Knowing When to Fold 'Em: An Essay on Evaluating the Impact of *CEASEFIRE, COMSTAT, and EXILE*

Richard A. Berk

Department of Statistics
UCLA

July 19, 2005

1 Introduction

Professor Rosenfeld and his colleagues are to be congratulated for their courage. They have tackled a very controversial and visible set of issues addressed previously by a number of notable criminologists. They have also chosen to enter the debates with an analysis based solely on observational data. Observational data pose daunting problems when used to draw causal inferences. At a time when these problems have been thoroughly aired in the statistical literature (e.g., Freedman, 1985; 1987; 1991; Rubin, 1986; Rosenbaum, 2002; Berk, 2003) and candidly addressed by world-class econometricians (Manski, 1994; Heckman, 2000), the authors of this paper proceed nevertheless.

One has to wonder, however, whether in this case discretion would have been the better part of valor. Some empirical problems are just too difficult because of the data likely to be available and the existing methods of analysis. Trying to tease out the impact of the crime control interventions addressed in this paper may be one example.

Over the next several pages I will try to elaborate on this point. In so doing, I will touch briefly on several statistical concerns about the analyses reported, which may be useful to note, but are of secondary importance. The goal is to focus on some larger issues.

Although the paper by Rosenfeld and his colleagues is lengthly and complicated, there are surely important details not included. Page constraints being what they are, there is always lots that goes unsaid. I am operating under the same constraints. I provide below a large number of references to help fill the gaps.

2 Response Schedules and Causal Effects

It will prove useful to start with some fundamentals. In a very important and recent paper, David Freedman lays out in a deceptively simple manner what is required for causal effects to be defined (Freedman, 2004). Absent a clear definition, there is no way to determine what a causal model is meant to accomplish. The result can be pages of computer printout with no manifest purpose.

Freedman's formulation integrates a great deal of earlier work on causal effects (Neyman, 1923; Holland, 1986; Rubin, 1986, 1990). It also employs a framework criminal justice researchers will likely find accessible. We consider first a single response variable and a single intervention. For a *given* observation i, the response $y_{i,x}$ is taken to be a function of x, and of a perturbation $\delta_{i,x}$. Freedman's response schedule then takes the form of

$$y_{i,x} = \alpha_0 + \alpha_1 x + \delta_{i,x}. \tag{1}$$

Equation 1 can be used to define how an intervention represented by x is causally related to a response. For now, the response is taken to be a linear function of the intervention value and a perturbation.[1] But there's lots more. To make the discussion concrete, suppose that the intervention is the number of patrol officers assigned to "order-maintenance" policing, and the response is the homicide rate (i.e., homicides per 100,000).

[1] For the points to be made, one could have as easily been working with a non-linear function; x in equation 1 simply would be replaced with $f(x)$. In addition, x may be measured in a quantitative metric or a qualitative category. For the former, one can refer to the "value" at which the intervention is set (as in a "dose"). For the latter, one can refer to the "level" (as in experimental or control group). Finally, Freedman actually provides two kinds of response schedules. But, equation 1 is more more consistent with the applications to follow.

Here is what a causal relationship requires. The value of the intervention is first determined. For example, a decision is made to assign a particular number of patrol officers to a order-maintenance policing. Next, the social processes that determine the homicide rate multiply the number of police officers by a constant equal to α_1 and add a constant equal to α_0. Then, independent of the value of x, $\delta_{i,x}$ is generated as a random draw from a distribution with a mean of 0.[2] Finally, to get the response, $\delta_{i,x}$ is added to the linearly transformed value of x. In other words, equation 1 coupled with these side conditions conveys *how* the response is produced when x is the cause of $y_{i,x}$. Thus, one can think of equation 1 as the "schedule" by which a response is generated.

To move from causal *relationship* to a causal *effect*, the response needs to be compared to something. Suppose in the *same* city at the *same* time, the control condition is no patrol officers assigned to order-maintenance policing. The value of the intervention is now set to 0. The value of 0 is a reasonable quantification of a "control" condition, but the response schedule for a comparison deployment policy works for any plausible number of patrol officers.

The social processes that determine the homicide rate now produce the same constant α_0 as when the number of patrol officers is set to any other value. The constant α_1, multiplied by 0, drops out. To α_0 is added another perturbation generated as a random draw from a distribution with a mean of 0. As before, the value of the perturbation drawn is independently of the value of x. A new response results. The causal effect is then defined as the difference between the two responses.

It is important to emphasize that a *definition* of a causal effect has been provided. The definition determines what one is trying to estimate in a data analysis. It follows that for the data analysis to make sense, it must be undertaken in a manner that is consistent with the definition. When all one has is observational data, this can be very difficult.[3]

[2] For the definition of a causal relationship, the mean of the distribution could be any constant. In effect, the value of α_0 is then shifted up or down by the value of that constant. It is usually cleaner, however, to assume a mean of 0, which is no more constraining than assuming that the mean were some other constant.

[3] A key problem is that one can only observe the response of a given city at a given time for a single value for the number of patrol officers assigned to order-maintenance policing. This more generally has been called the "fundamental problem of causal inference." The goal of any attempt to estimate a causal effect is to find a good proxy for the response(s)

Equation 1 is too easily skimmed, and its key implications missed. At the risk of repeating what may be obvious to some readers, four points can be stressed. First, the requirement that the disturbance is generated independently of the value or level of the intervention is essential if there is any hope of isolating the causal effect of the intervention. If the two are related, the two are confounded. Moreover, in practice $\delta_{i,x}$ is unobservable. All one gets to see in the data are $y_{i,x}$ and x. So, the independence assumption is both weighty and impossible to directly verify.[4] Its plausibility depends on theoretical justifications grounded in the substance of what is being studied.

A second point is that because α_0 and α_1 are not subscripted, all units in the study are affected in the same way by the intervention. So, if the units in the study are entire cities, the number of patrol officers assigned to order-maintenance policing alter the homicide rate in exactly the same manner in each. Alternatively, there can be several different kinds of cities, and within each group, the impact is the same. Ideally, these groups need to be defined *a priori*. Post hoc exploration of the data to find such groups can lead to overfitting.[5]

A third point is that the intervention value assigned to some other unit j has no impact on how unit i responds. This is the "no interference" assumption called the "stable unit trait value assumption" (SUTVA) in some statistical circles. Thus, the number of patrol officers assigned to order-maintenance policing in one city has no impact on the homicide rate in another city. This too is difficult to examine empirically but here, geographical distance may provide a sufficient buffer. On the other hand, for a study within a given city based on longitudinal data, the i can refer to time, and then no apparent buffer may exist.

A final point is that because the value or level of the intervention is "set," the intervention needs to be manipulable. In this case, the number of patrol officers assigned to order-maintenance policing has to be subject to direct control. The credibility of such claims would depend on knowing the history of the policing intervention in each city.

The requirements of equation 1 are not a mere technicality. They go to

that cannot be observed (Berk, 2005).

[4] Claims are sometimes made that there are "specification tests" that can address the independence assumption, but these tests simply trade one set of untestable assumptions for another (Berk, 2003: section 9.5).

[5] When there is substantial overfitting, the results obtained do no generalize well. They are data-set specific.

the heart of whether, even in principle, coherent causal inferences can be made. Moreover, the burden is on the researcher to make a convincing case that the definition holds up for the application on hand. And that case will usually require some combination of generally accepted theory or generally accepted empirical facts.

What might be required for this initial example? The researcher would need to provide convincing answers to at least the following questions.

1. How exactly does the way social processes generate homicide rates produce a random perturbation with an expected value of 0 (or any other constant)? How does that process also manage to generate the perturbation independently of the number of patrol officers assigned to order-maintenance policing? And how does that process manage to generate a perturbation in a manner that is identical to random sampling with replacement from a population of perturbations?

2. Why does this homicide rate process lead to a homicide rate that is a linear transformation of the number of patrol officers assigned to order-maintenance policing? If some other functional form is used instead, a case for that form would need to be made.

3. How do the social processes that generate the homicide rate constrain the impact of the intervention so that for a particular number of patrol officers assigned to order-maintenance policing, in each city (or each subset of cities) the effect is exactly the same? What is the mechanism that brings this about?

4. Why do the number of patrol officers assigned to order-maintenance policing in one city have no impact on the homicide rate of other cities?[6]

If the data come from a single city over time, the same issues arise, but some play through a bit differently. For example, why does the number of patrol officers assigned to order-maintenance policing have the exact same impact on homicides every year? Why do the number of patrol officers assigned to order-maintenance policing have no impact on the homicide rate at subsequent times?

Rosenfeld and his colleagues (hereafter referred to as RFB) address no questions of this sort, let alone answer them. The result is necessarily some

[6] As noted above, a good case for this might well be made.

variant of "assume-and-proceed" statistics. We will see, therefore, that it is not apparent what the statistical models are trying to estimate.

3 From Deceptively Simple to Manifestly Complicated

Freedman's response schedule lays the groundwork. What RFB actually do is far more complicated. Some might claim that these complications provide answers to the kinds of questions just raised, or at least make many of them moot. In fact, the fundamental principles from Freedman's response schedule still apply but are embedded in a far more elaborate structure. It can be difficult, therefore, to see what is really going on and effectively impossible to provide the justifications required.

RFB employ a special case of the hierarchical generalized linear model based on Poisson regression. They draw on the excellent exposition of hierarchical models for count data provided by Raudenbush and Bryk (2002, 309-317)[7] but they neglect some important details. Interested readers should consult the Raudenbush and Bryk treatment and not rely exclusively on RFB's Appendix A.

Let's turn now to the fancy model. Hierarchical models of the sort favored by RFB are just conventional generalized linear regression models with more complicated error terms. This is easy to see if the Level-2 equations are substituted into the Level-1 equation and the result simplified.[8]

For the RFB model, these manipulations lead to a response schedule of the same sort as equation 1, but with many more pieces. The time varying Level-1 predictors appear as additive terms. Each is associated with its own multiplicative constant. The same applies to Level-2 predictors associated with the Level-1 intercept. There are also a number of product variables constructed from the Level-1 time trends and the Level-2 predictors meant

[7] In service of full disclosure, with Jan de Leeuw I edit the book series for Sage Publication in which the Raudenbush and Bryk volume appears.

[8] Consider a very simple illustration. Suppose that α_1 in equation 1 is equal to $\beta_0 + \beta_1 z + \epsilon_{i,z}$, where z is a predictor and $\epsilon_{i,z}$ is a perturbation drawn at random independently of x, z and $\delta_{i,x}$. Substituting, one gets $y_{i,x} = \alpha_0 + \beta_0 x + \beta_1 xz + (x\epsilon_{i,z} + \delta_{i,x})$. This is just a conventional linear equation with a main effect, an interaction effect (a product variable), and a two-part error term with non-constant variance. Similar substitutions can be made using any Level-2 equations, although the details will vary a bit.

to explain variation in these trends over cities. These too have their multiplicative constants. Finally, all of the Level-2 perturbations enter additively, either alone or multiplied by a Level-1 predictor. In short, the x in equation 1 is replaced by many x's, α_0 and α_1 are replaced by many α's, and the $\delta_{i,x}$ is replaced by many $\delta_{i,x}$'s, with the notation altered accordingly.

Generalizations of the equation 1 requirements must hold for causal relationships to defined.

1. Each value or level for each x is set independently of the values or levels set for all other x's.[9]

2. The values of all α's are the same for all observational units.

3. The systematic component of the response is then a linear combination of α_0 and the predictors, each predictor weighted by its associated coefficient.

4. Each of the perturbations is drawn at random from normal distributions, independently of the values or levels of the predictors and of each other.

5. The perturbations are added to the linear combination of the predictors.

6. The values or levels set for any given unit have no impact on the response of any other unit.

The implications of these requirements play out much as they did for equation 1. And the credibility of the response schedule depends on making the case that each of its parts is a good approximation of how the world really works.

For example, why are the Level-1 intercept and the regression coefficients for the time trends made a function of Level-2 predictors but not the regression coefficients for the Level-1 time varying predictors (e.g., police per capita)? What is it about homicides that leads to this distinction? Or, what is it about homicides that leads to piecewise linear time trends? Why not, for instance, natural cubic splines? Or, why is a random perturbation shifting

[9] As Heckman emphasizes (2000: 54) "The assumption that the components of x can be varied independently is strong but essential to the definition of a causal effect. (Heckman's notation is changed slightly to be consistent with the notation used in this essay.)

up or down the overall homicide rate in a city at a particular point in time necessarily unrelated to another random perturbation making a time trend at that point in time more or less steep? What is it about the nature of urban homicides that requires this independence? Or, how does it come to be that the perturbations are all drawn from normal distributions, especially when most of the regression coefficients are bounded from above or below at 0?[10]

But there's more. The RFB model is not just a response schedule. It is also a statistical model, which is an attempt to represent the full stochastic process by which the *distribution* of the response is generated. What in addition does this require?

The hierarchical Poisson formulation begins with a response variable that is a count. The count is usually taken to be the sum of events that are independent of one another, with each event subject to the same probability of occurrence.[11] When the counts are a function of predictors, these characteristics hold conditional upon those predictors. For RFB, therefore, the homicides forming the basis of the response variable should be independent of one another, conditional on the intervention and all of the included covariates. Consequently, a homicide that occurs in a given city at a given time does not affect the probability that any subsequent homicides will occur. For example, a given gang-related killing has no impact of the probability of a subsequent gang-related killing. This would seem to rule out revenge or retaliation as motives in gang violence. What is the theoretical or empirical justification for proceeding as if the homicides in a city unfold in this manner?

RFB are not fully satisfied with the Poisson formulation. For the Poisson distribution, the conditional mean must equal the conditional variance. RFB relax this restriction and allow the conditional variance to be larger or smaller than the conditional mean. What is it about how the social world produces homicides that make this appropriate?[12]

There are no compelling answers to the vast majority of questions raised over the past several pages. RFB do not have the answers, nor does anyone

[10] RFB often behave as if they know the sign of the relationship.

[11] There are other ways a count can be Poisson, but the mechanisms are more complicated (Freedman, 1974) and no more applicable to homicides.

[12] In practice, overdispersion is a frequent consequence of omitted variables. The residual deviance is too large because important explanatory variables are not in the model. And if this is true, all bets are off.

else. The problem is that the statistical tools applied far outstrip current substantive knowledge. Absent a clear theoretical rationale and/or compelling results from past empirical research, the modeling effort undertaken by RFB necessarily rests on a large number of arbitrary decisions. If these inputs to the modeling process are arbitrary, so are the outputs. This is not the basis on which one wants to build important public policy.

In addition, is not apparent what RFB are trying to estimate. Different decisions about the structure of the hierarchical model imply different claims about how the world works. If these decisions are not grounded in any articulated rationale, there are a multitude of worlds, all with equal claims to being the world in which homicides are really produced. Which one are we trying to characterize? And it matters. For example, if the Level-2 perturbations are not independent of the predictors at either level, causal effects for the interventions are not defined. If the Level-2 perturbations are not independent of one another, the uncertainty in the model is defined inappropriately.

Concerns of the sort just raised are certainly not new (e.g., Freedman, 1985; 1987; Rubin, 1986; 1991; Holland, 1986; Berk, 1988; 2003; Berk and Freedman, 2003). Nor are hard-hitting critiques of how multilevel models are used in the criminal justice research (de Leeuw, 2005). The responses are old news as well. They include the following: "well, nothing is perfect;" "but, everyone does it;" "if we didn't do it, someone else would;" "the existing work is worse;" "and we'll do better next time." Even if such responses are sincere, they are beside the point in applied settings. Imagine similar words coming from a physician after a botched surgical procedure.

4 Some Statistical Concerns

To this point, I have focused on whether the modeling done by RFB meets the requirements of a serious response schedule and a credible statistical model. Clearly, it does not. As a result, *even before the data are examined*, the research enterprise is in serious trouble. However, there are several worrisome issues in how the data were analyzed that deserve brief mention.

The high multicollinearity that follows from including product variables and their components, even if regressors are centered, can reduce statistical power substantially and even lead to very unstable results (Kreft and de Leeuw, 1998: section 5.7). One has to wonder about the degree to which the null findings might reflect a lack of power. It would be useful to see some version of variance inflation factors.[13]

With all of the statistical tests reported, the usual multiple test problem could be serious. It would be useful to see some sensible discounting of p-values. For example, the Bonferroni correction would be simple to apply. It is usually too conservative, but in this case there was no doubt lots of data snooping that went on before final models were estimated. If anything, the Bonferroni correction would not be conservative enough.

Raudenbush and Bryk stress the importance of doing serious model assessments (2002: chapter 9). Although I am somewhat less sanguine about what can be learned because of the problems discussed earlier (Berk, 2003: chapter 9), model diagnostics are still a good idea. Moreover, the regression formulation that can be constructed by substitution means that a wide variety of diagnostic tools available for conventional linear regression can be used: Cook's distance to address influence, added variable plots to consider functional forms, quantile-quantile plot for evaluating normality and so on. There is no evidence that RFB took advantage of these tools or even the ones suggested by Raudenbush and Bryk.

5 What are the Alternatives?

There are three broad kinds of alternatives to the modeling undertaken by RFB: true experiments, quasi-experiments, and better statistical approximations of true experiments. The first two require better research designs, which implies that researchers usually need to be involved in how the intervention is implemented. The last requires abandoning conventional causal modeling as a data analysis strategy.

5.1 Randomized Experiments

The formal properties randomized field experiments have been known at least since R.A. Fisher's famous book, *The Design of Experiments*, published in 1935. Scores of textbook treatments have followed (e.g., Cox, 1958; Box, Hunter and Hunter, 1978). Thus, it is widely understood that random assignment can balance treatment groups on all possible confounders and

[13] But a post hoc power analysis is a bad idea (Hoenig and Heisey, 2001).

automatically provide a credible justification for statistical tests and confidence intervals. We have more recently come to appreciate that in practice randomized field experiments have weakness as well as strengths (Heckman, 1995). However, they are commonly the design of choice if a key goal is to make plausible causal inferences (Boruch, 1997; Berk, 2005).

Randomized experiments have been used to evaluate an enormous variety of social interventions (Greenberg and Shroder, 2004), many involving policing (Sherman, 1992; Skogan and Frydl, 2004; Weisburd, 2005). In short, randomized experiments for police practices are a genuine option and greatly reduce the kinds of problems undercutting the RFB analysis.

5.2 Quasi-Experiments

Quasi-experiments can range from close approximations of randomized experiments to designs that are no better than observational (Campbell and Stanley, 1963; Shadish et al., 2002). Among the strongest quasi-experiments for causal inference is the generalized regression-discontinuity design, which has proved useful in studies of police activity (Berk and Newton, 1985) and inmate classification procedures (Berk and de Leew, 1999). In randomized experiments, assignment to treatment groups is by a chance mechanism. In the generalized regression-discontinuity design, assignment is by a *known* deterministic mechanism. Knowledge of the deterministic assignment mechanism can lead directly to unbiased estimates of treatment effects, although with somewhat less statistical power than when random assignment is employed. Strong quasi-experiments can be an effective alternative when random assignment is not feasible.

5.3 Estimating Treatment Effects without Conventional Causal Models

Partly in response to the kinds of difficulties faced by RFB, there have been a number of useful attempts over past 20 years to develop causal effect estimators that do not depend conventional causal modeling (e.g., Rosenbaum and Rubin, 1983; 1984; Heckman and Hotz, 1989; Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996; Abadie, 2003; Abadie and Imbens, 2004; Imbens, 2004; Hirano and Imbens, 2005). The basic idea is to concentrate on obtaining useful estimates of an average causal impact of an intervention, treating potential confounders as nuisance variables. Several different kinds of averages may be used depending on the units for which causal inferences are to be made. A key distinction is often whether causal inferences are to be made only for the units in the sample or for the population from which the sample was drawn.

This overall strategy is illustrated well in Paul Rosenbaum's *Observational Studies* (2002) in which a variety of procedures are discussed that allow one to analyze observational data within the spirit of randomized experiments. In randomized experiments, random assignment on the average severs any association between confounders and the intervention. It is not necessary to know what the confounders are, let alone represent them in some statistical model. Random assignment treats them as a nuisance whose impact needs to be eliminated.

Building on the work of a number of statisticians, Rosenbaum approaches observational studies from a similar point of view. The key is trying to represent with covariates the process by which some units get one intervention and other units get another intervention. Statistical approaches that attempt to do this are often called "selection models." If the probability of selection into the various interventions can be accurately estimated for each observation, the probabilities (called a "propensity scores") can by themselves serve as an effective covariate, matching variable or case weights when treatment effects are estimated. In any of these roles, propensity scores can eliminate associations between the intervention and the covariates used to construct the propensity scores. As a result, the impact of confounders can be eliminated without having to represent in a causal model how the covariates affect the response variable.[14]

There are now many variations on this basic strategy. But each depends fundamentally on knowing the variables that affect the selection process, having those variables in the data set, and then representing those variables properly in a model for the propensity scores. These can be demanding requirements and at the very least, Rosenbaum recommends the use of special sensitivity tests for any such models developed.

[14] The formal requirements are defined within the Neyman-Rubin formulation of causal effects that has come to dominate this literature. They require "unconfoundedness:" conditional upon the covariates, the intervention is orthogonal to the potential responses. They also require "overlap:" for any set of covariate values, the probability of receiving a given intervention is not 1.0 or 0.0 (Imbens, 2004: 7-9). Both are very strong assumptions that can be difficult to fulfill in practice.

A key advantage of propensity score approaches over the kinds of causal modeling favored by RFB is relative simplicity. The assumptions required can be plainly articulated, and in some cases usefully explored with data. In addition, selection models reduce the role of many covariates to a single variable represented by the propensity score. A much easier and transparent data analysis will often result. The variety of methods discussed in the references listed at the beginning of this section have many of the same strengths.

5.4 Just Say No

What if random assignment, a strong quasi experiment, or a convincing analysis of observational data are not in the cards. Even if the policy questions are vital, it may be wise to throw in the hand. Suspect science, even the best that can be done under the circumstances, does long run damage to the credibility of all science. The position taken here is that under these circumstances, responsible researchers should withdraw until stronger studies are possible. It may even be possible to help make those stronger studies more likely.

It is important to stress that the call for quality is not a call for perfection. All research necessarily has flaws. Nor is there a requirement that the research be done by the book. The book is often wrong or at least out of touch with reality. The point is that a more responsible balance must be struck between rapid responsiveness to the pressing social problems of the day and research that the social science community can proudly stand behind.

6 Conclusions

The difficulties faced by RFB are widespread in criminal justice research and social science more generally (Berk, 2003). Rapid technical developments in causal modeling have produced a gaping hole between what the models require and the subject-matter knowledge available. Proceeding with modeling nevertheless risks undermining the credibility of all social science research, even research resting on far stronger methodological foundations.

Fortunately, there are options. One can shift much of the burden from suspect causal models to the research designs through which the data are generated. Randomized experiments and strong quasi-experiments can often lead to credible results when analyses of observational data will fail. If observational data are the only information that can be obtained, there are data analysis strategies that place fewer demands on existing social science knowledge.

There is also the option of just saying no. Often the responsible decision is not to engage. One can then wait until there are data and statistical procedures able to produce credible results. In long run, this will lead to stronger research that policy makers can rightfully take seriously.

References

Abadie, A. (2003) “Semiparametric Instrumental Variable Estimation of Treatment Response Models.” *Journal of Econometrics* 113: 231-263.

Abadie, A. and G.W. Imbens (2004) “Large Sample Properties of Matching Estimators for Average Treatment Effects” NBER Technical working paper no. 283.

Angrist, J.D., Imbens, G.W., and D. Rubin (1996) “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91: 444-472.

Berk, R.A. (1988) “Causal Inference for Sociological Data,” in N. Smelser (ed.) *Handbook of Sociology*, Newbury Park, Sage Publications.

Berk R.A., and P.J. Newton (1985) “Does Arrest Really Deter Wife Battery? An Effort to Replicate the Findings of the Minneapolis Spouse Abuse Experiment.” *The American Sociological Review* 50: 253-262.

Berk, R.A. and J. de Leeuw (1999) “An Evaluation of California’s Inmate Classification System Using a Generalized Regression Discontinuity Design.” *Journal of the American Statistical Association*, 94: 1045-1052.

Berk, R.A., and D.A. Freedman (2003) “Statistical Assumptions as Empirical Commitments.” In T.G. Blomberg and S. Cohen (eds.), *Punishment and Social Control*, second edition. New York: Aldine de Gruyter: 235-254.

Berk, R.A., (2003) *Regression Analysis: A Constructive Critique.* Newbury Park: Sage Publications.

Berk, R.A. (2005) “Randomized Experiments and the Bronze Standard.” UCLA preprint #425, Department of Statistics, UCLA (http://preprints.stat.ucla.edu/)

Boruch, R.F., (1997) *Randomized Field Experiments for Planning and Evaluation: A practical Guide.* Newbury Park, CA: Sage Publications.

Box, G.E.P, Hunter, W.G., and J. S. Hunter (1978) *Statistics for Experimenters.* New York: John Wiley.

Campbell, D.T., and J.C. Stanley (1963) *Experimental and Quasi-Experimental Designs for Research.* Boston: Houghton Mifflin.

Cox, D.R. (1958) *Planning of Experiments.* New York: John Wiley.

de Leeuw, J. (2005) “Comments on Pardoe and Weitner: Sentencing Convicted Felons in the United States: a Bayesian Analysis Using Multilevel Covariates.” *Journal of Statistical Planning and Inference*, forthcoming.

Fisher, R.A. (1935) *The Design of Experiments.* New York: Hafner Press.

Freedman, D.A. (1974) “The Poisson Approximation of Dependent Events.” *The Annals of Probability* 2: 256-269.

Freedman, D.A. (1985) “Statistics and the Scientific Method,” in *Cohort Analysis in Social Research: Beyond the Identification Problem*, W.M. Mason, and S.E. Fienberg (eds.), New York: Springer Verlag.

Freedman, D.A. (1987) “As others See Us: A Case Study in Path Analysis.” (with discussion). *Journal of Educational Statistics* 12: 101-223.

Freedman, D.A. (1991) “Statistical Models of Shoe Leather,” in *Sociological Methodology, 1991*, P.V. Marsden (ed.). Oxford: Basil Blackwell.

Freedman, D.A. (2004) “Graphical Models for Causation and the Identification Problem.” *Evaluation Review* 28: 267-293.

Greenberg, D. and M. Schroder (2004) *The Digest of Social Experiments*, Third Edition. Washington, DC: The Urban Institute Press.

Heckman, J. (2000) “Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective.” *Quarterly Journal of Economics* 115: 45-97.

Heckman, J. (1995) “Assessing the Case for Randomized social Experiments. ”*Journal of Economic Perspectives* 9: 85-110.

Heckman, J. and J. Hotz (1989) “Alternative Methods for Evaluating the Impact the Training Programs” (with discussion). *Journal of the American Statistical Association* 84: 862-874.

Hirano, K., and G.W. Imbens (2005) “The Propensity Score with Continuous Treatments,” in *Missing Data and Bazyesian Methods in Practice: Contributions by Donald Rubin's Statistical Family*, New York: John Wiley, forthcoming.

Hoenig, 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.

Holland, P.W. (1986) “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81: 945-960.

Imbens, G.W., and J.D. Angrist (1994) “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 61: 467-476.

Imbens, G. (2004) “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review.” *The Review of Economics and Statistics*: 86: 4-29.

Kreft, I. and J. de Leeuw (1998) *Introducing Multilevel Modeling.* Thousand Oaks, CA: Sage Publications.

Neyman, J. [1923] (1990) “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.” Translated and edited by D.M. Dabrowska and T.P. Speed, *Statistical Science*, 5: 465-471.

Manski, C.F. (1994) “The Selection Problem,” in *Advances in Econometrics, Sixth World Congress*, edited by Christopher Sims. Cambridge, England: Cambridge University Press.

Raudenbush, S.W. and A.S. Bryk (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*, second edition. Thousand Oaks, CA: Sage Publications.

Rosenbaum P., and D. Rubin (1983) “The Central Role of the Propensity Score in Observational Studies of Causal Effects,” *Biometrika* 70: 41-55.

Rosenbaum P., and D. Rubin (1984) “Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score.” *Journal of the American Statistical Association* 79: 516-524.

Rosenbaum, P.R. (2002) *Observational Studies*,second edition. New York: Springer-Verlag.

Rubin, D. B. (1986) “Which Ifs Have Causal Answers.” *Journal of the American Statistical Association* 81: 961-962.

Rubin, D.B. (1990) “Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies.” *Statistical Science* 5: 472-480.

Shadish, W.R., Cook, T.D., and D.T. Campbell (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sherman, L.W. (1992) *Policing Domestic Violence*. New York: The Free Press.

Skogan, W., and K. Fryd (eds.). (2004) *Fairness and Effectivness in Policing: The Evidence*. Washington, D.C.: The National Academy Press.

Weisburd, D. (2005) “Hot Spots Policing Experiments and Criminal Justice Research: Lessons from the Field.” *The Annals* 599: 220-245.