

When Second Best Is Good Enough: A Comparison Between A True Experiment and a Regression Discontinuity Quasi-Experiment*

Richard Berk
Department of Statistics
Department of Criminology
University of Pennsylvania

Geoffrey Barnes
Department of Criminology
University of Pennsylvania

Lindsay Ahlman
Ellen Kurtz
Philadelphia Adult Probation and Parole Department

July 8, 2009

Abstract

In this paper, we compare the results from a randomized clinical trial to the results from a regression discontinuity quasi-experiment when both designs are implemented in the same setting. We find the that results from the two approaches are effectively identical. We

*Geoffrey Barnes' work on this project was funded in part by a grant from the Smith Richardson Foundation to the Jerry Lee Center of Criminology at the University of Pennsylvania. For this support we are grateful. Thanks also go to Larry Sherman for suggestions that helped to shape the randomized experiment.

attribute the comparability in part to recent statistical developments that make the model required for the analysis of data from a regression discontinuity design easier to determine. These developments make an already strong quasi-experimental design even stronger.

1 Introduction

When properly implemented, randomized experiments are the surest way to obtain unbiased estimates of treatment effects. Yet, it is sometimes difficult to mount such studies, so it can be useful to have other approaches. The regression discontinuity design has long been recognized as a powerful alternative to randomized experiments (Thistlewaite and Campbell, 1960; Campbell and Stanley, 1963; Rubin, 1974; Boruch and DeGracie, 1975; Trochim, 2001), has been used successfully in a few criminal justice applications (Berk and Rauma, 1963; Berk and de Leeuw, 1999; Chen and Shapiro, 2007), and is currently enjoying new-found popularity (Imbens and Lemieux, 2008). There are good reasons. Regression discontinuity designs can obtain the unbiased treatment effect estimates of true experiments, but with some loss of statistical power. In compensation, it is often relatively easy to increase sample sizes because the design can be less intrusive. There can also be benefits for external validity when the regression discontinuity selection mechanism corresponds well to how treatment group membership is likely to be determined in practice. Its main drawback is that the analysis of regression discontinuity designs depends fundamentally on a statistical model. If that model is substantially wrong, treatment effect estimates will be biased, sometimes badly. And for any statistical model, it can be difficult in practice to determine if the requisite assumptions are effectively met.

One can never know with full confidence whether a model used to analyze a regression discontinuity design is sufficiently accurate. But there are situations in which a strong case can be made. In this paper, we consider just such a situation. We are able to directly compare the estimates from a regression discontinuity design to those of a randomized experiment, both implemented in the same setting. And in that comparison, there are lessons about when the second best design can be good enough. Related work is discussed in a recent paper by Thomas Cook and his colleagues (2008).

2 The Setting

Law enforcement agencies across the country must operate under resource constraints. Often those constraints are very severe. The Philadelphia Adult Department of Probation and Parole (APPD) is no different, and its administrators have recognized for some time that supervisory resources should be allocated more efficiently. One avenue would be to allocate supervision intensity and support services by demonstrable need.

It is well known that risks to public safety are not the same for all individuals on probation or parole. Indeed, there is at least anecdotal evidence that many of the individuals supervised are no more likely to commit crimes than members of the general population living in similar neighborhoods. At the same time, there is a relatively small number of individuals who pose a genuine threat (Berk et al., 2009). It just makes good sense, therefore, to allocate scarce supervisory resources differentially. Within a fixed budget, less should be allocated to low risk cases and more should be allocated to high risk cases.

Two questions naturally follow. First, how can one at intake determine with reasonable accuracy which cases are low risk and which cases are not? Clearly, an effective forecasting procedure is required. Estimates of probation and parole risks have long been used in criminal justice decision making (Burgess, 1928; Borden, 1928, Ohlin and Duncan, 1949; Ohlin and Lawrence, 1952; Dean and Dugan, 1968; Wilkins, 1980; Glaser, 1987; Farrington, 1987). Our forecasting procedures, described in the next section of the paper, capitalize on several very recent developments in statistics and computer science.¹

Second, if supervisory resources are to be allocated away from low risk cases, what are the consequences for public safety? In particular, do resource reductions for individuals forecasted to be low risk lead to an increase in their arrests for new crimes, compared to what would have happened under business as usual? That is, are the probation or parole failure rates higher when supervisory resources are substantially reduced? This question was addressed by the APPD by mounting a randomized experiment with a regression discontinuity design built in. It is a comparison between these two approaches that is the focus of this paper.

¹For recent reviews of current practice in parole and probation forecasting, see Gotfredson and Moriarty (2006) and Berk(2008b).

	Not Low Risk Forecasted	Low Risk Forecasted	Error
Not Low Risk	2385	971	.29
Low Risk	10144	16500	.38

Table 1: Confusion Table for Low Risk Forecasts

3 Forecasting “Low Risk”

In order to test whether one of the two supervision strategies would be better for low risk probationers and parolees, it was necessary to first determine which individuals were “low risk.” Low risk was defined as not being a significant threat to public safety and was operationalized as not being arrested and charged with any of the following crimes within two years of intake: murder, attempted murder, robbery, sex offenses, or aggravated assault.

Because the forecasting activities are not a central part of this paper, we will be brief. Random forests (Breiman, 2001; Berk 2008) was applied to a training data set of 30,000 observations. Predictors included the full range of variables available to APPD staff at intake when supervisory decisions are initially made: age, the number of prior arrests, the age at which the earliest contact with the adult courts occurred, gender, the number of prior arrests for gun-related crimes and so on. A full discussion of the methods and data can be found in Berk et al., (2009) where a very similar forecasting application is reported.

A cost ratio of 10 to 1, determined by the APPD, was applied so that incorrectly forecasting a parolee or probationer as low risk was 10 times more costly than incorrectly forecasting a parolee or probationer as high risk. Parolees or probationers who committed serious crimes while being treated as low risk cases was an outcome to be actively avoided even if that meant increasing the number of low risk individuals who were not identified as such. The statistical bar for getting into the low risk group, therefore, was set relatively high.

Table 1 is a conventional confusion table showing how accurate the forecasting was. The first two columns are counts of observations (N=30,000). The last column is the proportion of cases incorrectly forecasted. When individuals were in fact not low risk, they were incorrectly forecasted 29% of the time. When individuals were in fact low risk, they were incorrectly forecasted 38% of the time. The error rates represent real forecasting error,

not classification error in which the data used to construct the forecasting algorithm are used to also evaluate the forecasting algorithm. The level of forecasting accuracy was acceptable to the APPD.

4 The Experiment

The experiment was motivated by the desire to evaluate two ways of supervising low risk parolees and probationers, one of which required fewer resources than the other. Beginning on October 1st, 2007, 1,559 currently-serving probationers and parolees were assigned at random to one of two interventions. All of these offenders had been forecasted to be low risk, had been under supervision for at least 3 months, and had at least one month remaining in their current sentence(s). Individuals assigned at random to the less resource intensive experimental group were placed with parole/probation officers having cases loads of approximately 400 other low risk individuals. The supervision was governed by the following protocols, as described by Ahlman and Kurtz (2009: 5).

- Office Reporting: office visits were scheduled once every six months, during which the officer would review residence, employment, payments on fines/costs and restitution, and compliance with other conditions.
- Phone Reporting: phone reports were to be scheduled every six months, approximately midway between office visits.
- Drug Testing: drug testing during the scheduled office visit was to be administered only if it was required by court order. A Forensic Intensive Recovery (FIR) evaluation was to be ordered after no more than three positive urinalysis results, and low risk offenders could be sent to drug treatment at the offender's request. ²
- Missed Contacts: Arrest warrants were to be issued if there had been no case contact for six months. If the offender surrendered positively,

²FIR is basically treatment for drug dependence. Offenders are evaluated for drug problems either in prison or before sentencing, and if eligible, are sent to in-patient or out-patient drug treatment as a part of their probation/parole sentence. Offenders can also be paroled through FIR, so that if an inmate can credibly claim to have a drug problem, and if there are treatment beds available, the inmate can be paroled to an FIR program supervised by the APPD.

the warrant could be removed with no criminal sanction.³

The control intervention was “standard supervision,” nominally more intrusive and implemented with case loads of around 150. Standard supervision had the following features (Ahlman, personal communication).

- Standard supervision relied heavily on officer discretion, but general practice included office visits once a month. The supervising officer could increase reporting to as much as weekly, usually in response to positive drug tests or new arrests. Office visits could also be as infrequent as once every three months for offenders whose officers consider them (on an intuitive basis) to be low risk.
- Field visits would be rare, but in some instances, officers doing standard supervision could spend one day a month visiting offenders in their homes.
- Officers providing standard supervision would generally issue warrants for offenders who failed to report or make contact for 90 days.
- The usual array of support services would be delivered in a highly discretionary manner.

In summary, with a caseload that was more than twice the size of those in general supervision (from about 150 to about 400) offenders, the probation officers handling the experimental group were, by design, simply unable to provide as much supervision as they had prior to the study. Clearly, costs of providing supervision and services were to be dramatically less as well.

All study subjects were followed through local county criminal justice records for a full 12 months. As already noted, to be included in the study, an individual had to have at least 30 days of supervisory time remaining. If the sentence ended before the 12 month followup was finished, 12 months of followup data were collected nevertheless. The same principle applied if an individual was re-incarcerated. For these individuals “time at risk” was reduced *de facto*. However, the reduction occurred *after* the individual had already failed and is irrelevant for the analyses to follow.⁴ Finally, even

³“Positive surrenders” occur when parolees or probationers for whom an arrest warrant has been issued turn themselves in to the APPD.

⁴There could be very serious biases if one tried to control statistically for such time at risk. The most obvious problem is that time at risk can be affected by incarceration, even

though study subjects could vary in the length of time they were exposed to either the treatment or control condition, no bias was introduced as a result. Random assignment guaranteed that on the average the length of the exposure period was the same for the experimental and control groups.⁵

The key outcome variables were arrests for any new crimes, arrests for serious crimes, and arrests for crimes of violence. Serious crimes included murder, attempted murder, aggravated assault, sexual crimes, and robbery. Only the most significant forms of violent behavior were included in this definition. Simple assaults, verbal altercations, and threats were excluded as serious crimes. Violence, on the other hand, included any use or threatened use of force, no matter how slight. Although nearly all serious crimes were included in the definition of violence, not all forms of violence were considered serious.

5 Implementation

The experiment was implemented largely as planned. Perhaps the most demanding problem was maintaining as sufficient number of low risk individuals consistent with prior power calculations. A substantial number were dropped from the low risk pool before random assignment because of mandatory and competing supervisory requirements. For example, some individuals were assigned to Philadelphia’s Youth Violence Reduction Partnership (YVRP), the Sex Offender Unit, or the Mental Health Unit. Other were processed through the city’s Gun Court. Nevertheless, statistical power targets were reached with 759 low risk individuals randomly assigned the control group

if just for several days, so that failure under supervision affects time and risk. The causal direction goes the wrong way. One possible consequence is that individuals with less time at risk may be *more* likely to fail. But this would be an artifact of the incorrect causal direction. A more subtle problem is that random assignment does not justify any form of regression with covariates. If regression adjustments are introduced nevertheless, there is likely to be bias in any estimates of treatment effects and badly biases standard errors. And in the case of logistic regression, for instance, the treatment effect parameter being estimated is not what most researchers think it is (Freedman, 2006; 2008ab)

⁵There is, however, an external validity issue. It is possible that the results reported below would not fully apply to individuals who consistently had larger “doses” of the interventions. Within the dose range in this study, the results did not materially differ. But for the methodological aims of this paper, these concerns are not relevant.

and 800 low risk individuals randomly assigned to the experimental group.⁶

As anticipated, the content of the treatment and control interventions differed substantially. The control condition was far more intrusive. For example, members of the control group experienced an average of about 2.4 office contacts every three months while members of the experimental group experienced an average of about 1.0 office contacts every three months. Likewise, approximately 38% of the control group had drug tests ordered compared to about 15% of the treatment group. In short, the treatment and control interventions were delivered largely as planned so that it was appropriate to formally consider whether the less intensive and less costly approach for low risk individuals led to worse outcomes.

6 The Regression Discontinuity Design

The regression discontinuity design requires that assignment to treatment conditions is *fully* determined by a threshold on some quantitative covariate. It does not matter if that covariate is related to the response, although in practice it often is. Study subjects that fall above the threshold on the assignment covariate are assigned to one intervention, and study subjects who fall at or below the threshold on the assignment covariate are assigned to another intervention. There are generalizations to many treatments and many assignment covariates. Imbens and Lemieux (2008) provide an excellent and thorough exposition.

Within the usual parametric approach to the analysis of the regression discontinuity designs, if a regression function linking the assignment covariate to the response is the same on both sides of the threshold, a very simple form of regression analysis can produce unbiased estimates of the average treatment effect. In particular, one can use

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 f(x_i) + \varepsilon_i, \quad (1)$$

where i is the subject index, y_i is a quantitative response variable, t_i is an indicator variable denoting membership in either the treatment or comparison group, $f(x_i)$ is a function of the assignment covariate, and ε_i is a random disturbance much as in the usual linear regression model. Often the $f(x_i)$ is

⁶Each case was assigned a random number. The cases were sorted on that random number so that the order was necessarily random too. Cases were then just taken from the top of the list as needed to be in the experimental or control group.

assumed to be linear. Then, $f(x_i)$ is replaced by x_i , and equation 1 becomes the usual linear regression model.

The regression coefficient β_1 in equation 1 will capture the direction and size of any shift up or down in the $\beta_2 f(x_i)$ at the threshold value for x_i . As such, it provides an estimate of the average treatment effect. However, one must have a good approximation of the $\beta_2 f(x_i)$. If the estimated $\beta_2 f(x_i)$ is substantially incorrect, the ε_i may be correlated with t_i . Biased estimates of β_1 can follow.⁷

Equation 1 is not a causal model. Whatever the stochastic process may be that generates the observed y_i , the $\beta_2 f(x_i)$ simply describes how the response is associated the assignment variable. Why the function takes a particular form is unaddressed and is not formally relevant. There is also no need to make any causal attribution for the role of the assignment variable itself.

The treatment indicator provides the information needed to estimate a difference in level on either side of the threshold. The estimate is analogous to a comparison between the mean of the experimentals and the mean of the controls in a randomized experiment. Here too, there is no causal model. The key point is this: a good estimate of the $\beta_2 f(x_i)$ implies a sufficiently accurate empirical summary of how the response and the assignment variable are associated absent an intervention at the threshold. How to arrive at a good estimate of the $\beta_2 f(x_i)$ is a matter to which we will shortly return.

Equation 1 can be generalized in a number of ways (Imbens and Lemieux, 2008). For example, the response variable can be binary, in which case the regression can become logistic regression. There can be more than one intervention, in which case there can be more than one indicator variable for the interventions. However, the basic logic remains. Because one knows exactly how selection into treatments was accomplished, one can in principle use covariance adjustments to control of selection biases.⁸ That is, because the selection process is deterministic and built into the design, it is fully known, at least in principle.

The use of low risk forecasts to determine the experiment's subject pool provided the assignment covariate needed for a regression discontinuity design. Random forest classifies by a "vote" over a large number of classification trees. In this application, if more than 50% of the classification trees fore-

⁷It does not matter if ε_i is correlated with x_i .

⁸Imbens and Lemieux (2008) and Imbens and Kalyanaraman (2009) provide for some interesting nonparametric alternatives, but there is some debate about how useful they are (Berk, 2009), and they are unnecessary here in any case.

casted that an individual would be low risk, that was the class assigned. If 50% or fewer of the classification trees forecasted that an individual would be low risk, that individual was not classified as low risk. So, .50 was threshold on the “votes” covariate.

The randomized experiment was undertaken only with individuals who scored above the threshold of .50. Approximately half were assigned to the new, less intensive supervisory group, and the rest were assigned to the standard form of supervision. A comparison between the outcomes for these two groups would in principle lead to an unbiased estimate of any treatment effects.

But because of the way in which the forecasts were used, a second comparison was possible. One could compare the performance of the randomly assigned experimental group to *the performance of individuals who were not forecasted to be low risk and hence, were not part of the randomized experiment at all*. The key was that the practices guiding the pre-randomization exclusions from the experiment were applied in the same fashion to individuals who were not part of the experiment. For example, individuals sent to YVRP had to be identified and removed from the data. A random sample of 1000 individuals who were not low risk was drawn and subjected to much the same screening practices as those who participated in the experiment.⁹ These 1000 individuals constituted the comparison group in the regression discontinuity design. The same outcome measures for the comparison group were the same as those used for the experimental and control groups.

Because the regression discontinuity design was implemented as intended, the main vulnerability in the analysis was determining the $\beta_2 f(x_i)$. In conventional expositions of the regression discontinuity design, a convenient functional form, usually linear, is assumed. Regression diagnostics are examined and sometimes another convenient function is substituted. For example, x_i^2 with its own regression coefficient might be added to equation 1.

Despite some arguments in favor (Lee and Lemieux, 2009), a practice of limiting the $\beta_2 f(x_i)$ to a small number of convenient parametric forms can be a bad idea. The true function may not be well summarized in a convenient parametric form, and it may not be clear how one decides which parametric form is best, let alone approximately right. Searching over several parametric forms, therefore, may not be a good strategy.

⁹The screening process was complicated so there are no guarantees. But all of the major reasons for exclusions were applied to the random sample of 1000.

We turned, therefore, to the generalized additive model (GAM), whose right-hand side is the usual linear combination of regressors (Hastie and Tibshirani, 1990). However, relationships between quantitative predictors and the response may be inductively determined by smoothers. For example, with a quantitative response and three regressors, $y_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \varepsilon_i$. The formulation is much like conventional linear regression, but each regressor has its own empirically determined nonparametric function with the response. Note that there are no regression coefficients associated with each function. They are absorbed in the function itself.

The usual least squares estimation procedures are usually altered as well. For instance, a popular approach is to minimize a “penalized” error sum of squares of the form,

$$\text{RSS}(f, \lambda) = \sum_{i=1}^N [y_i - f(x_i)]^2 + \lambda \int [f''(t)]^2 dt, \quad (2)$$

where λ is a tuning parameter determining the degree of smoothing. The first term on the right-hand side is just the usual residual sum of squares. The second term introduces a cost for the complexity of the fit. The integral represents a roughness penalty, while λ determines the weight given to that penalty when the sum of squares is minimized.¹⁰

As λ increases without limit, the fitted values approach the usual least squares line. As λ decreases to zero, the fitted values approach an interpolation of the values of the response variable. Larger values of λ lead to smoother fitted values. Smaller values of λ lead to rougher fitted values. The goal is to empirically arrive at a function that is not “gratuitously” complicated. In practice, the value of λ is selected to minimize some function of forecasting error. The generalized cross-validation statistic is a popular choice. The generalized cross-validation statistic will not decline substantially if an increase in roughness is, in effect, solely a result of using up more degrees of freedom.¹¹

¹⁰The second derivatives of the function quantify how rapidly the first derivative (i.e., the slope) is changing. The more rapidly the first derivative changes, the less smooth the function. Integrating over the second derivative produces an overall measure of how smooth the function is.

¹¹If a linear function were to be the proper choice, the value of λ would become relatively large and the resulting function approximately a straight line. Likewise, if the proper function were approximately quadratic, that too would be likely to be found.

Equation 2 is easily altered so that the fitting penalty can be used with maximum likelihood estimation and, therefore, the entire generalized linear model. A formal discussion of the issues can be found in Green and Silverman (1994), and in Hastie and his colleagues (2009). The GAM implementation we used can be found in Wood (2008). Berk (2008: Chapter 2) provides a very accessible introduction to penalized fitting functions.

For a binary outcome within GAM, equation 1 can be reformulated as a generalization of logistic regression. That is,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 t_i + f(x_i), \quad (3)$$

where p_i is the probability of failure, and all else is as before. Ideally, the $f(x_i)$ should be relatively smooth. Then, an abrupt shift in level at the threshold can be easily identified and as before, β_1 provides an estimate of the average treatment effect. That effect is in a log-odds metric. One can exponentiate β_1 to obtain an estimate of the average treatment effect as an odds multiplier.¹²

Although as a practical matter, a relatively smooth $f(x_i)$ is desirable, one can never be certain if a relatively smooth $f(x_i)$ is effectively capturing how the assignment variable is associated with the response variable. Relying on a measure of forecasting accuracy, such as the generalized cross-validation statistic, can provide evidence but not proof. It can be useful, therefore, to shift where the burden of proof falls. One might ask not whether the $f(x_i)$ is reasonably smooth, but what reasons there are for believing it is not. One might also ask what reason one has to think that there would be an abrupt shift in the $f(x_i)$ exactly at the threshold, but not as a consequence of the intervention. In short, what are the substantive reasons to *not* believe results from equation 3. The answers will depend on the particulars of the application.

¹²Alternatively one can employ an approximate matching strategy using a subset of cases on either side of the threshold but relatively close to it. These cases are roughly matched on the assignment variable. If the mean (or proportion) of the cases just to the left of the threshold differs from the mean (or proportion) of cases just to the right of the threshold, there is evidence of a discontinuity at the threshold. The size of the difference can be a sensible estimate of the average treatment effect. This approach will usually discard a substantial number of observations and depend on what one assumes about the nature of the response function within the window chosen (Imbens and Lemioux, 2008; Berk, 2010). We did not use this approach because we expected the $f(x_i)$ to be substantially nonlinear within any reasonable window around the threshold.

7 Results

Table 2 show the proportion of individuals who failed by each of three outcome measures and by their study group membership. The three outcomes were (1) any criminal charge during the one year followup, (2) any serious criminal charge during the one year followup, and (3) any violence charge during the one year followup. The three study groups were (1) the predicted low risk individuals assigned at random to the experimental condition, (2) the predicted low risk individuals assigned at random to the control condition, and (3), a random sample of 1000 individuals not predicted to be low risk and who were excluded from the randomized experiment.

	Controls	Experimentals	Excluded
Any Charge	.149	.159	.268
Serious Charge	.030	.016	.065
Violence Charge	.040	.025	.093

Table 2: The Proportions of Individuals Who Failed By Each Study Group and Each of The Three Outcomes

Among the individuals forecasted to be low risk, charges for serious or violent crime were very rare. The proportions who failed were .04 or less. Charges for any criminal offense were more common at around .15, but still relatively low for big-city offenders under supervision. Taken in total, these figures are one indication that the individuals forecasted to be low risk really were.

For all three outcomes, the failure proportions for those individuals not forecasted to be low risk were substantially higher. This is what one would expect given the forecasting algorithm used. Clearly, there is strong systematic selection into the experiment so that a comparison between the experimental group and the group excluded from the study would, without selection adjustments, be badly biased in favor of the experimental intervention. This is precisely what the regression discontinuity analysis has to overcome if unbiased average treatment effect estimates are to be provided.

For each of the three outcomes, estimated differences between the randomly assigned experimentals and the randomly assigned controls are very

Smoothed Plot of Arrests by Votes

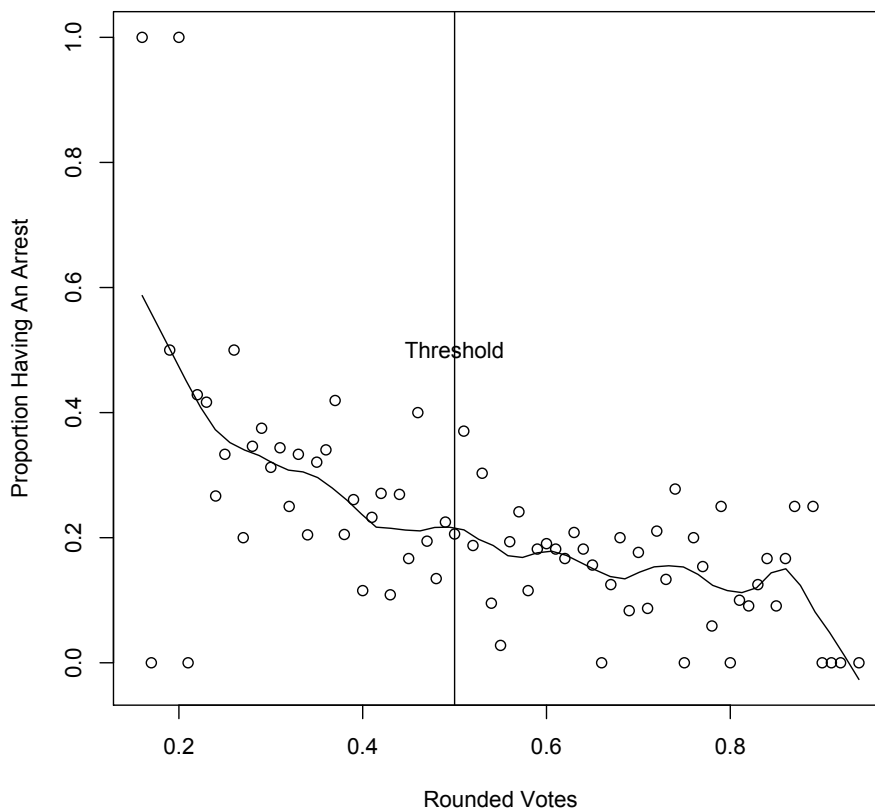


Figure 1: Vertical Slices Scatterplot

small, and one cannot reject the null hypothesis of equivalence.¹³ Can those effects be replicated in an analysis of the regression discontinuity design when there is at hand strong evidence for selection bias?

Ideally, a regression discontinuity analysis should begin with a scatter plot of the outcome against the assignment covariate. Often one can arrive at a first approximation of the proper functional form for the assignment variable,

¹³It is not at all clear how a joint test across the three outcomes can be done because the same cases are analyzed for each. But with three different outcomes, the chances of rejecting the null hypothesis when it is true are actually increased. The null findings are even more compelling than the individual p-values indicate.

spot anomalies in the data, and sometimes even find evidence for a treatment discontinuity. However, all of our outcomes are binary, so that a scatterplot of the raw data is not likely to be instructive. As an approximation, one can work with vertical slices in the scatter plot and compute the proportions failing within each slice. These proportions can then be plotted against the assignment covariate.

Figure 1 shows such a plot for a failure from any criminal charge. Proportions failing were computed within intervals of .01.¹⁴ Overlaid are a smoother and a vertical line for the threshold at .50. Four conclusions are readily apparent. First, there is a general downward trend from left to right, supporting the validity of the low risk forecasting algorithm. Individuals for whom there is more compelling evidence of low risk are less likely to fail. Second, there is no evidence of a break at the threshold as a treatment effect would require. Third, there is a bit of evidence for a nonlinear relationship, especially in the transition from low to moderate values of the assignment variable. Fourth, there is some evidence that a relatively smooth function summarizes the data well.

However, one must not make too much of Figure 1. Some information has been lost in the process by which proportions were computed, and some of the proportions were computed from very few cases. Consequently, we must return to the binary representation and undertake an analysis appropriate for categorical outcomes. The generalized additive model (GAM) was applied using a logistic link function and penalized regression splines to inductively seek the appropriate function for the assignment covariate (Hastie and Tibshirani, 1990). In effect, a generalization of logistic regression was employed. A single indicator variable was used to capture any discontinuity at the threshold.

Figure 2 shows the fitted values from the analysis plotted against the assignment covariate. This visualization conveys much of the information one would normally read from a conventional table of regression results, but is more readily assimilated. The experimental group does a little bit worse than the non-low risk comparison group, just as they did in the randomized experiment. The upward discontinuity at .50 is very small. Indeed, the treatment effect point estimates are almost identical: .01 for the true experiment and

¹⁴The raw proportions were simply rounded to two decimal places and then treated as intervals. For example, .80 implied an interval with a lower bound equal to or above .795 and an upper bound of less than .805.

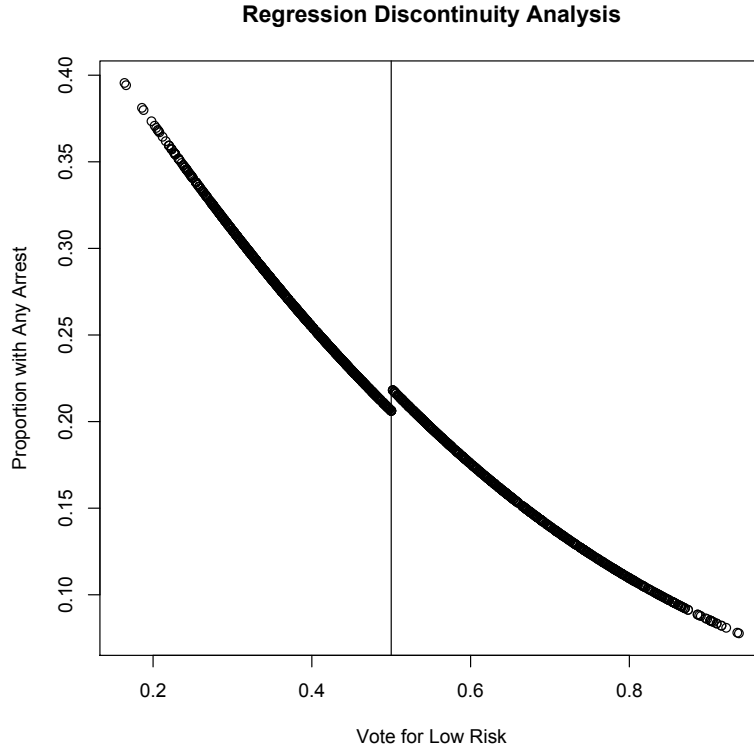


Figure 2: Any Charge: GAM Fitted Values Against Assignment Covariate

.008 for the quasi-experiment. And just as for the randomized experiment, one cannot reject the null hypothesis of equivalence in the failure proportions for the treated individuals and the comparison individuals. Note that the replication of the results from the randomized experiment is achieved even though on the average, the regression discontinuity comparison group had substantially higher failure rates. Strong selection biases were overcome.

The proportion of individuals who failed through a charge for a serious crime is very small. The strategy of computing proportions within vertical slices of a scatterplot structured like Figure 1 does not lead to a readily interpreted pattern and is not presented.¹⁵ But a smoother suggested a function much like that in Figure 1, with perhaps a more strongly nonlinear form. And there was no visible evidence of a treatment effect.

¹⁵The majority of the slices had proportions equal to 0.0.

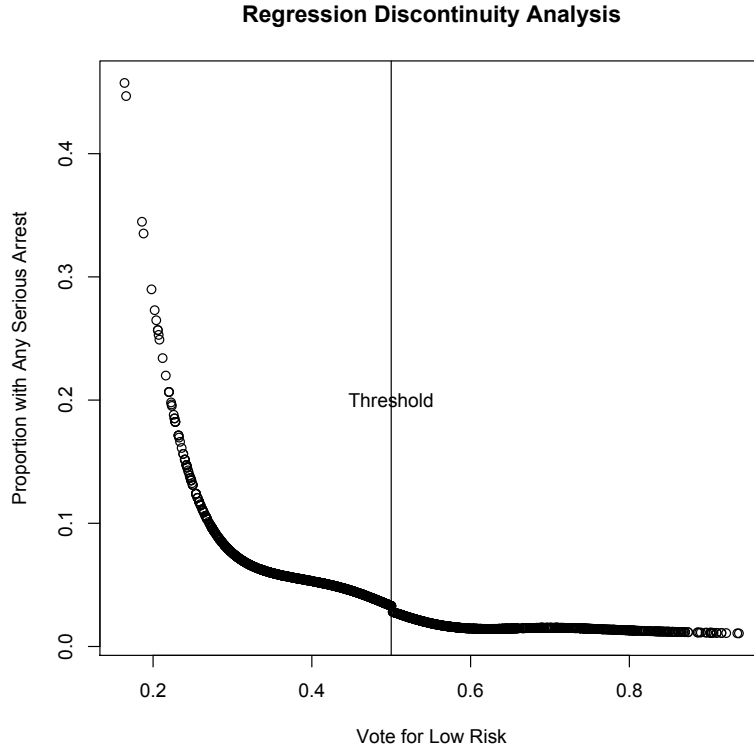


Figure 3: Serious Charge: GAM Fitted Values Against Assignment Covariate

Figure 3 shows the result of the GAM statistical analysis of the raw data (not aggregated within vertical slices). The more strongly nonlinear pattern is apparent. In the randomized experiment, the experimentals did a tiny bit better than the controls. The estimated treatment effect was .014. In the regression discontinuity analysis, the estimated treatment effect also favors the experimentals, and the effect is again tiny: .008. As before, we fail to reject the null hypothesis of equivalence, just as in the randomized experiment.

For the violent crimes outcome measure, we again applied the slicing strategy, and as before do not show the results.¹⁶ Suffice it to say, there was again evidence for a highly nonlinear relationship between the assignment variable and the response and no evidence of a treatment effect.

¹⁶Again, the majority of proportions were equal to 0.0.

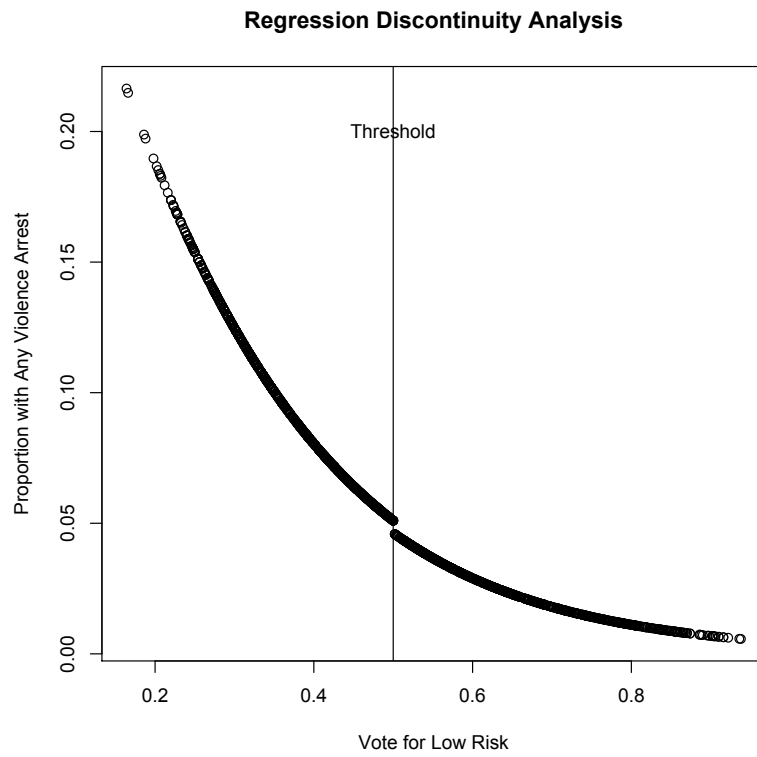


Figure 4: Violence Charge: GAM Fitted Values Against Assignment Covariate

Figure 4 presents, as before, the fitted GAM values plotted against the assignment variable. Just as for serious crimes, the experimentals did ever so slightly better than the controls. The difference in the proportion who failed was .015 (i.e., from .040 to .025), a disparity that was not remotely large enough to reject the null hypothesis of no treatment effect. The regression discontinuity analysis finding is much same. For the comparison group just to the left of the threshold, the failure proportion is .052. For the experimental group just to the right of the threshold, the failure proportion is .045. The difference is .007, and the null hypothesis of equivalence cannot be rejected.

In summary, it is clear that the randomized experiment and the regression discontinuity design arrived at virtually the same point estimates and the same overall conclusions. So, if one accepts the randomized experiment, one must accept the regression discontinuity quasi-experiment. Moreover, we can find no concrete reason based in the experiment or setting for believing that a smooth $f(x_i)$ is inappropriate or that a break at the threshold could materialize other than in response to the intervention. Recall that the assignment variable is the proportion of votes over classification trees in favor of a low risk forecast. The vote can be viewed as how reliable the forecast is. How often for a given case do the classification trees arrive at the same conclusion? There seems to be no substantive reason why the relationship between the proportion of votes in favor of a low risk forecast should be other than relatively smooth. There seems also to be no substantive reason why there should be a discontinuity at a relative vote of .50 other than as a result of the intervention.

8 Discussion

Recent work would seem to confirm that in the absence of random assignment, strategies for removing selection bias by relying on covariates are at best risky (Shadish et al., 2008). A key problem is that both the requisite covariates and the selection process into treatment groups are typically unknown. For the regression discontinuity design, however, the requisite covariates and the selection process into treatment groups are known exactly. It may not be surprising, therefore, that for the research reported in this paper, one would have arrived at the virtually the same empirical results whether one had implemented the true experiment or the regression discontinuity quasi-experiment. Cook and his colleagues (2008) reach similar conclusions.

Insofar as equation 3 accurately summarizes the data, very similar results are mathematically required.

At the same time, however, comparisons between randomized experiments and the alternatives can raise a number of subtle issues (Hill, 2008; Little et al., 2008; Rubin, 2008). To begin, it is important that the causal effects being estimated are defined in the same manner for each of the studies whose results are compared. In particular, it can matter a great deal if the treatment effect estimates are for all of the study subjects or for only a subset of them. For example, in a regression discontinuity study, one might only be interested in estimating the treatment effect for subjects near the assignment threshold, not all of the study subjects (Imbens and Lemieux, 2008). In this paper, the estimated average treatment effects are for all of the study subjects regardless of design.

It can also be important that comparisons be made “within” a given study. If a randomized experiment and regression discontinuity design are applied in different settings, with different subjects, and using interventions that are not effectively identical, it would not be surprising if the findings differed as well. That is, even when the intervention being tested is nominally the same across research designs, there can be important differences that might be mistaken for method artifacts. The results reported here are “within” study.

There remains, however, the matter of statistical power. For a given sample size, randomized experiments will often have more power than regression discontinuity quasi-experiments. In practice, however, it is difficult to know how important any difference in power really is. Power for the regression discontinuity design will depend on the correlation between the treatment indicator and the assignment variable and on the correlation between the assignment variable and the response variable. Both will depend on the unknown $f(x_i)$ and before the data are collected and the analyses undertaken, both correlations will be unknown. After the data are collected, any formal power analysis will be *post hoc*, and it is increasingly understood that *post hoc* power analyses can be very misleading (Hoenig and Heisey, 2001). As noted earlier, it is often easy to obtain very large samples in a regression discontinuity context, in which case power will not be an issue. This will usually be one’s safest strategy.

We stress that it is impossible to prove in general that any regression discontinuity summary of the $f(x_i)$ is sufficiently accurate. When there is a choice, therefore, one should favor random assignment if causal inference is

the primary goal. But sometimes random assignment is not on the table, or other research criteria such as external validity dominate the design process. This can quite properly lead to a regression discontinuity design, especially if stakeholders find the assignment mechanism congenial. For example, innovative interventions can be allocated by “need.” Then, an accurate summary of the $f(x_i)$ is front and center.

It remains an open question, at least to us, about what estimation procedures should in general be used to analyze regression discontinuity designs. We suspect that all of the nonparametric and semiparametric procedures will produce about the same results if sensibly applied. We are less sanguine about parametric regression procedures unless there is widely accepted theory and credible empirical studies specifying in advance a particular functional form.

9 Conclusions

With results that are so clear, four conclusions are easily stated. First, the downward sloping assignment functions strongly confirm the usefulness of the random forests forecasts. There is no reason why the forecasting approach taken in Philadelphia would not be appropriate for other jurisdictions, assuming that data of at least equal quality were available.

Second, there is strong evidence, at least for the APPD, that a less resource intensive form of probation and parole supervision, applied to individuals forecasted to be low risk, does not jeopardize public safety. Important savings can follow if resources are reallocated accordingly. Whether similar results would materialize in other jurisdictions is an empirical question. A lot would depend on the mix individuals being supervised and the nature of the oversight.

Third, the results of the randomized experiment and the regression discontinuity quasi-experiment are virtually identical. The point estimates are effectively the same, the test results are effectively the same, and the overall conclusions are effectively the same. The comparability of the point estimates is especially compelling because it rests on the most demanding criterion and does not depend on the null findings. In short, there is a strong case for the regression discontinuity model employed. By inference, there is a strong case for the methods used to arrive at that model. Nonparametric regression procedures, guided by good approximations of out-of-sample performance,

have promise as a way to inductively determine the relationship between the assignment covariate and response.

Fourth, it may not be especially difficult in practice to obtain the same results from regression discontinuity quasi-experiments as from randomized experiments. Our results have much in common with those reported by Cook and his colleagues (2008). But because of the reliance on a model of the assignment process for the regression discontinuity approach, this is not an argument for favoring regression discontinuity designs when randomized experiments can be undertaken. It is an argument for seriously considering a regression discontinuity design when a quasi-experiment is the only choice or when internal validity concerns do not dominate the research. Recent statistical developments have made an already strong research design even stronger.

Shortly after the outcome of the randomized experiment was known, and building on earlier forecasting results for very high risk offenders (Berk et al., 2009), a new forecasting effort was launched. Three mutually exclusive and exhaustive outcomes were forecasted simultaneously: arrests for serious crimes defined largely as above, no arrests for any crimes whatsoever, and arrests for crimes not defined as serious. Predictors were taken from the same administrative data as before. Forecasting accuracy was excellent for all three outcomes. With these results in hand, the Philadelphia Adult Department of Probation and Parole began a reorganization to move supervisory resources away from individuals who pose a smaller risk to public safety toward individuals who pose a larger risk to public safety.

References

- Ahlman, L.C. and E.M. Kurtz (2009) "The APPD Randomized Controlled Trial in Low Risk Supervision: The Effects on Low Risk Supervision on Rearrest." Philadelphia: Adult Probation and Parole Department.
- Berk, R.A. (2008a) *Statistical Learning from a Regression Perspective*. New York: Springer.
- Berk, R.A., (2008b) "Forecasting Methods in Crime and Justice." *Annual Review of Law and Social Science*, J. Hagan, K.L. Schepple, and T.R. Tyler (eds.), Palo Alto: Annual reviews.
- Berk, R.A., (2010) "Recent Perspectives on the Regression Discontinuity Design." *Handbook of Quantitative Criminology*, A. Piquero and D. Weisburd (eds.), New York: Springer, forthcoming.
- Berk, R. A., Brown, L. and Zhao, L. (2009) "Statistical Inference After Model Selection." University of Pennsylvania, Department of Statistics, Working Paper (under review).
- Berk, R.A., and Rauma, D. (1983) "Capitalizing on Nonrandom Assignment to Treatments: A Regression Discontinuity Evaluation of a Crime Control Program." *Journal of the American Statistical Association* 78(381): 21-27, 1983.
- Berk, R.A., and de Leeuw, J. (1999) "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design." *Journal of the American Statistical Association* 94(448): 1045-1052.
- Berk, R.A., Sorenson, S., and Y. He (1995) "Developing a Practical Forecasting Screener for Domestic Violence Incidents." *Evaluation Review* 29(4): 358-382.
- Berk, R.A., Kriegler, B. and J-H Baek (2006) "Forecasting Dangerous Inmate Misconduct: An Application of Ensemble Statistical Procedures." *Journal of Quantitative Criminology* 22(2): 131-145.
- Berk, R.A., Sherman, L., Barnes, G., Kurtz, E., and L. Ahlman, (2009) "Forecasting Murder within a Population of Probationers and Parolees:

- A High Stakes Application of Statistical Learning.” *Journal of the Royal Statistical Society* (Series A) 172, part 1: 191-211.
- Borden, H.G. (1928) “Factors Predicting Parole Success.” *Journal of the American Institute of Criminal Law and Criminology* 19: 328-336.
- Boruch, R.F., and J.S. DeGracie (1975) “Regression-Discontinuity Evaluation of the Mesa Reading Program: Background and Technical Report.” Evanston, Illinois, Northwestern University, NIE Project on Secondary Analysis, working paper.
- Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5-32.
- Burgess, E.W. (1928) “Factors Determining Success or Failure on Parole.” In A.A. Bruce, A.J. Harno, E.W. Burgess and J. Landesco (eds.) *The Working of the Indeterminant Sentence Law and the Parole System in Illinois*. Springfield, Illinois, State Board of Parole: 205-249.
- Campbell, D.T., and J.C. Stanley (1963) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Chen, M.K. and Shapiro, J.M. (2007) “Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-based Approach.” *American Law and Economics Review* 9(1): 1-29.
- Cook, T.D., Shadish, W.R., and V.C. Wong (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* 727(4): 724-750.
- Dean, C.W. and T.J. Dugan (1968) “Problems in Parole Prediction: An Historical Analysis.” *Social Problems* 15: 450-459.
- Farrington, D.P. (1987) “Predicting Individual Crime Rates.” In D. M. Gottfredson and M. Tonry (eds.), *Prediction and Classification*. Chicago: University of Chicago Press.
- Freedman, D.A. (2006) “Statistical Models for Causation: What Inferential Leverage Do They Provide?” *Evaluation Review* 30: 691713.
- Freedman, D.A. (2008a) “On Regression Adjustments to Experimental Data.” *Advances in Applied Mathematics* 40:180193.

- Freedman, D.A. (2008b) "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23:237-249.
- Glaser, D. (1987) "Classification for Risk," in D. M. Gottfredson and M. Tonry (eds.) *Prediction and Classification*, Chicago, University of Chicago Press.
- Gottfredson, S.D. and L.J. Moriarty (2006) "Statistical Risk Assessment: Old Problems and New Applications." *Crime & Delinquency* 52(1): 178-200.
- Green, P.J. and B.W. Silverman (1994) *Nonparametric regression and Generalized Linear Models*. New York: Chapman & Hall.
- Hastie, T.J., and Tibshirani (1990) *Generalized Additive Models*. New York: Chapman and Hall.
- Hastie, T., Tibshirani, R. and J. Friedman (2009) *The Elements of Statistical Learning*, Second Edition. New York: Springer.
- Hill, J. (2008) "Comment." *Journal of the American Statistical Association* 103(484): 1346-1350.
- Hoening, J.M. and D.M. Heisey (2001) "The Abuse of Power: The Pervasive Fallacy of Power Calculation for Data Analysis." *The American Statistician* 55: 19-24.
- Imbens, G., and Kalyanaraman, K. (2009) "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." Harvard University, Department of Economics, Working Paper.
- Imbens, G., and Lemieux, T. (2008) "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142: 611-614.
- Lee, D.S, and Lemieux, T. (2009) "Regression Discontinuity Designs in Economics." National Bureau of Economic Research: working paper #14723.
- Little, R.J., Long, Q, and X. Lin (2008) "Comment." *Journal of the American Statistical Association* 103(484): 1344-1346.

- Ohlin, L.E. and O.D. Duncan (1949) "The Efficiency of Prediction in Criminology." *American Journal of Sociology* 54: 441-452.
- Ohlin, L.E. and R.A. Lawrence (1952) "A Comparison of Alternative Methods of Parole Prediction." *American Sociological Review* 17: 268-274.
- Rubin, D. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688-701.
- Rubin, D. (2008) "Comment: The Design and Analysis of Gold Standard Randomized Experiments." *Journal of the American Statistical Association* 103(484): 1350-1353.
- Shadish, W.R., Cook, T.D., and D.T. Campbell (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin.
- Shadish, W.R., Clark, M.H. and P.M. Steiner (2008) "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignment." *Journal of the American Statistical Association* 103(484): 1334-1356.
- Thistlewaite, D.L., and Campbell, D.T. (1960) "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Design." *Journal of Educational Psychology* 51: 309-317.
- Trochim, W.M.K. (2001) "Regression Discontinuity Design," in N.J. Smelser and P.B. Bates (Eds.) *International Encyclopedia of the Social and Behavioral Sciences*, volume 19: 12940-12945.
- Wilkins, L.T. (1980) "Problems with Existing Prediction Studies and Future Research Needs." *The Journal of Criminal Law and Criminology* 71: 98-101.
- Wood, S.N. (2008) "Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models." *Journal of the Royal Statistical Society, Series B*, 70(3):495-518.