SAGE CLASSICS

The goal of the **Sage Classics Series** is to help bring new generations of social scientists and their students into a deeper, richer understanding of the roots of social science thinking by making important scholars' classic works available for today's readers. The Series will focus on those social science works that have the most relevance for impacting contemporary thought, issues, and policy—including increasing our understanding of the techniques, methods, and theories that have shaped the evolution of the social sciences to date.

The works chosen for the Series are cornerstones upon which modern social science has been built. Each volume includes an introduction by a preeminent leader in the discipline, and provides an historical context for the work while bringing it into today's world as it relates to issues of modern life and behavior. Moving into the future, opportunities for application of the findings are identified and references for further reading are provided.

It is our hope at Sage that the reissuance of these classics will preserve their place in history as having shaped the field of social science. And it is our wish that the Series, in its reader-friendly format, will be accessible to all in order to stimulate ongoing public interest in the social sciences as well as research and analysis by tomorrow's leaders.

Sage Classics 1

SOCIAL EXPERIMENTATION

Donald T. Campbell
M. Jean Russo

SAGE Publications
International Educational and Professional Publisher
Thousand Oaks London New Delhi

OVERVIEW OF CHAPTER 14

Suppose you are an administrator of a state-run program in your county that provides counseling and social services to juvenile offenders. You and your staff develop a special approach that you believe will reduce recidivism, and you are given permission to implement this approach in your county. If it is successful, the program director would like to disseminate your approach to the other counties in the state. How will you demonstrate that your program is successful?

This chapter, coauthored by Trochim and Campbell, may provide an inexpensive yet valid way to determine whether your approach had an impact. The authors discuss the Regression Point Displacement Design (RPDD), which can deal with large units such as schools, cities, or counties using aggregated instead of individual data. A single unit receiving an ameliorative treatment can be compared to a number of control units. Pretests and posttests are required, although it is not necessary that the same measure be used. The RPDD regresses the posttest on the pretest and fits a regression line for the control units only. A *t*-test determines if the distance (or displacement) of the treated unit is significantly different from its expected value on the regression line. If there was no treatment effect, one would not expect the treatment group to differ at greater than

chance levels from the regression line. In the situation described previously, this design could be applied rather effortlessly, because it is likely that recidivism rates at regular intervals are already available for every county in the state and can provide the pretest and posttest measures.

The RPDD is characterized by four features: A single treatment unit is often used, aggregated data are generally used rather than individual data, the groups do not have to be equal prior to treatment, and the observed regression line is used for the analysis rather than as a statistical adjustment for pretreatment differences.

The usefulness of the design is demonstrated in a discussion of the Medicaid Regression Point Displacement Design and the Schizophrenic Reaction Time Study. In both of these applications, the RPDD showed that the treatment groups differ significantly from the regression line. The authors warn, however, that because this is a quasi-experimental design, rival explanations for the noted effect must be ruled out. In the Schizophrenic Reaction Time Study, for example, the effect is not apparent when the data from the individuals (rather than the grouped data) are analyzed. The authors argue that the RPDD can never have greater power than such a micro-level analysis, and that the significant effect noted using the RPDD is a result of the small error term that occurred when the regression line was fitted to only three points.

The basic requirements for the RPDD are (a) multiple control groups, and (b) pretest and posttest measurements. There are many variations of this design, however, based on five dimensions: how the treatment group is chosen, the entity that will be measured, the number of treatment groups, same or different measures for pre- and posttest, and the number of covariates.

The authors devote the final section of the chapter to the threats to validity that are most problematic when using this design. Both selection bias and regression artifacts can be ruled out if the treatment groups are chosen randomly or selected based on a sharp cutoff point. Doing so makes the RPDD akin to the Randomized Experiment or the Regression Discontinuity Design, respectively. Bias can be introduced, however, if there is measurement error in the pretest. When this occurs, the regression line is rotated slightly, and the bias becomes more extreme as the distance increases between the treatment group pretest mean and the

overall pretest mean. Instrumentation can be a problem if the measurement process has changed in the treatment group between the pretest and the posttest, and local history should be examined to determine if differences exist in the treatment group setting that might affect posttest performance. The power of the RPDD is diminished somewhat because of the few points typically used; however, this is offset by a gain in power resulting from more reliable aggregated data. The authors feel that this issue needs further investigation. The *t*-test used in the statistical analysis of this design assumes that the control groups are a random sample of the population. Because this assumption can never be met, control groups should be selected based on their variability.

Unfortunately, the external validity of this design is low. What does this mean, then, if you found that your approach to reduce recidivism for juvenile offenders was effective? You might suggest that the approach be tried first in counties with similar pretest recidivism rates, because you can generalize with more confidence to similar groups. Unfortunately, you cannot be sure that your approach would be as effective in markedly different counties.

FOURTEEN

DESIGN FOR COMMUNITY-BASED DEMONSTRATION PROJECTS

This chapter describes an old but neglected quasi-experimental research design. Figure 14.1 reprints its most widely distributed exemplar, used by Riecken et al. (1974, p. 115), and Cook and Campbell (1979, pp. 143–146), neither source presenting any statistical analysis. They cite Fleiss and Tanur (1972) from whom we borrow Figure 14.2, in which is claimed an effect significant at the $p < .05$ level. They in turn cite H. F. Smith (1957) and Ehrenberg (1968) as at least partial predecessors. In the examples of Figure 14.1 and 14.2, the measures employed on the x- and y-axes are quite different. This option is shared with the regression–discontinuity (RD) design (Trochim, 1984), which is related to the design described in this chapter. Our Figure 14.3 illustrates this application, drawn from the same data set as Figure 14.1. Although generally the interpretability of the design will be

Trochim, W. M. K., & Campbell, D. T. (1996). *The regression point displacement design for evaluating community-based pilot programs and demonstration projects.* Unpublished manuscript.

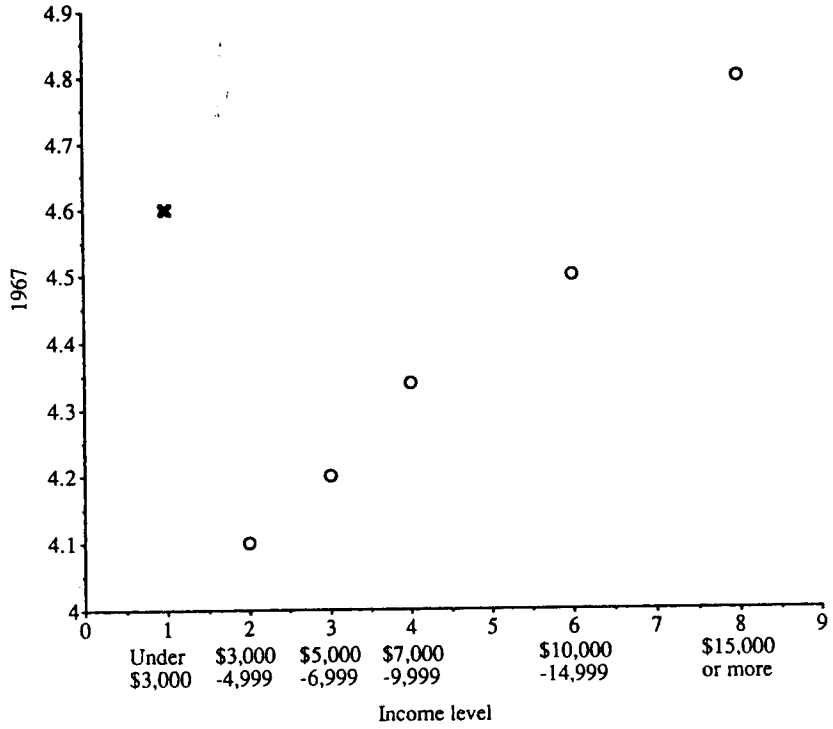

Figure 14.1. Medicaid example modified from Riecken et al. (1974). Used by permission of Academic Press.

substantially greater when the same measure is used before and after, we will not argue that this is so for Figure 14.3.

We envisage a typical application as employing repeated (e.g., annual) rate measures, for cities or some larger or smaller reporting units, with one (or several) units receiving an intensive ameliorative effort not given in the others. Because of this anticipated usage, we shall refer to the x-axis as the *pretest* and the y-axis as the *posttest* in what follows. Some treatment or program, very often likely to be a community-based pilot program or demonstration project, is administered to the treated unit. Although in Figure 14.1 the treated unit was the most extreme on the pretest, the method is applicable no matter where in the distribution of pretest values the "experimental" unit falls. Indeed, the statistical power and interpretability is likely to be better for mid-distribution

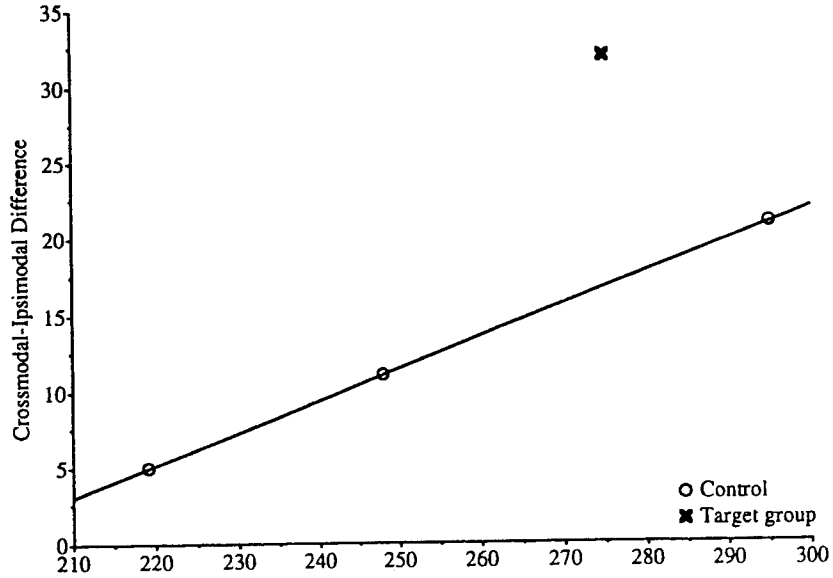

Figure 14.2. Fleiss and Tanur (1972) graph.

demonstration sites. The untreated units we will identify as *control units* or *control groups*. The analysis fits a regression line to the control units and tests the significance of the departure of the experimental unit from that regression line. The name suggested for this design is the Regression Point Displacement Design (RPDD).

Demonstration programs and pilot projects usually receive only very weak and methodologically suspect evaluation. Often there is only a comparison of rates for that one unit for a time period before the special effort and a time period afterward. Or a single comparison unit (e.g., another city similar to the demonstration one) is employed, inevitably differing in many ways. The potential power of the RPDD comes from using numerous untreated units as "controls," and from not requiring pretreatment equality between any one of them and the experimental unit.

The RPDD remains very much a quasi-experimental design, for which many of the common threats to validity must be examined on the basis of contextual information not included in the statistical analysis. If a statistically

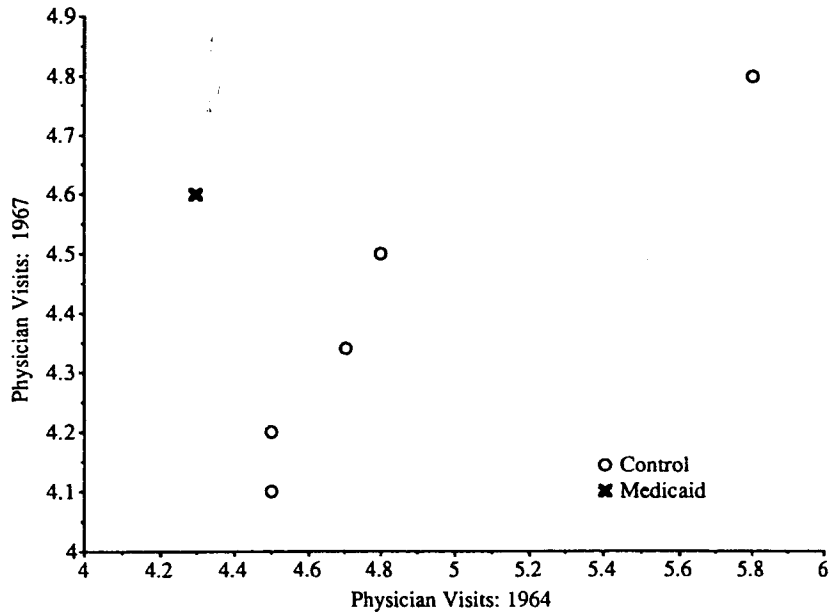

Figure 14.3. Medicaid RPDD with physician visit for both pre- and posttreatment from Riecken et al. (1974). Used by permission of Academic Press.

significant displacement is shown, there are many other possible causes that need to be considered, over and above the demonstration program.

The RPDD can be illustrated with a simple hypothetical example. Consider a single site at which a treatment is administered. Furthermore, assume that there are arbitrarily ten other sites that will not get the treatment (control sites) but will be measured. All available persons at both the treatment and control sites are measured before the treatment and at some specific time after the treatment. Note that as a result of normal turnover rates and absenteeism at each site, the persons measured at the pretest may not be the same as those measured at the posttest. The resulting data might look like the simulated values depicted in Figure 14.4.

The figure shows the linear regression line of the ten control group pretest–posttest pairs of means. The vertical line indicates the posttest displacement or "shift" of the treatment group from the regression-line predicted value.

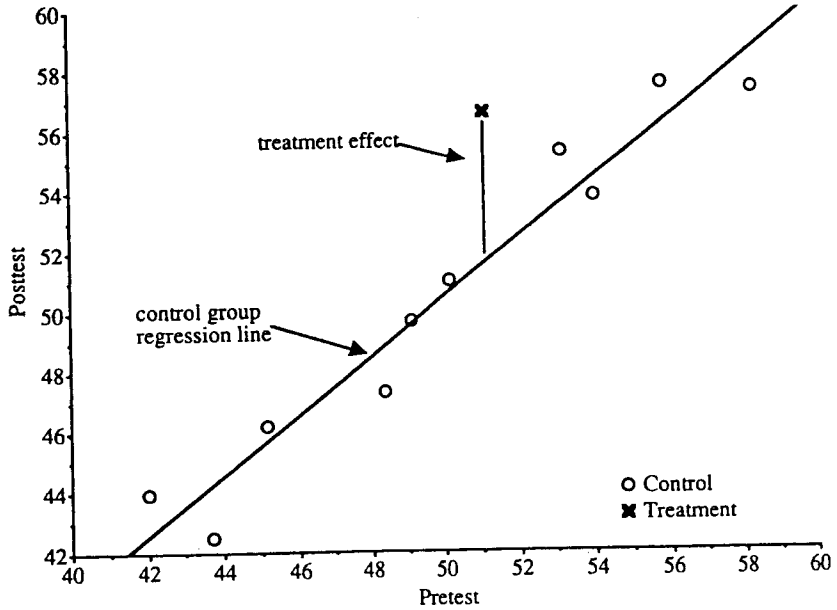

Figure 14.4. Hypothetical RPDD using simulated data.

In this case, it is visually clear that the displacement probably exceeds the normal variability one might expect around the regression line and indicates a likely treatment effect.

The central idea of the design is that in the null case, one would not expect the treatment group to differ at greater than chance levels from the regression line of the population. There is evidence for a treatment effect when there is a posttest (vertical) displacement of the treated group point from the control group regression line. This led to the name *regression point displacement.* Of course, this evidence does not imply that the treatment of interest is what "caused" this vertical regression shift. One must always assess the plausibility of other potential causes for such a shift.

The RPDD is characterized by four major features: (a) the use of a single treated unit instead of many; (b) the use of aggregate-level data instead of individual-level; (c) the absence of any need to ensure (or attempt to achieve by statistical adjustment) pretreatment equality between treated and control

groups; and (d) the avoidance of "regression artifacts" or underadjustment as a result of "errors in variables" by employing the observed regression line, rather than using it as a means of adjustment. Feature d, and to a lesser extent c, are shared with the regression–discontinuity (RD) design. The first two are not absolutely necessary. We would still probably classify a study with two or even three demonstration sites as an RPDD. The key distinction is that in alternative designs there are enough points to allow one to fit the same model to both groups, whereas the RPDD typically does not. For instance, in the RD design, one usually has enough points in both groups to be able to estimate a within-group slope, whereas in the RPDD, that is not possible or justifiable, given the few available treatment group points. The higher aggregation level (e.g., cities) is also a typical RPDD characteristic, although it is by no means required. One can envision an RPDD design involving only a single person who receives a treatment, with multiple control persons.

THE MEDICAID REGRESSION POINT DISPLACEMENT DESIGN

The first example comes from the Medicaid study discussed in Riecken et al. (1974, p. 115) and Cook and Campbell (1979, pp. 143–146), part of which was shown in Figures 14.1 and 14.3. The original data are shown in Figure 14.5 (taken from Lohr, 1972; Wilder, 1972, p. 5, Table B).

The figure shows the average number of physician visits per person per year in the United States for the years 1964, 1967, and 1969, broken out by family income ranges. The Medicaid program was introduced in 1964. The legislation mandated that only those families with an annual income of less than $3,000 were eligible to receive Medicaid. Overall, it appears that the annual average number of physician visits is declining for most income groups with two notable exceptions—the lowest income group shows an increase over both time intervals and the second lowest group increases between 1967 and 1969.

The central question is whether the introduction of Medicaid is associated with a significant increase in the average number of physician visits per year. Several RPDDs can be constructed from these data. The first is identical to that shown in Riecken et al. (1974, p. 115) and Cook and Campbell (1979, pp. 143–

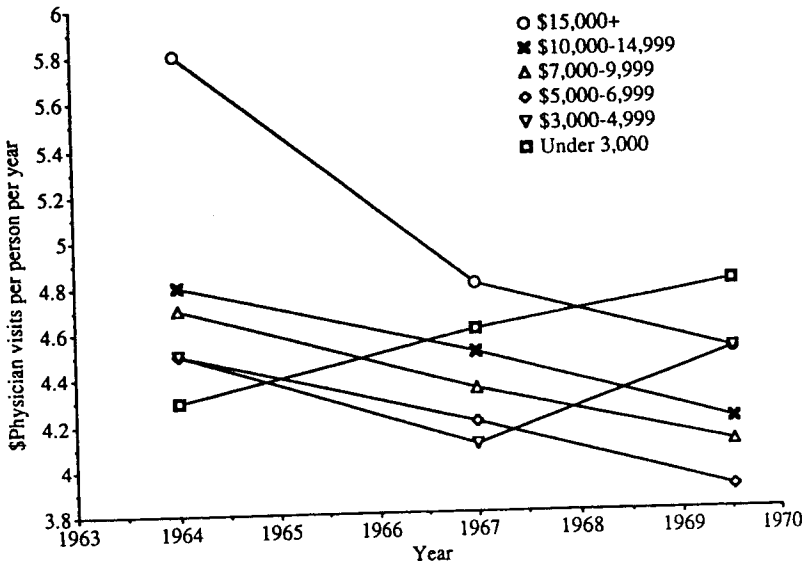

Figure 14.5. Average number of physicians per person per year for the years 1964, 1967, and 1969, by family income ranges (taken from Lohr, 1972; Wilder, 1972, p. 5, Table B).

146) and displays income group along the horizontal axis and physician visits on the vertical as shown earlier in Figure 14.1. One problem in analyzing these data concerns the metric for the pretest. We know that income distributions tend to be nonnormal and as a consequence we may need to transform the pretest variable before conducting the analysis. We also see that the highest income group has no upper-income limit given. To analyze these data, we decided to use the logarithm of the upper and lower limit for each pretest income interval (with an upper limit for the high pretest group set arbitrarily at $50,000 and the lower limit for the low income group set to $1000) and then use the midpoint between these logs as the pretest value for each group. The transformed data are graphed in Figure 14.6. The ANCOVA estimate of effect (in log pretest units) is $\beta_2 = .824$ ($t = 21.03$, $p = .0002$).

A second RPDD can be constructed from these same data by graphing posttest (1967) physician visits against pretest (1964) ones for each income group as shown earlier in Figure 14.3. The lowest income group is by definition

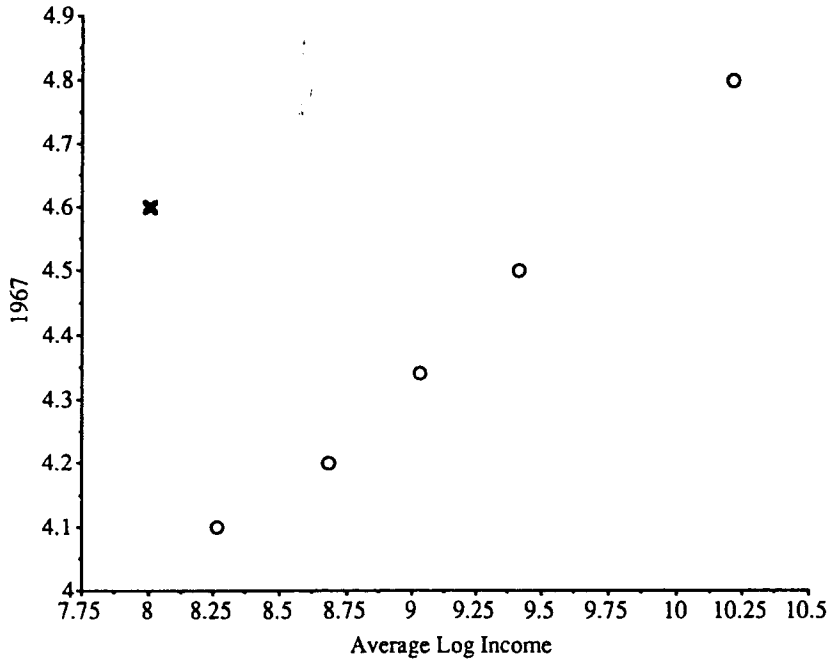

Figure 14.6. RPDD for the study of the effects of Medicaid on physician visit rates with logs of income as pretest.

the Medicaid treatment group indicated by an x on the graph. The other income groups are shown with an o and can be considered comparison or control groups. The Medicaid group had the lowest pretest average number of physician visits. The question is whether their posttest level is significantly higher than would be predicted given the control group pre–post levels. The ANCOVA estimate of effect is $\beta_2 = .479$ ($t = 3.63$, $p = .036$).

Medicaid appears to be associated with a significant rise in annual physician visits, but one still cannot conclude that Medicaid is what caused this rise. In order to reach this conclusion, one has to rule out any plausible alternative causal explanations for the observed effect (Cook & Campbell, 1979). Several possibilities suggest themselves. First, it could be that the regression line that is fitted to the data does not accurately reflect the true regression for the population in question. The question is whether the apparent significant effect results

from specification of the wrong regression model. This could arise in some contexts because the control groups do not represent the population of interest, an unlikely event in this case because the control group means include the entire U.S. population income ranges (although the use of a single group to indicate the annual physician visits for all persons with incomes of more than $15,000 may very well distort the shape of the graph). The more plausible problem is that the control group pre–post relationship is not linear in the population, but is instead quadratic or some other functional form. Following Darlington (1990, p. 295), the polynomial regression (including both the linear quadratic terms in the regression model) is fitted to the data. The resulting equation is

$$Y_i = -13.56 + 6.6X_i + -.59Y_i^2$$

Neither of the X-coefficients is statistically significant (at $p < .05$), a result that is probably attributable to the small number of points and the consequent low statistical power associated with each estimate (note, however, that the linear term alone is significant in the original ANCOVA model). This linear plus quadratic regression line is shown in Figure 14.7.

The polynomial regression fits the control group points better than the linear one did. However, it is impossible to know whether it is the better model for the population with so few points available for prediction (remember, one could fit the observed points perfectly with a fourth-order polynomial model). In judging plausibility one must examine the assumption that the straight-line fit is misleading in that a curvilinear plot would reduce the departure from expectancy. In this example, even if the linear model is not the best, it is clear from visual inspection that no reasonable model would rule out the sensibility of concluding that there is a significant effect for the Medicaid group. The linear fit reduces the $Y_0 - \hat{Y}_0$ discrepancy from that which any reasonable curvilinear fit would produce. With the polynomial model, the observed treatment group posttest will be even further from the predicted value. We feel that in this case, the linear fit is conservative, rather than misleading. A curvilinear fit should of course be the privileged fit in cases in which it reduces the apparent effect. Note that these problems are minimized in cases in which the experimental unit falls in the middle of the controls in as much as the $Y_0 - \hat{Y}_0$ discrepancy will vary little as a function of the curvilinear plot chosen.

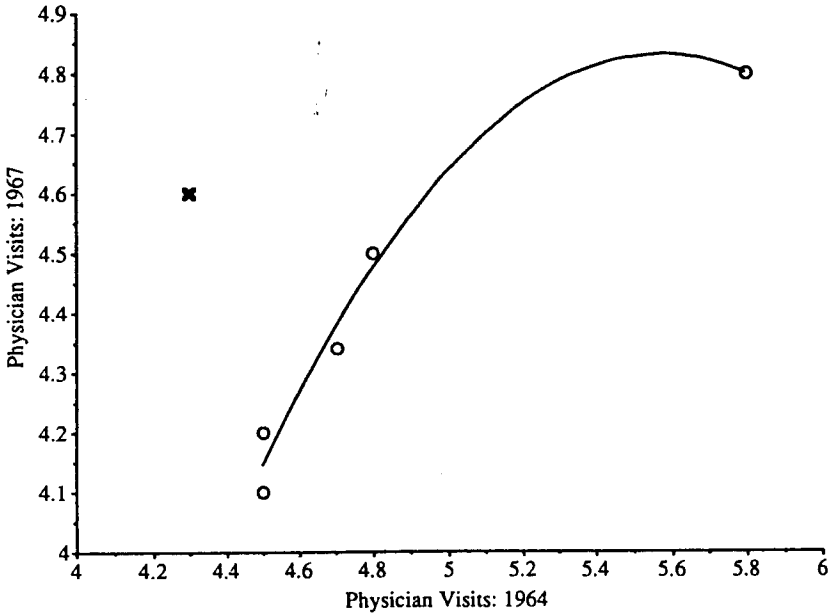

Figure 14.7. Second-order polynomial regression line for the RPDD data from the study of the effects of Medicaid on physician visit rates.

A second threat to the causal inference is an issue of internal validity. There may have been some other factor affecting only the lowest income group that increased their annual physician visit rate. For instance, if there is another low income subsidy program, such as the WIC nutritional supplement program, it may be that the rise in physician visits is attributable to the increased gynecological care subsidized by that program rather than to the Medicaid subsidies. There is no way to rule out that threat with these data, although one could examine it by comparing physician visit rates for WIC versus non-WIC recipients if such data were available. The problem in this context is that there are likely to be many federal or state programs that are targeted to the lowest income group (i.e., those who fall below the official poverty line). Although we would usually argue that an RPDD that gives the treatment to an extreme case is preferable to uncontrolled treatment assignment, it is not preferred when the implicit cutoff is a well-publicized, frequently used value such as the federal

poverty level, which is used as the criterion for assignment in many national programs that constitute alternative potential causes for any observed treatment effect. Thus, in this case, it is impossible to be confident that the observed effect is a result of Medicaid alone. Nor does the apparent jump in physician visits for the next highest income group from 1967 to 1970 (evident in Figure 14.5) solve the problem because it too may very well result from raises in the Medicaid eligibility cutoff values, raises in the official poverty income level (and the consequent eligibility for other federal programs), or both.

SCHIZOPHRENIC REACTION TIME STUDY

Fleiss and Tanur (1972) described a version of a Regression Point Displacement analysis that explored reaction time in schizophrenics. The purpose of this analysis was to examine whether clear-cut schizophrenics differ from other groups in their cross-modal–ipsimodal reaction time difference. The data are shown in Figure 14.2. The ANCOVA estimate of effect is $\beta_2 = 16.11$ ($t = 109.59$, $p = .0058$).

Many readers will join us in being surprised that such power can be obtained from an application with just three control group points. There are several variables affecting such a p-value. One is the degrees of freedom (in this case, $N - 2 = 1$ df). A second is the dispersal of values in the control groups from the fitted line. A third is the magnitude of the departure of the treated group from the center of the fitted line (i.e., the error term is larger the greater the distance of the treatment group pretest mean from the mean of the control group average).

Let us consider the second component. In this case, the linear fit is essentially perfect, producing a very small error term, and hence a very large t-value. But with only three points, are not such perfect fits going to happen by chance very frequently? Or, to put it another way, in repeated samplings from the same universe, is not the error term going to fluctuate widely from replication to replication? Can we be sure that the small-sample values for the t-test are still appropriate for this application, for $df = 1$? Note that Fleiss and Tanur failed to find a significant effect when using within-group variance, in contrast to the Figure 14.2 analysis using only group means. We would argue that the

relative power of an RPDD *can never be greater* than the power of a more microlevel analysis (e.g., using individual data points instead of group means) on which it is based, even though we may serendipitously find extremely significant estimates in a given RPDD, as in this example. Thus, although our presentation has been inspired in part by Fleiss and Tanur's (1972) seminal paper, we regard the specific application with caution just because the *perfect* linear fit is so out of line with ordinary experience. Someone interested in applying their finding would be well advised to explore the relationship between the ipsimodal versus cross-modal–ipsimodal differences over a larger number of diagnostic groupings in an effort to get a more plausible error term.

REQUIREMENTS FOR THE RPDD

To qualify as an RPDD, there must be multiple comparison or control groups and pre–post measurement. But given this restriction, there are many alternative versions of the design that are possible. Many of the variations can be described in terms of five major dimensions, where a different RPDD can be constructed for different combinations of these dimensions. The dimensions follow.

1. Method of Assignment of Treated Unit

For almost any two-group pre–post design it is possible to construct a RPDD analogue. If the single experimental unit is randomly assigned (from a pool of potential candidate units), the RPDD is analogous to a Randomized Experimental (RE) design. When the experimental unit is chosen because it has the most extreme value (high or low) on the pretest, this is essentially equivalent to assignment by a cutoff (the cutoff in this case is usually implicit and consists of the pretest values that distinguish this extreme case from the others), making the RPDD analogous to the regression–discontinuity (RD) design. Finally, when the treatment group is chosen arbitrarily, for political reasons or personal favoritism, or for any other unspecified reason, we can consider the assignment to be by an unknown or unspecifiable rule, and the RPDD is most analogous to a Nonequivalent Group Design (NEGD). For most of the experimental or quasi-experimental designs, it is possible to construct RPDD analogues, and

useful to do so because consideration of analogous designs and the literatures that have grown around them will help raise validity issues that ought to be considered in the RPDD analogue.

2. The Unit of Measurement

The unit of measurement refers to the entity represented in each pre–post point in a RPDD. Usually, these will be broad units—states, cities, communities, socioeconomic groups, diagnostic groups—not individual persons. However, an RPDD design can be constructed using either aggregated individual data or group data. For instance, many readily available databases consist of already aggregated frequencies, rates, proportions, or averages across geographically or demographically defined groups. In the Medicaid example of Figures 14.1 and 14.3, the average number of physician visits per year per person for six different income groups is used. The RPDD analysis does not require physician visit rates by individual (nor changes in such rates)—it operates in this case on the group averages. Restricting the data analyzed to repeated measures from the same individuals adds power to some statistical analyses. If the group means in each case are based on the same persons in each year, we might expect a smaller error term (see Cook & Campbell, 1979, pp. 115–117). Yet in most instances, if a survey is conducted in a number of communities before and after an educational intervention in one of the communities, the people measured on the pretreatment survey are not likely to be the same as those measured afterward (unmatched). The RPDD is perhaps the strongest design available for studying community-level interventions in which different persons are sampled on each occasion. It is also possible to use the RPDD when the pretest and posttest scores are based on the same person. For example, the same (matched) patients could be measured before and after, but for the RPDD, all such patients at a given site (or clinic, hospital) would constitute a unit and their average scores used. This pre–post same-individual RPDD might arise in the case in which the same students within a school are measured before and after some treatment is implemented in a single classroom. Classroom average scores could be used in conducting a RPDD analysis, or individual scores (grouping all control group cases together) could be used in a more traditional nonequivalent group analysis. One must be careful in using data from separate pre- and posttest

samples because of the potential bias that can arise. For instance, if pretest and posttest average scores for a group are used, it is likely that those nondropouts present on the posttest are unrepresentative of the original pretest group. Contrast this with two different but repeated random samplings over time from the same community. Although even this may be biased in the sense that the basic demographic structure of the community may have evolved between measurements, it is much less plausible in this case, especially when the time span is relatively short. Random, rather than opportunistic, sampling on each occasion can help to ensure some degree of equivalence when repeated measures are not obtained for the same group of people.

3. The Number of Treatment Groups

In the simplest case, the RPDD involves only a single treated group or site. The treatment would be administered in one community or classroom. When the design involves enough treated points that it is reasonable to estimate the same functional form as for the controls, the RPDD essentially transforms into one of the other pre–post designs—randomized experiment, nonequivalent group design, or regression-discontinuity—depending on the method used to assign units to treatment or control condition.

4. The Same Versus Different Pre–Post Measures

In general, the same variable is measured before and after the treatment (matched measures), but different measures can be used. For instance, if one is looking at the effect of an educational program, it might not make sense to measure content-related performance on the pretest because students would not be expected to know any of the content (and may not even understand the questions). One might use a general measure of prior intelligence or academic achievement (GPA, standardized achievement test scores) as the premeasure, with a treatment content-specific outcome measure. Thus, pretest and posttest in this case are different, or unmatched, measures. In Figure 14.1, the pre- and posttest measures are still less similar. Whether matched or not, statistical power is likely to be greater when the premeasure has a strong linear or monotonic relation with the outcome variable.

5. The Number of Covariates

In the simplest case, the RPDD uses a single pretreatment variable. But it is also possible to use multiple pretreatment variables that can simultaneously be entered as covariates in the model. The major problem with multiple pretreatment covariates is that, because each covariate costs one degree of freedom, using multiple covariates requires more control groups. However, the use of multiple covariates when many control groups exist will be an important mechanism for improving the statistical power and efficiency of the treatment effect estimate.

THREATS TO VALIDITY

Selection Bias

Probably the most important threat to validity in the RPDD is the potential for selection bias that stems from initial between-group differences that affect the posttest and are unrelated to the treatment. The plausibility of such a threat rests on the method used for assigning (or selecting) the treated group in the RPDD. The method of assignment determines which traditional multiple unit pre–post designs the specific RPDD is most like. If the RPDD units are randomly assigned, the design is most analogous to a RE. In the RPDD case, however, random assignment is not used to ensure probabilistic pretest equivalence as much as to minimize the chance that a unit might be opportunistically chosen because it is well-suited, politically favored, likely to be successful, or any other number of factors that could bring about an apparent effect even if the program is never administered. Random assignment helps to guard against the many pretreatment correlates that might bias the outcome, wittingly or unwittingly.

If the RPDD experimental unit is assigned solely on the basis of its extremity on the pretreatment measure, this is analogous to a RD design because the assignment is by means of an implicit *cutoff rule*. This is the case with the Medicaid study as described in Figure 14.1. There, Congress allocated Medicaid explicitly using an income cutoff rule. Note, however, that this is not the case for other RPDDs constructed from the Medicaid data (shown in Figure 14.3) where, for instance, 1964 average physician visits constitute the pretest

and 1965 values the posttest. Even though in this case the experimental group also turns out to be the lowest in pretest average physician visits, *physician visits were not the basis for the allocation of the Medicaid treatment.*

Where the RPDD is structured like the RE or the RD designs, the selection bias problem is largely mitigated by the fact that we know perfectly the rule that determines the assignment to treatment (probabilistic in the RE case and cutoff-based in RD). Just as in those designs, only a factor that correlates perfectly with the known assignment rule poses a legitimate selectivity threat. Of course, as the Medicaid study shows, there can be many such factors because the same implicit cutoff (i.e., the poverty rate) is used to allocate multiple programs.

This should be contrasted with the third assignment strategy—uncontrolled assignment—that yields an RPDD most analogous to the NEGD. In this case, the rule for assignment (i.e., selection) of the experimental unit is not explicit or able to be controlled for perfectly in the statistical model. As a consequence, one is less sure that the observed treatment effect is attributable to the treatment as opposed to any of the many possible selection factors that might also affect the posttest. For instance, assume a study in which there are ten possible treatment sites for some presumably beneficial treatment. Further assume that the selection of the experimental site is intensely political with each site lobbying to be the first to receive the experimental. The city that is ultimately selected is likely to differ in many ways from those that were not selected. It may be more highly motivated, have greater resources, have more political clout, and so on. If these (and other) factors can affect posttest scores, it will not be possible to say with great confidence whether any observed treatment effect is a result of the treatment or of these inherent differences between this city and the controls. Measuring all cities, including the experimental, on a pretest is likely to improve our inferential ability because posttest differences can be adjusted for pretest ones, but our experience with such adjustments for selection bias warns us that we should be cautious about attaching too much credibility to treatment-effect inferences in this case.

In an ideal situation, the demonstration site for a pilot program in a RPDD would be chosen purely at random, perhaps in a public lottery. Although with an N of 1 in the experimental group one would not be getting the benefit of

plausible pretreatment equalization, one would be reducing the plausibility that a systematic difference on other variables not only determined the choice of the pilot site, but also determined the exceptional departure from expectancy on the outcome variable.

The discussion so far has assumed that the experimental unit was *not* selected *after* its eccentricity on the outcome variable was known. Nonetheless, that possibility needs discussion. Consider a case in which ten cities have measures on HIV-positive rates over successive years, and one notices that one of them is exceptionally far below expectancy for the second year. The interpretive problem is that each city has had AIDS prevention programs, all slightly different, so that there is an "experimental program" to be credited with the effect, no matter which city is exceptional, even if that exceptionality is a result of chance.

Although such interpretive opportunism is to be discouraged (especially if it is disguised from the reader), the strategy of locating a "truly exceptional" city (or site) first on the basis of posttest scores, and then speculating on what "caused it" should not be prohibited entirely. But in this case, the ordinary p-value for a given t-value cannot be used. Instead, a correction on the order of that for "error-rate experimentwise" is needed. The simplest approach would be to use a Bonferoni correction of the p-values (Darlington, 1990, pp. 250–257; Dunn, 1961; Ryan, 1959, 1960). If for a specified in-advance site, for a given df (e.g., $df = 8$ for the ten cities assuming a linear fit) a t-value of 2.306 is required for $p < .05$, when we want a comparable p-value (i.e., 1/20 for testing the exceptionality of any one of the ten points from the regression line determined by the nine others (not specifying which one in advance), we need a t-ratio corresponding to $1/(20 \times 10)$ or 1/200, or $p < .005$.

The RPDD, unlike other quasi-experiments, does not require pretest equivalence between the treated group and the controls. The treated group theoretically could come from anywhere along the pretest continuum. The design rests on the assumption that the treated group posttest mean does not differ significantly from the regression line prediction. As a consequence, the traditional concerns about selection bias take on a slightly different form in this context. Here, the key issues are whether the control groups yield an unbiased estimate of the true population regression line and whether the

treatment unit is a member of the control group population. This could be assured by randomly sampling control groups from the population, a circumstance that will not be feasible in many situations. If the sample of control groups is not representative of the theoretical population or the regression line is incorrectly estimated, the estimate of the treatment effect will be biased. There is no solution to this problem, although it might best be minimized by selecting many control groups with wide pretest variability. For instance, in their study of schizophrenics, Fleiss and Tanur (1972) gave this advice for selecting control groups:

> A more efficient approach would call for the identification of many of the factors that distinguish schizophrenics from normals: having a mental disorder, being hospitalized, having been treated with drugs some time in the past, and so on. Samples of subjects from groups defined in terms of various combinations of such factors would be drawn and studied. These samples would have one feature in common: they would all consist of subjects who are not schizophrenic. (p. 525)

Measurement Error and Regression Artifacts

The general problem of regression artifacts (or error in independent variables) is taken care of in the RE or RD analogues of the RPDD, because such regression is displayed and accounted for by the inclusion of X in the regression analysis (Cappelleri, Trochim, Stanley, & Reichardt, 1991; Trochim, 1984; Trochim, Cappelleri, & Reichardt, 1991). Nonetheless, when choosing the experimental unit involves unknown but systematic variables on which the experimental group differs from the control group in ways that would affect the posttest differentially from the pretest, one might mistakenly conclude that the treatment was effective when, in fact, the apparent effect should be attributed to measurement error and the resulting regression to the mean.

This is similar to the misleading interpretations that can occur in relation to the "fuzzy" RD design (see chapter 13, this volume). If in fact, the choice of the experimental unit had been based on a latent decision variable related to the pretest by the addition of a pretest random error component, and to the posttest by the addition of a posttest random error component, and if the award of the experiment is based on extremity on the latent "true score," then a mistaken inference comparable to that graphed in Figure 13.3 is possible.

The problem is a manifestation of the familiar effect of pretest measurement error in NEGDs as described by Reichardt in Cook and Campbell (1979, Fig. 4.4, p. 161). Reichardt showed that pretest measurement error attenuates the within-group slope. In the RPDD, however, because there is only a single experimental group point, it is impossible to estimate a treatment group slope (and thus, the "slope" cannot be attenuated because of pretest measurement error). However, the true control group regression line would appear to be rotated clockwise slightly (assuming a positive relationship). The further the experimental group is away from the control group pretest mean, the greater will be the deleterious effects of such measurement error—bias will be greater.

The issue of how the treated unit is chosen is so central to the interpretability of the RPDD that it warrants some belaboring. The central concern is this: Can the treated unit be considered a member of the control unit population prior to treatment, or does it come from some different distribution? If we select the treated unit randomly or because of its extremity (implicit cutoff), it is reasonable to infer that the unit is sampled from the control group population. Here, just as in the RE or RD designs, estimates of treatment effect will be unbiased by the measurement error. The regression of the posttest onto pretest accurately describes the amount of regression to the mean expected for all units, treated and control. But when selection of the treated group is not controlled, it is plausible that the treated unit does not come from the control unit population, but rather from some population that differs systematically from the controls. In this case, we must assume that the two populations may differ in their overall pretest averages and, as a consequence, measurement error would affect the populations differently and there would be regression to different population means, just as in the NEGD.

In any regression analysis, random measurement error on the pretest will attenuate pre–post regression line slopes. This is not likely to be a serious problem in the RPDD, because presumably the group means are less influenced by random error than individual data. Nevertheless, this needs further investigating. It is likely, for instance, that such an investigation would lead to the conclusion that random-measurement error introduces greater bias in treatment effect estimates when the treatment group pretest mean is located further away from the overall pretest mean, where the attenuation most affects point predictions. Traditional adjustments for random measurement error have to

be modified for the RPDD and may be problematic, especially when there are relatively few control points for estimating reliability.

However, we can state *unequivocally* that the deleterious effects of measurement error will be less manifest in the RPDD than in an individual-level NEGD analysis of the same data because the average values used in the RPDD must (by definition) have less variability or error than the individual data on which they are based. We expect that the power and efficiency of the NEGD will be the upper limit for a comparable RPDD (because the loss of degrees of freedom will outweigh the gains in reliability), but that measurement error will be reduced and the bias in estimates because of it will be correspondingly less in the RPDD.

Another regression artifact comes where the choice of the experimental unit is triggered by the error component in the pretest. During 1956, Connecticut endured an extreme crackdown on speeding and subsequently claimed a dramatic reduction in fatalities. But we know that the 1954–1955 increase in Connecticut's traffic fatalities was the largest in its history and that the 1954–1955 increase caused Governor Ribicoff to initiate the crackdown. Campbell and Ross (1968) concluded that the purported effects were merely a return to trend, a regression artifact. They also present the effect in the context of other nearby states, as in Figure 14.8.

Were one to use the 1955 and 1956 data as an RPDD, one potentially could get a significant pseudo-effect, as plotted in Figure 14.9. In this case, there are too few control states to produce significance, but the danger is illustrated (the actual t-value is -1.50, $p = .26$).

Instrumentation

As for any quasi-experimental design, the range of rival hypotheses should be examined in case a significant effect is found. For pilot studies and demonstration sites, one of the most frequently troubling will be that the program effort has changed uniquely the measurement process between the pretest and posttest for the experimental unit. The pressure on the law enforcement system to show a good effect may lead, for example, to the downgrading of felonies to misdemeanors (e.g., Seidman & Couzens, 1974). Equally frequently, program attention to a problem such as child abuse may lead to increased thoroughness

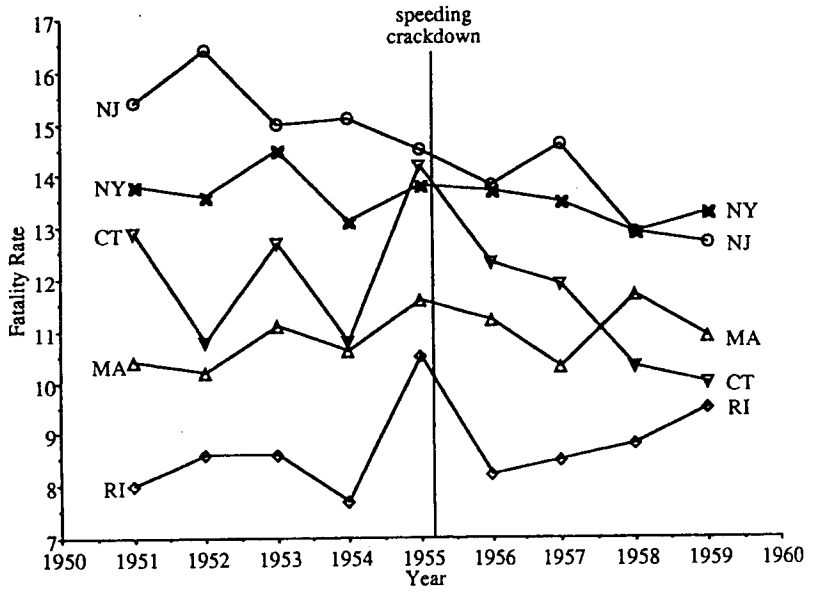

Figure 14.8. Traffic fatalities for five Northeastern states: 1951–1959.

of reporting, and a pseudo-increase (a pseudo-harmful effect) specific to the experimental unit.

Statistical Power

Although at first glance it appears that the RPDD design suffers from low statistical power because of the relatively few pre–post points that are typically used, group means are generally more stable and precise than within-group data. Fleiss and Tanur (1972) compared a traditional pre–post ANCOVA with the RPDD analysis using the same data and found that the ANCOVA results were not significant, whereas the RPDD results were. They commented,

> The difference between the analysis of covariance performed at the beginning of this chapter, where significance was not found, and the regression analysis just performed, where significance was found, is that predictability in the former was determined by covariation within groups, whereas predictability in the latter was determined by covariation between groups. (p. 525)

Two major issues related to statistical power need to be investigated. First, what is the power of the RPDD design as it stands? This should be relatively simple to determine and would make it possible to estimate the needed number of control points given some initial estimates of probable treatment effect size and desired level. An important factor in statistical power is where the treatment group scores on the pretest continuum. Statistical power will decline as the treated group pretest occurs further from the overall pretest mean. Second, an analysis needs to be done of the power of the RPDD relative to within-group ANCOVA alternatives. This should reveal whether one would ever gain statistical power in the trade-off between within-group variability in the ANCOVA framework and the presumed lower variability in the between-group-oriented RPDD.

Violating Assumptions of Statistical Tests

The RPDD may also be subject to violations of the assumptions of the *t*-test that is used. Fleiss and Tanur (1972) pointed out that the analysis is technically valid only when the control groups used to estimate the regression are a random sample from a population of such groups. The population in this case would be hypothetical (there are an infinity of potential groups that could be entered as controls) and as a consequence, this assumption can technically never be met. Instead, as Fleiss and Tanur pointed out, "One must be sure to select groups defined by the presence or absence of enough factors to assure that the variability of their mean responses is high" (1972, p. 525).

One benefit that accrues in the RPDD derives from the usually higher aggregate values used for the data. For instance, when group means are used (as opposed to individual-level values), we can more reasonably expect that the statistical assumption of normally distributed variables is likely to be met. This is because of the well-known statistical property of the central limit theorem, which holds that with sufficient sample sizes, sampling distributions are normally distributed. The advantage, of course, is that one needs to worry less about this distributional assumption, which is critical to many statistical tests.

Local History

A key threat to internal validity in this design is "local history" (Cook & Campbell, 1979). Whenever the treatment group consists of persons who are

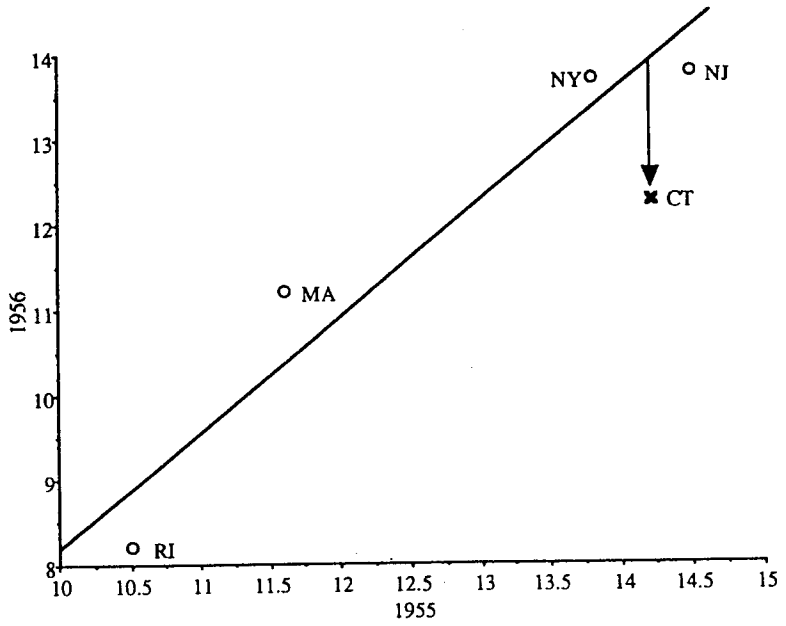

Figure 14.9. Connecticut Speeding Crackdown, RPDD.

treated together (such as at the same site) and distinct from control group persons, any factor in the setting that affects posttest performance can lead to a pseudo-effect in the data. If treated persons receive multiple treatments, or experience a markedly different setting from controls, or have a change in instrumentation (e.g., clinicians change their implicit judgement standards at the treatment site between pre- and posttest, but not at control sites), a pseudo-effect can result. These threats are not as serious an issue for the control groups because, with lots of such groups, setting variability is increased and the potential for systematic bias declines.

External Validity

Finally, the RPDD is not strong in external validity or generalizability. Although generalizing to other potential treatment groups that have the same pretest level may be reasonable, it is impossible to know whether any observed treatment effects would hold for groups with other pretest levels. Put another way, it is impossible with this design to study treatment interaction effects. If

the treatment effect changes for different pretest levels, it will not be possible to know from the RPDD. Nevertheless, if the treatment group is typical of the potential target treatment group of interest (especially if it is a unit randomly sampled from the population of interest), it will be reasonable to generalize to other similar target groups. One might not be interested in whether the treatment might work for groups having markedly different pretreatment levels (see the Fleiss and Tanur example).

CONCLUSION

The RPDD has a very specific potential range of applicability. It is limited to contexts in which one has pre- and postmeasurement and multiple control groups. It is strongest when applied to routinely collected multiwave administrative data, where it is either too costly to match individual cases or may not be possible. Because such data are widely available and cost constraints a constant factor in our society, it is likely that the RPDD would be widely applicable. In fact, there are probably many instances in which the requirements of the design have already been met and for which a post hoc analysis could be simply constructed.

The design has several important weaknesses that need to be anticipated. Where few control groups are available, one is likely to have low statistical power. It is recommended that a power analysis be routinely reported with any analysis of the RPDD that fails to show significant treatment effects. In terms of internal validity, there are several possible threats that could lead to pseudo-effects in various situations. Care needs to be taken in selecting a heterogeneous set of control groups. Whenever possible, the treated unit should be randomly selected from the population. This will tend to minimize any deliberate selection factors that might threaten internal validity and is also likely to be the variation that has the greatest statistical power. Failing that, selection of the most extreme case is preferred over arbitrary or convenience-based selection, especially when the same measure is used for before and after measurement.

The RPDD has great potential for enhancing our ability to conduct research in natural social contexts. It is relatively inexpensive to apply where appropriate administrative data exist. It is based on well-known statistical

models that can be estimated with almost any statistical computing package. It extends our ability to evaluate the effects of community-level programs where other designs are often not readily available. Although much work is yet needed to explore the implications and variations of the RPDD, it is clearly a useful addition to the methodological tool kit of the researchers of the experimenting society.

NOTE: This is an abridged version of an original article, "The Regression Point Displacement Design for Evaluating Community-Based Pilot Programs and Demonstration Projects." The complete version can be accessed through the World Wide Web at the following address: http://trochim.human.cornell.edu/research/rpd/rpd.htm