cash benefit for which eligibility for the program was determined by a low score on the QAV of income.

Alternatively, the experimental treatment could be a reward (such as a scholarship), the QAV could be a measure of merit, and the treatment could be given to those who exhibit the greatest merit on the QAV. For example, the RD design was first introduced by Thistlewaite and Campbell (1960) who used it to assess the effects of receiving a certificate of merit, which was awarded on the basis of superior performance on a Scholarship Qualifying Test. Seaver and Quarton (1976) assessed the effects of making the dean's list on subsequent academic performance, where the QAV for making the dean's list was the grade point average from the prior academic term. And more recently, van der Klaauw (2002) assessed the impact of financial aid offers on college enrollment, where financial aid was awarded on the basis of a QAV of academic ability.

In addition to measures of need or merit, other types of QAVs could be used to determine eligibility for a treatment or admission to a program in an RD design. For example, a treatment allocated on the basis of first-come, first-served could be assessed in an RD design using either time of arrival or time of application for the treatment as the QAV. If different treatments are made available to people residing in different geographic regions that have sharp boundaries, the physical distance from the boundary could be used as the QAV in an RD design. DiNardo and Lee (2004) assessed the effects of unionization using the vote counts in favor of unionization in individual companies as the QAV, where a majority vote in favor of unionization was the cutoff score. Cahan and Davis (1987) estimated the effects of the first year of school using age as the QAV and the minimum age required to enroll a child in school as the cutoff score. Outcomes were assessed at the end of the academic year by comparing children who had been old enough to enter first grade the year before with children who had not been old enough to enter first grade the year before (also see Gormley & Gayer, 2005, Gormley, Gayer, Phillips, & Dawson, 2005; Gormley, Phillips & Gayer, 2008; Wong, Cook, Barnett, & Jung, 2008).

Statistical precision and power increase as the correlation between QAV and the outcome measure increases. So using a QAV that is operationally identical to the outcome measure is often advantageous because it maximizes the correlation between the two. An example of operationally identical measures would be a pretest measure of cognitive ability being used to assign students to a remedial education program and a posttest measure of the same test being used as the outcome assessment. But the logic of the RD design holds regardless of the correlation between the QAV and the outcome measure. The RD design will still work, for example, even if the QAV is completely uncorrelated with the outcome measure, in which case the design takes on many of the properties of a randomized experiment.

The QAV could be derived from subjective judgments as long as those judgments are given numerical values so individuals can be ordered and a cutoff value specified. If desired, the QAV could be a composite of several separate measures for which each measure is differentially weighted. For example, Henry, Fortner, and Thompson (2010) created an RD design for which the lowest scoring school districts were assigned to treatment using an index of education advantage composed of four separate variables: teacher stability, teacher experience, children not living in poverty, and students meeting state proficiency standards. All that is required is that the separate measurements be combined quantitatively into a single index, which is used as the QAV. Neither the QAV nor any of the separate measures used in its composition need be free of measurement error. The only requirement is that the QAV be used to assign participants to treatment conditions according to a cutoff value.

It may be more difficult to select an appropriate cutoff score when, as sometimes occurs, participants "trickle" into a study rather than being assigned to treatment conditions all at one time. With trickle processing, it may be necessary to alter the cutoff score as the study progresses because the flow of participants and their QAV scores are not as anticipated with, for example, either too few or too many being eligible to receive the experimental treatment. In this case, different groups of participants could be assigned to conditions using different cutoff scores for which each group is conceptualized as a separate RD design. The data from these multiple designs could be analyzed either separately or as a single large sample if the QAV scores in each separate group are standardized so the different cutoff scores are aligned at the same location on the standardized QAV scale.

Patterns of Treatment Effects

An effect of the experimental condition is evidenced in an RD design by a discontinuity in the relationship between the QAV and the outcome variable that occurs at the cutoff score on the QAV. The presence of a discontinuity is assessed by comparing regression lines in the two treatment groups, in a way that is best explained pictorially. Figures 27.1, 27.2, and 27.3 present three outcomes for an RD design. Each of the figures displays scatterplots of the scores of the individuals in the treatment and comparison conditions. In each figure, scores on the QAV are plotted along the horizontal axis and scores on the outcome variable are plotted along the vertical axis. The cutoff score falls at the value of 30 on the QAV and is represented in the figures by a vertical line at the score of 30 on the horizontal axis. Individuals with a score on the QAV below 30 were assigned to the experimental condition, and their scores in the figures are denoted by squares. Individuals with a QAV score at or above 30 were assigned to the comparison condition, and their scores in the figures are denoted by circles. The sloped lines in the figures are the regression lines that pass through the scatter of the data points in the experimental and comparison conditions.

In Figure 27.1, the experimental treatment has no effect compared with the comparison treatment because the regressions in the two conditions fall on the same line. In Figure 27.2, the experimental

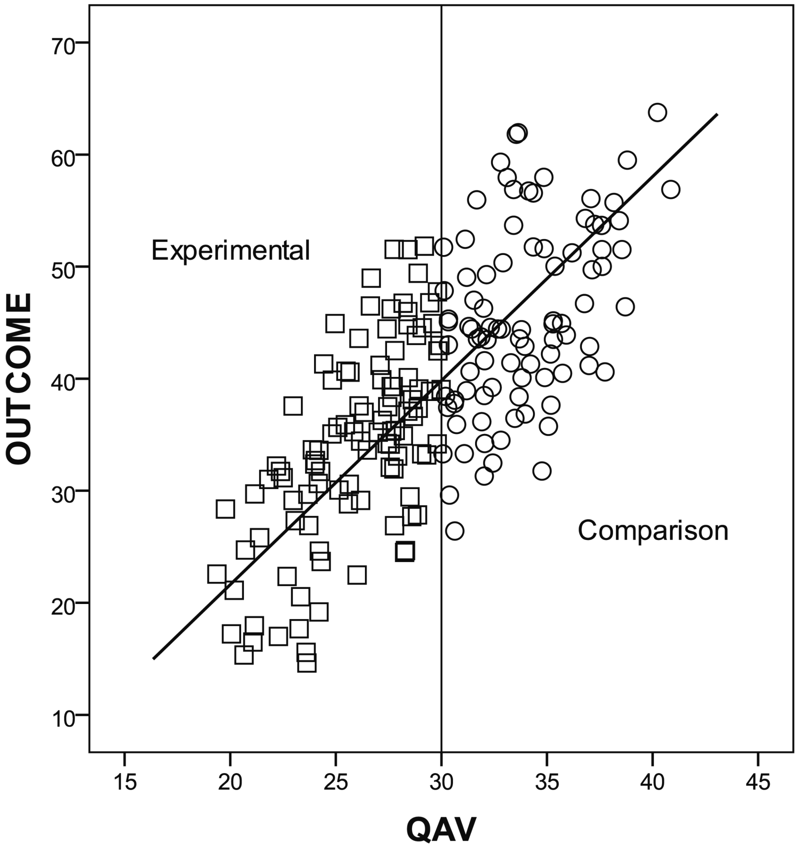

FIGURE 27.1. Hypothetical data showing no treatment effect.

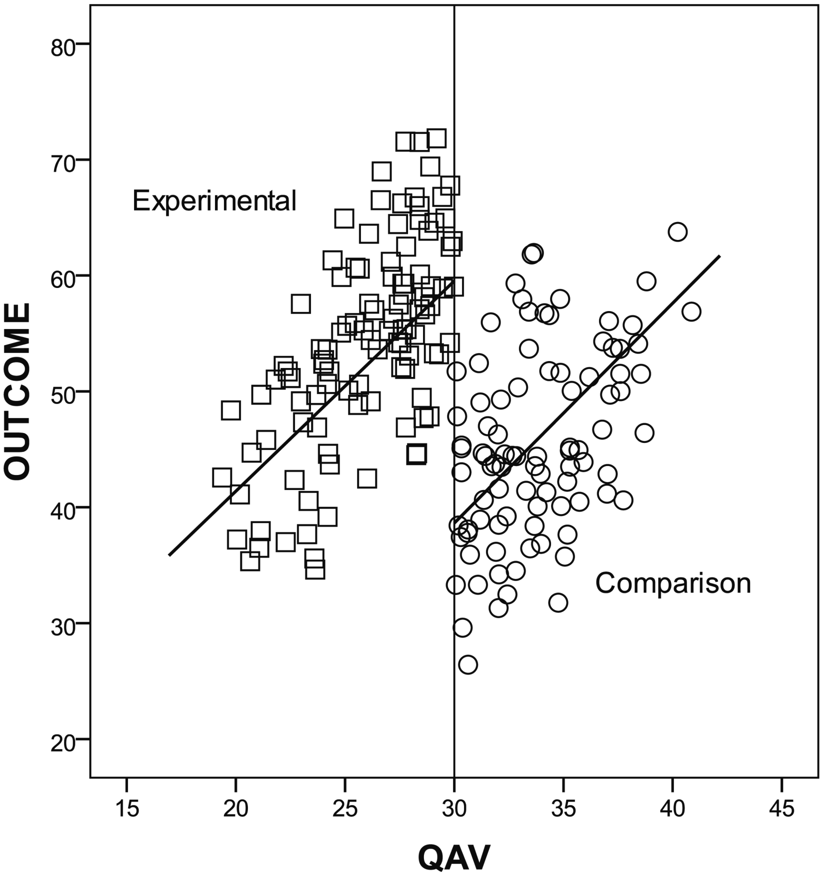

FIGURE 27.2. Hypothetical data showing a positive treatment effect.

treatment has a positive effect compared with the comparison treatment because the regression line in the experimental group is displaced upwardly (positively) compared with the regression line in the comparison group. In other words, the presence of a break or discontinuity between the regression lines in the experimental and comparison conditions at the cutoff score reveals the presence of a treatment effect. Absence of a discontinuity indicates the lack of a treatment effect at the cutoff score.

Figure 27.3 depicts yet another potential outcome of an RD design. As in Figure 27.2, the regression line from the experimental group in Figure 27.3 is displaced vertically compared with the regression line for the comparison group, which reveals that the experimental condition has a positive effect compared with the comparison condition. But unlike in Figure 27.2, the regression lines in the two groups in Figure 27.3 are not parallel, which reveals the experimental treatment has a different effect for individuals with different QAV scores. That is, the degree of vertical displacement between the two regression lines in Figure 27.3 depends on the score on the QAV. Were the regression line from the comparison group to be extrapolated to the left of the cutoff score, the vertical discrepancy between the regression line from the experimental group and the extrapolated regression line from the comparison group would be larger for individuals with lower QAV scores. This means an interaction between the treatment and the QAV is present. A discontinuity or break in the regression lines at the cutoff point, as illustrated in both Figures 27.2 and 27.3, is called an effect of a *change in level*, whereas nonparallel regression lines, as illustrated in Figure 27.3, are called an effect of a *change in slope*.

The Statistical Analysis of Data From RD Designs

The analysis of covariance (ANCOVA) model (which is a special case of multiple regression) is the classic method for analyzing data from an RD design. We

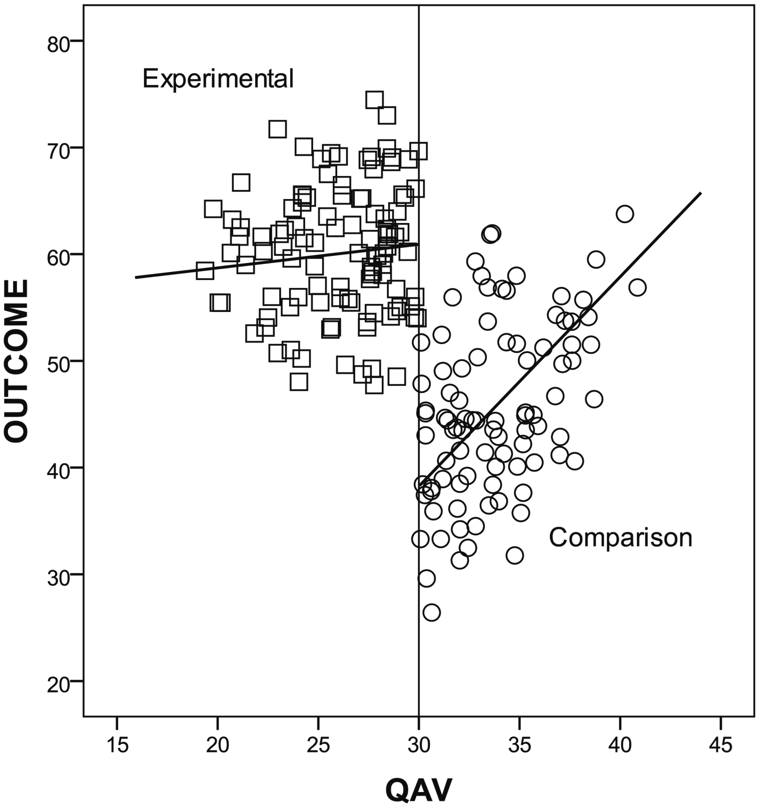

FIGURE 27.3. Hypothetical data showing a treatment effect that produces a change in level and a change in slope.

introduce the ANCOVA model in its simplest form and then describe a variety of embellishments that can be added to the elementary model.

Change in level. The simplest ANCOVA model for the analysis of data from an RD design is

$$Y = a + B_t\,T + B_x\,(X - X') + E. \qquad (1)$$

The model contains three observed variables: Y, T, and X. The Y variable is the individuals' scores on the outcome measure, T is a dummy variable representing assignment either to the treatment condition ($T = 1$) or to the comparison condition ($T = 0$), X is the individuals' scores on the QAV, and X' is the value of the cutoff score on the QAV. In Figures 27.1, 27.2, and 27.3, the cutoff score is 30, so X' equals 30. The notation $(X - X')$ means that the value of X' is subtracted from all the QAV scores before these scores are entered into the model as an independent variable. The model could be fit using either an ANCOVA option in a statistical package or a multiple regression option where Y is regressed onto T and $X - X'$.

The model in Equation 1 specifies that the regression of Y onto $X - X'$ (which represents the QAV scores) is a straight line that has the same slope in the two treatment groups. If the experimental treatment has an effect, it does nothing more than displace the regression line in the experimental group upward or downward compared with the regression line in the comparison group. This means the treatment effect is constant across the QAV. The value of B_t is the size of the vertical displacement of the regression line in the experimental group compared with the comparison condition. The value of B_x is the slope of the two parallel regression lines. This model would well fit the data in Figures 27.1 and 27.2, where the regression surfaces in the two groups are straight lines and parallel, and the treatment has either no effect or a constant effect across the values of the QAV. The value of the a parameter in Equation 1 is the intercept of the regression slope in the comparison condition and is usually of little interest in the analysis. The E variable in the model is the disturbance or error term, which represents all the factors not included in the regression that influence an individual's score on Y and allows the individual data points, such as in Figures 27.1 and 27.2, to scatter around the regression lines.

The precision of the estimate of the treatment effect (and the power of the statistical analysis to detect a treatment effect) in Equation 1 will generally be greatest when the cutoff score is specified so that equal numbers of participants are in the two treatment groups. But the design can still be implemented if circumstances demand the cutoff be an extreme, rather than middle, score along the QAV. In any case, if an extreme cutoff score is used, data from one of the treatment conditions should not be omitted to make the sample sizes equal. More data are better than fewer data, even if it means the treatment groups will be unequal in size.

Treatment effect interactions. A slightly more complex model is required to fit the data in Figure 27.3, in which the treatment effect alters the level as well as the slope of the regression line in the experimental group. The more complex model is as follows:

$$Y = a + B_t\,T + B_x\,(X - X') + B_{tx}\,T(X - X') + E. \quad (2)$$

The notation $T(X - X')$ means a variable is created that is the product of the dummy variable T and the $(X - \mathrm{X}')$ variable. In this model, the value of B_t is the size of the vertical displacement of the regression lines at the cutoff score. If the regression lines are not parallel, as in Figure 27.3, the size of the vertical displacement between the two regression lines depends on where the displacement is measured along the QAV. In Figure 27.3, for example, the displacement is greater for smaller values of the QAV than for larger values. By including the QAV scores scaled as $(X - X')$ in Equation 2, the model estimates the vertical displacement at the cutoff score on the QAV. If the value of X' is not subtracted from the value of X, the vertical displacement between the regression lines will be estimated at the point at which the QAV equals zero, which is likely to be uninformative or even misleading. To estimate the vertical displacement at a different value along the QAV, such as at QAV = X'', create the new variable $X - X''$ and enter that variable in place of the $X - X'$ variable in the two places it appears in Equation 2.

The value of B_x in Equation 2 is the slope of the regression line in the comparison group. The value of B_{tx} is the difference between the slopes in the experimental and comparison groups, and it represents the effect of a treatment interaction. A positive value of B_{tx} means the slope in the experimental group is steeper than in the comparison condition. For more details on fitting the statistical model and interpreting the results in the presence of an interaction, see Reichardt, Trochim, and Cappelleri (1995).

In the presence of a treatment effect interaction, methodologists often suggest placing more emphasis on the estimate of the vertical displacement at the cutoff score on the QAV than on the estimate of the vertical displacement at any other point along the QAV. This is because estimating the vertical displacement at the cutoff score requires minimal extrapolation of the regression lines in the two groups. To estimate the vertical displacement of the regression lines at any other point along the QAV would require extrapolation of the regression line from one of the treatment groups into a region where there are no data from that group. For example, to estimate the vertical displacement for a value of the QAV less than X' in Figure 27.3, the regression line for the comparison group would have to be extrapolated to the left of the cutoff point where there are no individuals from the comparison group. Estimates of treatment effects on the basis of such extrapolations tend to be both less powerful and less credible than estimates of effects at the cutoff score. On the other hand, in some instances, the experimental group might be composed only of individuals who fall at one of the extreme ends of the distribution of QAV scores, so little or no variation among the QAV scores occurs in that group. In that case, the estimate of the regression slope in the experimental group may be unstable, which could make an estimate of the vertical displacement of the regression lines at the cutoff point imprecise. Under these conditions, a more precise estimate of the vertical displacement might be obtained by replacing X' in Equation 2 with the mean of the QAV in the experimental group. With that change, the model would estimate the average effect of the treatment for the individuals in the experimental group rather than the effect of the treatment for individuals at the cutoff score.

If the regression lines in the treatment groups are parallel, fitting Equation 2, which includes the $T(X - X')$ interaction term, rather than Equation 1, can reduce the power of the analysis. But fitting Equation 1 rather than Equation 2, when the regression lines are not parallel, can bias the estimate of the treatment effect. Researchers often drop the interaction term if it is not statistically significant so as not to suffer a loss in power, but we recommend caution in dropping the term *if* doing so substantially alters the size of the estimate of the treatment effect, for fear a bias might be introduced.

Curvilinearity. To obtain an estimate of the treatment effect unbiased by selection, the relationship between the QAV and the outcome must be modeled correctly. The models in Equations 1 and 2 fit a straight regression line to the data in each treatment group. However, the regression surfaces might be curvilinear rather than linear. Figure 27.4 illustrates how fitting straight lines in the presence of curvilinearity could bias the estimate of the treatment effect. To make the illustration as simple as possible, the scatter in the data has been removed, so the data points all fall directly on top of the regression lines. As Figure 27.4 shows, the regression surface is curvilinear with no discontinuity at the cutoff score of 40 on the QAV. A model that fits the proper curvilinear regression surfaces would correctly estimate the treatment effect to be zero. But if straight lines were fit to the data, as shown in Figure 27.4, a discontinuity at the cutoff would be found along with a treatment effect interaction. In other words, an improper analysis that fit straight, rather than curved, regression lines would produce both a discontinuity and a treatment effect interaction when, in fact, neither of these effects is present.

To avoid such biases in the estimates of a treatment effect, curvilinearity in the regression surfaces must be modeled correctly. One approach to modeling

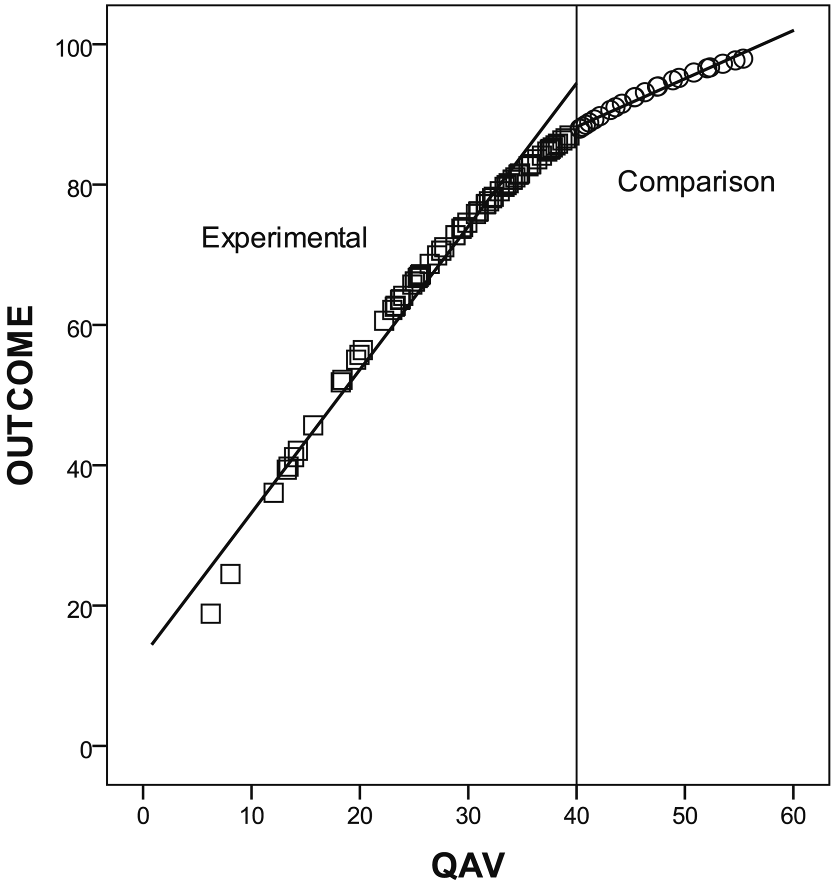

FIGURE 27.4. Bias resulting from fitting straight lines to a curvilinear relationship.

curvilinearity is to apply a nonlinear transformation to the scores on either the QAV or the outcome variable to turn a curvilinear relationship between the untransformed scores into a linear relationship between the transformed scores (Draper & Smith, 1998). The alternative, and more common, approach is to model curvilinearity by adding polynomial terms to the ANCOVA equation. For example, the following model adds a quadratic term:

$$Y = a + B_t\,T + B_x\,(X - X') + B_{tx}\,T(X - X') + B_{x2}\,(X - X')^2 + E. \tag{3}$$

Equation 3 is the same as Equation 2 except the quadratic term $B_{x2}\,(X - X')^2$ has been added. This term allows the regression surfaces in the two groups to take on a quadratic curvature. Equation 4 adds an interaction term to allow the quadratic curvature in the regression surfaces to differ across the treatment groups, which would be evidence of a treatment effect interaction:

$$Y = a + B_t\,T + B_x\,(X - X') + B_{tx}\,T(X - X') + B_{x2}\,(X - X')^2 + B_{tx2}\,T(X - X')^2 + E. \tag{4}$$

Higher order polynomial terms with or without interaction terms could be added as well. For example, both a cubic term $B_{x3}\,(X - X')^3$ and a cubic interaction term $B_{tx3}\,T\,(X - X')^3$ could be added to allow the regression surface to take on a cubic curvature and to allow for a treatment effect interaction in the shape of the cubic curvature. All polynomial terms are entered with the QAV variable scaled as $(X - X')$. Norming the X variable in this fashion means any change in level caused by the treatment effect is estimated at the cutoff score. To estimate the change in level at an alternative location along the QAV variable (say at X'' rather than X'), the value of X' should be replaced everywhere by X''.

In theory, any curvilinear shape can be fit perfectly if enough polynomial terms are added to the model. In practice, however, there are limits to the number of polynomial terms that can reasonably be added because of limits imposed by sample size and multicolinearity. Adding polynomial terms can increase the power of the statistical analysis to the extent they cause the fitted regression surface to more closely model the true regression surface. But adding polynomial terms also tends to reduce statistical power because of multicolinearity and because the polynomial terms are correlated both among themselves and with treatment assignment. Underfitting the model by including too few polynomial terms can bias the estimates of treatment effects but overfitting can severely reduce statistical power. Hence, the researcher must walk the narrow line between under- and overfitting the statistical model. Prevailing practice and advice from methodologists seems to be to fit additional, higher order polynomials as long as each additional polynomial term is statistically significant.

Plotting the relationship between the QAV and outcome scores can help in modeling curvilinearity properly. Plot the data before analyses are conducted to diagnose the nature of any curvilinearity that exists and plot the residuals after analyses have been conducted to see whether the models have properly fit the curvilinearity that was present. Including both straight and best-fitting lines in the plots can help point out departures from linearity. Best-fitting lines can be plotted using locally weighted polynomial (loess) regression, which uses subsets of the data to plot the best-fitting regression line for each point along the independent variable. Alternatively other forms of smoothing, such as mean or median smoothing can be used to view the relationship between the QAV and outcomes variable. A general rule is that the degree of the polynomial function fit in Equation 4 should be the number of inflection points in the smoothed curve plus one.

Supplementary sources of data can sometimes also be used to diagnose curvilinearity. For example, if an operationally identical measure of the outcome variable was collected before the implementations of the treatments, the plot of this prior measure versus the QAV scores could be examined. Because treatment effects cannot be present in the pretest data, any lack of smoothness in the relationship between the QAV and the pretest scores is due to curvilinearity from one source or another. Assuming the true outcome scores would behave much the same as the operationally identical pretest scores, the same curvilinearity would be presumed to arise in the true outcome scores as well. Marcantonio and Cook (1994; also see Riecken et al., 1974) provided an example of this strategy in a study of the effects of Medicaid. In 1967, families with incomes less than $3,000 were eligible for Medicaid payments, and a plot of the outcome variable in 1967 versus income revealed a dramatic discontinuity. A plot of the outcome variable versus income in a year before the start of Medicaid revealed a straight-line relationship between the two with no discontinuity, thereby increasing one's confidence that the discontinuity discovered in the 1967 data was due to the effects of Medicaid and not improperly modeled curvilinearity. Alternatively, a researcher might obtain data on both the QAV and outcome scores from a comparable cohort of individuals, perhaps from a neighboring locale, where the experimental treatment was not available. Again, the pattern of curvilinearity that was or was not present in the auxiliary data would suggest the pattern of curvilinearity to suspect in the real data of interest.

Curvilinearity can arise because the true underlying relationship between two constructs is curvilinear or because anomalies of the measurement process introduce twists or turns into an observed relationship. For example, measurement inconsistencies because of floor and ceiling effects can make an otherwise linear relationship curvilinear. It pays to be sensitive to, and on the lookout for, data to confirm or disconfirm the existence of such measurement problems. For example, curvilinearity because of floor or ceiling effects is often evidenced by a buildup of scores at the high or low ends of univariate distributions of the QAV and outcome scores. If everything else is the same, the larger the treatment effect is the less plausibly it can be explained as being caused by bias introduced by curvilinearity.

Covariates. Covariates measured before the administration of the treatment can be added to the statistical model to increase statistical power without biasing the treatment effect estimate (Imbens & Lemieux, 2008). Equation 5 adds the covariate Z to Equation 4:

$$Y = a + B_t\,T + B_x\,(X - X') + B_{tx}\,T(X - X') + B_{x2}\,(X - X')^2 + B_{tx2}\,T(X - X')^2 + B_z\,Z + E. \quad (5)$$

Interaction or polynomial terms for the covariate could also be added to model a treatment interaction or a curvilinear relationship between the covariate and the outcome measure. The power of the statistical analysis is maximized when the covariate is highly related to the outcome variable but little related to both the QAV and the other independent variables in the model. In other words, power is greatest when covariates are added that predict the outcome above and beyond any of the other variables in the model. Adding covariates that do not well predict the outcome and are correlated with the other variables in the model could reduce, rather than increase, power because of multicolinearity.

Fitting models to the data. Our presentation began with the simplest ANCOVA model and built up to more complex models. Underfitting the model by including too few terms can lead to bias, but overfitting by including too many terms can reduce the power of the analysis because of multicolinearity. Common practice is to include or exclude terms on the basis of their statistical significance. Because overfitted models may suffer from low power because of multicolinearity, we would caution against excluding terms solely because they are not statistically significant. It is important to use diagnostics such as the variance inflation factor (VIF) and to understand the effects of multicolinearity. Perhaps most important, researchers should attend to the size of the treatment effect estimates. The analyst can place the most confidence in the results if the treatment effect estimates vary little across different models as terms are added or omitted. If the sizes of the treatment effect estimates vary meaningfully as terms are dropped, even if the dropped terms are not statistically significant, it is possible that the term is not statistically significant because of its low power to detect its importance rather than because the term is not needed to model the data correctly.

In this vein, note how difficult it can be to distinguish between a model that fits a curvilinear relationship and a model that fits straight lines plus a treatment effect interaction. Such a difficulty would arise, for example, with the relatively subtle degree of curvilinearity that exists in Figure 27.4. If the data points in that figure scattered widely around the best-fitting line rather than falling directly on top of the line, both a curvilinear model and a linear model with an interaction would account for the data quite well if these two models were fit separately, therefore making it difficult to choose between them. In addition, it would be difficult to choose between the two model specifications if both curvilinear and linear interaction terms were included in the same model simultaneously because of the multicolinearity between these terms. Because of the difficulty of distinguishing between linear interactions and curvilinearity, some methodologists have suggested that an apparent interaction should not be taken as evidence of a treatment effect unless a discontinuity in level exists at the cutoff score. But such a restriction would not solve the problem of misinterpretation in Figure 27.4, in which curvilinearity, if improperly modeled, can masquerade as both a linear treatment interaction and a discontinuity in level.

Uncertainty about which model correctly fits the data will virtually always be present. In such cases, researchers should report results from a range of plausible models and draw conclusions on the basis of the range of treatment effect estimates thereby produced (Reichardt & Gollob, 1987). Sifting through models and reporting only the treatment effect estimates that are most desirable is not an appropriate analysis strategy. In addition to bracketing the size of the treatment effect by using a range of plausible models, analysts should also bracket the size of the treatment effect by repeating the statistical analyses using only the slices of data most proximate to the cutoff score (Imbens & Lemieux, 2008). For example, analyses could be repeated using data from only that half of the participants in each treatment condition whose scores on the QAV are closest to the cutoff score. Then the analysis could be repeated again using data from only that quarter of the participants in each treatment condition whose scores on the QAV are closest to the cutoff. The intuition for restricting the data to participants with QAV scores closest to the cutoff is that (a) these participants are most similar on other characteristics so the treatment effect estimates are less likely to be biased by confounding variables and (b) the treatment effect estimates are least likely to be sensitive to misspecification of the relationship between the QAV and the outcome variable. Of course, the obvious disadvantage of reducing the sample size is a loss of power and reliability.

Discontinuities in the Absence of a Treatment Effect

The analysis of data from an RD design assumes the regression of the outcome variable on the QAV would be continuous (rather than discontinuous) at the cutoff score in the absence of a treatment effect. The results of the RD analysis would be biased if the regression surface would have been discontinuous at the cutoff point even in the absence of a treatment effect. It is difficult to test the assumption of no discontinuity in the absence of a treatment effect directly. But the assumption can be tested indirectly (Imbens & Lemieux, 2008). One approach is to look for discontinuities in the relationship between the outcome variable and the QAV at locations other than the cutoff score. A discontinuity at the cutoff point that is no greater than at other points on the QAV reduces the plausibility of a treatment effect at the cutoff point, if everything else is the same. Another approach, which is often referred to as a *falsification test*, is to look for discontinuities in the relationship at the cutoff value of the QAV and the outcome variables collected before the treatment was implemented. Because discontinuities in such relationships cannot be a result of the treatment, the presence of such discontinuities raises the suspicion that any discontinuity in the relationship between the QAV and the real outcome variable is also not due to the treatment.

A discontinuity in the absence of a treatment effect could be introduced if another treatment were implemented concurrently with the treatment under study (which is called a violation of the assumption of no hidden treatments; Rubin, 2005). For example, imagine estimating the effects of Medicaid payments that are made available to anyone with income below the poverty line when, at the same time as Medicaid is introduced, other transfer payments, such as food stamps, are also introduced using the same eligibility criteria of income below the poverty line. In that case, the RD design would estimate the joint effects of Medicaid and the other transfer payments rather than the effects of Medicaid alone. Researchers should explicitly consider the following three other potential sources of discontinuity.

No shows. If individuals eligible to participate in the study know their scores on the QAV and know the cutoff score required to place them into a desired treatment, those who fail to qualify for a desired experimental treatment may decide not to show up for the study. This can introduce a discontinuity between the regression lines in the treatment groups, which makes the experimental treatment look more effective than it really is because the most attentive individuals are removed from the less desirable treatment condition more than from the desirable condition. To avoid this bias, it is best to keep the cutoff score confidential or make it impossible to determine one's QAV score ahead of time.

Attrition. Participants sometimes drop out of research studies once they have begun or fail to complete the outcome measurements. Such participants produce *incomplete* or *missing* data. Estimates of the treatment effects in the RD design can be biased because of missing data, especially when data are missing because participants drop out differentially across the treatment conditions. The best strategy is to prevent attrition. The means of preventing attrition and coping with missing data are much the same in randomized experiments as in RD designs, and readers are advised to consult the literature in that area for advice (Schafer & Graham, 2002).

Misassignment to treatment conditions. Participants sometimes "cross over" from one treatment condition to another to receive a treatment to which they should not have been assigned according to the cutoff rule in the RD design. For example, administrators or researchers might respond to pressure to admit participants into a desired treatment when their QAV scores fall just below the cutoff score needed to obtain that treatment because those participants are particularly deserving or demanding of the desired treatment. Alternatively, misassignment might arise when the cutoff score is known to the participants ahead of time, the values of the QAV scores are reported by the participants (rather than measured independently by the researchers), and participants lie about, or otherwise manipulate, their QAV scores to push their score across the cutoff and thereby receive the treatment they most desire. (This provides another reason to keep the cutoff score hidden from the participants before the treatments are assigned.) Or individuals assigned to a less desirable treatment might arrange to receive the more desirable treatment from a source outside the study. Often participants with QAV scores nearest the cutoff are most likely to cross over, resulting in what has come to be known as a *fuzzy* assignment.

Treatment crossovers can bias the estimates of the treatment effect because the more motivated or desperate participants tend to cross over from one treatment to the other and are therefore underrepresented in one group and overrepresented in the other. Methods developed to address the problems introduced by treatment crossovers in randomized experiments are applicable to RD designs as well. If the true values of the QAVs are known, the simplest strategy is to analyze the data according to how participants should have been assigned to the treatment conditions, rather than according to the condition they actually received (Boruch, 1997). Such an analyze-as-assigned-rather-than-as-treated strategy will tend to produce an underestimate of the treatment effect and is called the intent-to-treat estimate. The intent-to-treat estimate may be conservative, but it is considered better than using an analyze-as-treated estimate, which is more likely to produce biases of unknown direction.

Methodologists have also suggested the following four additional strategies for coping with the effects of fuzzy assignment in the RD design. First, if misassignments appear to be restricted to a narrow range near the cutoff score, bias can be avoided by conducting the statistical analysis and omitting all the scores inside this range. Second, if misassignment would occur because researchers or administrators insist subjective criteria be used to determine treatment assignment, these subjective assessments can be quantified and made a part of the QAV. By making the subjective assessments that would be responsible for a misassignment part of the QAV, one presumably removes the incentive for researchers or administrators to misassign participants. Third, those individuals who are likely to be assigned to a given treatment even if their QAVs fall on the "wrong" side of the cutoff score can be omitted from the study. If this last strategy is used, individuals should be omitted from the study before examining their QAVs, otherwise there will be a tendency to drop participants differentially from the two groups, which could introduce a bias. That is, if QAVs are known, there will be a tendency to omit from the analysis only those with QAVs that fall on one side of the cutoff, namely, the side that failed to qualify for the most desirable treatment. Fourth, using the QAV as an identifying instrument in a two-stage least squares, instrumental variable analysis is also an available strategy but beyond the scope of the current chapter (Foster & McLanahan, 1996; van der Klaauw, 2002, 2008).

The distribution of QAV scores. If one of the treatment conditions is more appealing than the other, the three sources of bias (i.e., no shows, attrition, and crossovers) will tend to alter the distribution of scores on the QAV on one side of the cutoff score as compared with the other. That is, no-shows, attrition, and crossovers would be expected to produce either a localized bulge in the height of the distribution of the QAV scores on one side of the cutoff score or a localized dip in the height of the distribution on the other side, or both. Therefore, evidence of these three sources of bias can be obtained by plotting the frequency distribution of the QAV scores and looking for a discontinuity at the cutoff score. McCrary (2008) provided a test of the statistical significance of such a discontinuity in the QAV frequency distribution.

Elaborations of the Prototypical RD Design

The preceding sections have considered only the simplest RD design. The simple RD design, however, can be embellished in a variety of ways to better tailor the design to the demands of the research setting.

More than one cutoff score. The prototypical RD design compares two treatment conditions and assigns participants to those conditions using a single cutoff score on the QAV. In addition, an RD design could be used to compare two treatment conditions using two cutoff scores in which case one of the treatment conditions is assigned to participants with scores in between the two cutoff scores and the other condition is assigned to participants with scores beyond either of the two cutoff scores. Alternatively, two cutoff scores could be used to compare three different treatment conditions. The statistical model for data from RD designs with more than one cutoff score would include additional dummy variables and interaction terms to estimate changes in level and slope at each cutoff score.

RD designs combined with randomized experiments. An RD design can be combined with a randomized experiment (Boruch, 1975; Shadish, Cook, & Campbell, 2002). When comparing two treatments, one combination of designs would use two cutoff scores to create two extreme groups of participants on the basis of their QAV scores. One of the extreme groups on the QAV measure would receive the comparison condition, the other extreme group would receive the experimental condition, and those in the middle (in between the two cutoff scores) would be randomly assigned to the treatment conditions. This design could satisfy a desire by administrators to assign most participants to treatments on the basis of need or merit, while acknowledging that, because measures of need or merit are fallible, it would be most equitable to give all the participants who fell within a middle range on the QAV an equal chance to receive the experimental treatment.

Another design option, using a single cutoff score, would be to assign individuals with QAV scores on one side of the cutoff to one of the treatment conditions and assign individuals with QAV scores on the other side of the cutoff score to one of the two treatment conditions at random. For example, everyone with scores at one end of the QAV could be given the experimental treatment, whereas those with scores on the other side of the cutoff could be assigned to the experimental and comparison conditions at random. Designs that combine randomized experiments with RD designs are likely to be more powerful and produce results that are more credible than those produced either by RD designs without random assignment or by random assignment without the additional data from the RD portion of a design for which individuals are not assigned to treatments at random.

Cluster designs. Cluster RD designs are analogous to cluster randomized experiments in that groups or clusters of individuals, such as schools, classrooms, or clinics, are assigned to treatment conditions. In cluster RD designs, assignment to treatment conditions is based on a QAV measured at the cluster level so all individuals in a given cluster are assigned to the same treatment condition. For example, Henry et al. (2010) used a cluster design to assess the effects on student achievement of supplemental funding awarded to schools at the district level, where the QAV was a measure of district educational disadvantage. In cluster designs in which outcomes are measured at the individual level, but treatments are assigned at a higher level, multilevel models can maximize the power of the analysis to detect treatment effects. For example, Henry et al. used a multilevel model to assess the effect of supplemental funding at the district level on outcomes measured at the individual student level, controlling for the students' prior achievement as well as other individual, classroom, and school characteristics. However, Schochet (2008) has shown that the power of cluster designs is less than the power of noncluster designs and that power tends to increase as the number of clusters, rather than the number of participants within a cluster, increases.

RD Designs Employing Comparisons Across Settings, Outcome Variables, or Times

In classic RD designs, as described in the preceding sections, participants are measured and assigned to treatment conditions (either individually or in clusters) on the basis of a QAV. Because the treatment effect is estimated by drawing a comparison across participants, such designs are called RD designs comparing participants. Three other types of RD designs are also possible based on drawing comparisons across either settings, outcome variables, or times (Reichardt, 2006). Each of these three types of RD designs is described in the following paragraphs.

First, to determine whether adding traffic lights to highway intersections reduces traffic accidents, imagine a design in which traffic lights are installed at intersections using a quantitative assignment rule based on the volume of traffic passing through the intersection during the previous month. In other words, imagine a design in which a representative set of intersections without traffic lights is selected, traffic lights are added to the intersections that had the heaviest traffic during the preceding month, and the number of traffic accidents at each intersection during the ensuing year is tallied. The effect of adding traffic lights is then estimated by regressing the number of accidents during the ensuing year onto the QAV of traffic volume at the intersections, and measuring the size of any discontinuity in the regression lines at the cutoff point that demarcates those intersections that received a traffic light and those that did not. That is, settings that receive a traffic light are compared with settings that do not receive a traffic light to determine how people behave differently in the different settings. Because the treatment effect is estimated by drawing a comparison across different settings, such a design is called an RD design comparing settings.

Second, consider a design in which letters of the alphabet are assigned to different treatment conditions using a quantitative assignment variable. For example, imagine assessing the effectiveness of a new educational television series designed to teach prereading skills. During the 1st year of production, the show is able to teach only half the letters of the alphabet, and the producers of the show want to teach the most important letters. The frequency with which each letter appears in the English language is measured, and this measure is used as the QAV. The 13 most frequently appearing letters are taught during the 1st year of the show. At the end of the 1st year, a group of children who have been viewers of the show are tested to assess their knowledge of all 26 letters. The effect of the show is estimated by regressing the scores on the outcome measures onto the QAV and looking for a discontinuity between the regression lines across the two groups of letters. Because the treatment effect is estimated by comparing performances across letters of the alphabet (where each letter is a different outcome variable), such a design is called an RD design comparing outcome variables.

Third, the *interrupted time-series* (ITS) design (see Chapter 32 of this volume) is an RD design in which different temporal occasions are assigned to different treatment conditions using time as the QAV (Marcantonio & Cook, 1994). Hence, an ITS design is called an RD design comparing times.

ITS designs and classic RD designs (i.e., RD designs comparing participants) are widely recognized in the literature on quasi-experimentation. The other two types of RD designs (i.e., RD designs comparing settings and RD designs comparing outcome variables) are not nearly as well known but could often be used to advantage. For example, the original evaluations of the effects of *Sesame Street* (Cook et al., 1975) employed a relatively weak nonequivalent group (NEG) design (see Chapter 26 of this volume) but could have used a more credible RD design comparing outcome variables as in the preceding example of an evaluation of an educational television show comparing letters of the alphabet.

RELATIVE STRENGTHS AND WEAKNESSES OF THE RD DESIGN

NEG designs tend to be easier to implement than RD designs. The reason is that NEG designs place no restrictions on how units are assigned to treatment conditions, whereas RD designs require that units (i.e., participants, times, settings, or outcome variables) be assigned according to a quantitative assignment rule. In addition, an RD design may have less statistical power than an NEG design. Because an RD design allows no overlap between the treatment groups on the QAV whereas the treatment groups in a NEG design could overlap substantially on covariates, the power of RD designs tends to be reduced by multicolinearity more than the power of NEG designs.

Estimates of effects from RD designs tend to be more credible than estimates from NEG designs, especially in light of recent evidence comparing estimates of effects from RD designs to those from randomized experiments and NEG designs (Cook, Shadish, & Wong, 2008). The nature of selection differences is known in RD designs because RD designs impose a quantitative assignment rule. In contrast, the nature of selection differences is usually unknown in NEG designs, so the specification and modeling of selection differences is more difficult and leads to less credible estimates (Reichardt, 1979).

In contrast, estimates of treatment effects derived from randomized experiments tend to be more credible than estimates from RD designs. The reason is twofold. First, the effects of selection differences must be modeled in RD designs by using the QAV as a covariate in a regression analysis. Using an improper model (such as fitting linear regression surfaces when the true regression shape is curvilinear) can bias the estimates of treatment effects. In randomized experiments, selection differences between groups are random, which can be modeled without using covariates to fit a regression surface, so there is less chance of error and bias. Second, estimating treatment effect interactions (in which case the effect of the treatment varies across QAV scores) in RD designs involves extrapolating the regression line from the comparison condition into a region on the QAV that contains no data from the comparison group and extrapolating the regression line from the experimental condition into a region on the QAV that contains no data from the experimental condition. In contrast, treatment effect interactions in randomized experiments can be estimated without extrapolating regression surfaces into regions that do not contain relevant data.

The results from RD designs are also less precise and powerful than from randomized experiments because the QAV and the treatment-assignment dummy variable are correlated in the RD design. In contrast, any covariates included in the analysis of data from a randomized experiment are uncorrelated with the treatment-assignment dummy variable and therefore cannot diminish precision and power because of multicolinearity. To obtain the same precision and power as in a randomized experiment, an RD design must have more than two times as many participants (Cappelleri, Darlington, & Trochim, 1994; Goldberger, 1972; Schochet, 2008). In addition, a larger sample size is needed in an RD design, as compared with a randomized experiment, to ensure that the regression surface between the QAV and outcome score is modeled correctly.

Randomized experiments can be more difficult to implement than RD designs. Whether for ethical or practical reasons, situations arise in which administrators and participants are more likely to resist the random assignment of a desirable treatment than assignment on the basis of a measure of need or merit. Under such circumstances, it can be easier for researchers to implement an RD design than a randomized experiment.

CONCLUSION

NEG designs are often easier to implement than RD designs and can have more power than RD designs. But RD designs tend to produce treatment effect estimates that are more credible than estimates from NEG designs. The credibility of the RD design derives from the fact that units (i.e., participants, settings, outcome variables, or times) are assigned to treatment conditions on the basis of a quantitative assignment rule. Knowing the quantitative rule by which units are assigned to treatments allows the researcher to model the effects of selection differences between the treatment groups with greater confidence than is possible in NEG designs in which the rule by which units are assigned to treatments is not explicit or quantitative.

Conversely, RD designs tend to produce less credible results than do randomized experiments. And randomized experiments are also more powerful. But RD designs can sometimes be implemented in cases in which randomized experiments cannot.

In spite of its potential advantages compared with NEG designs and randomized experiments, the RD design has been used relatively infrequently in psychological research in the past. The primary reason, we suspect, is that many psychological researchers are simply unaware of the design and its advantages. Because of its relative strengths, the RD design is receiving increased attention and emphasis from funding agencies, such as the Institute for Educational Science (Cook & Wong, in press). The design has also received a great deal of attention in the economics literature in recent years (Cook, 2008). We suspect the RD design is poised for a similar surge of interest among research psychologists, especially in applied areas of research in which randomized experiments are not always practical.

References

Boruch, R. F. (1975). Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods and Research*, *4*, 31–53. doi:10.1177/004912417500400103

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Buddelmeyer, H., & Skoufias, E. (2004). *An evaluation of the performance of regression discontinuity design on PROGRESA*. Bonn, Germany: Institute for the Study of Labor.

Cahan, S., & Davis, D. (1987). A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal*, *24*, 1–12.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review*, *18*, 141–152. doi:10.1177/0193841X9401800202

Cook, T. D. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics, and econometrics. *Journal of Econometrics*, *142*, 636–654. doi:10.1016/j.jeconom.2007.05.002

Cook, T. D., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). *"Sesame Street" revisited*. New York, NY: Russell Sage.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750. doi:10.1002/pam.20375

Cook, T. D., & Wong, V. C. (in press). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique*.

DiNardo, J., & Lee, D. S. (2004). Economic impacts of new unionization on private sector employers: 1984-2001. *Quarterly Journal of Economics*, *119*, 1383–1441. doi:10.1162/0033553042476189

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York, NY: Wiley.

Foster, M. E., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods*, *1*, 249–260. doi:10.1037/1082-989X.1.3.249

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

Gormley, W. T., Jr., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of

Tulsa's pre-K program. *Journal of Human Resources*, *40*, 533–558.

Gormley, W. T., Jr., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*, 872–884. doi:10.1037/0012-1649.41.6.872

Gormley, W. T., Jr., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science*, *320*, 1723–1724. doi:10.1126/science.1156019

Henry, G. T., Fortner, C. K., & Thompson, C. L. (2010). Targeted funding for educationally disadvantaged students: A regression discontinuity estimate of the impact on high school student achievement. *Educational Evaluation and Policy Analysis*, *32*, 183–204. doi:10.3102/0162373710370620

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635. doi:10.1016/j.jeconom.2007.05.001

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*, 226–244. doi:10.1162/003465304323023778

Marcantonio, R. J., & Cook, T. D. (1994). Convincing quasi-experiments: The interrupted time series and regression-discontinuity designs. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 133–154). San Francisco, CA: Jossey-Bass.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*, 698–714. doi:10.1016/j.jeconom.2007.05.005

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

Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods*, *11*, 1–18. doi:10.1037/1082-989X.11.1.1

Reichardt, C. S., & Gollob, H. F. (1987). Taking uncertainty into account when estimating effects. In M. M. Mark & R. L. Shotland (Eds.), *Multiple methods for program evaluation* (New Directions for Program Evaluation, No. 35, pp. 7–22). San Francisco, CA: Jossey-Bass.

Reichardt, C. S., Trochim, W. M. K., & Cappelleri, J. C. (1995). Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review*, *19*, 39–63. doi:10.1177/0193841X9501900102

Riecken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. K., Jr., Pratt, J. W., . . . Williams, W. (1974). *Social experimentation: A method for planning and evaluating social intervention*. New York, NY: Academic Press.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331. doi:10.1198/016214504000001880

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037/1082-989X.7.2.147

Schochet, P. Z. (2008). *Technical methods report: Statistical power for regression discontinuity designs in education evaluations*. Washington, DC: Institute for Education Sciences, National Center for Education Evaluation and Regional Assistance.

Seaver, W. B., & Quarton, R. J. (1976). Regression-discontinuity analysis of dean's list effects. *Journal of Educational Psychology*, *66*, 459–465.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post-facto experiment. *Journal of Educational Psychology*, *51*, 309–317. doi:10.1037/h0044319

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

van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression discontinuity approach. *International Economic Review*, *43*, 1249–1287. doi:10.1111/1468-2354.t01-1-00055

van der Klaauw, W. (2008). Regression discontinuity analysis: A survey of recent developments in economics. *Labour*, *22*, 219–245. doi:10.1111/j.1467-9914.2008.00419.x

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*, 122–154. doi:10.1002/pam.20310