ANNUAL REVIEWS **Further**

Click here for quick links to Annual Reviews content online, including:

- Other articles in this volume
- Top cited articles
- Top downloaded articles
- Our comprehensive search

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Credible Causal Inference for Empirical Legal Studies

Daniel E. Ho[1] and Donald B. Rubin[2]

[1]Stanford Law School, Stanford University, Stanford, California 94305; email: dho@law.stanford.edu

[2]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138; email: rubin@stat.harvard.edu

Annu. Rev. Law Soc. Sci. 2011. 7:17–40

The *Annual Review of Law and Social Science* is online at lawsocsci.annualreviews.org

This article's doi: 10.1146/annurev-lawsocsci-102510-105423

Copyright © 2011 by Annual Reviews. All rights reserved

1550-3585/11/1201-0017$20.00

Keywords

research design, policy evaluation, matching, regression discontinuity

Abstract

We review advances toward credible causal inference that have wide application for empirical legal studies. Our chief point is simple: Research design trumps methods of analysis. We explain matching and regression discontinuity approaches in intuitive (nontechnical) terms. To illustrate, we apply these to existing data on the impact of prison facilities on inmate misconduct, which we compare to experimental evidence. What unifies modern approaches to causal inference is the prioritization of research design to create—without reference to any outcome data—subsets of comparable units. Within those subsets, outcome differences may then be plausibly attributed to exposure to the treatment rather than control condition. Traditional methods of analysis play a small role in this venture. Credible causal inference in law turns on substantive legal, not mathematical, knowledge.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

1. MOORE'S LAW OF PARKING

It would be easy to dismiss the parking studies of Underhill Moore. From 1933 to 1937, the famed Yale Law professor sought to quantify the causal effect of law. He worked with a cadre of research assistants to count over 13,000 instances of parked cars spanning 15 New Haven areas, dispatched police officers to place tags for dollar fines on over 3,400 cars, and painted large white ovals to simulate a would-be roundabout in the middle of an intersection. While the research entailed the minutiae of defining when a car had parked (when the wheels stopped moving), the goals were lofty: nothing short of a "general theory of human behavior" in relation to law (Moore & Callahan 1943, p. 2). Moore himself admitted that the venture walked a fine line between the avant-garde and the absurd: "The[y] ridicule my project. They do not understand it.... I am writing for [those] groping for ways of applying the scientific method to the social sciences.... [Y]ears from now a kindred soul may find in my crude researches some clue to the solution" (Douglas 1950, p. 188).

Despite the rather obvious mismatch of legal theory and empirical data, as a matter of methodology, the ridicule is misplaced. Moore's research, like much of the first wave of empirical legal studies in the 1920s and 1930s, grappled with thorny methodological challenges to drawing inferences about the causal effects of laws, all while modern foundations for experiments were only beginning to take shape (Schlegel 1995). It was not until 1925 that Fisher offered randomization as the "reasoned basis for inference" for experiments (Fisher 1925, 1935). How then could one infer the causal effect of a parking regulation? As William O. Douglas and collaborators contemporaneously noted, "[P]roblems will center around the development of more adequate techniques for controlling errors and the production of data from which inferences as to the causal connection of these various factors...will emerge" (Clark et al. 1930). More generally, how could the first empiricists quantitatively assess the impact of law?

Moore's approach was pioneering, if not downright modern. He reasoned that when an "experimental situation[] could not be manufactured at will," one could "tak[e] advantage of the terms of the ordinance itself" (Moore & Callahan 1943, pp. 88–89). This insight was crucial. Simply examining streets with or without limits would be comparing the incomparable. Moore's solution capitalized on the geographic or temporal arbitrariness of when a parking time limit applied. On Crown Street, the limit applied on one side of the street for one month and on the other side the next. On Church Street, the 15-minute limit applied only until 7:00 PM. And so Moore collected data immediately before and after 7:00 PM to isolate the impact of the parking regulation. To assess whether the differences in the time of day affected inferences, Moore further checked for similarity of traffic flows and driver activities, the latter monitored by research assistants following subjects to destinations.

Figure 1 presents the data that Moore collected for one street, with minutes parked on the *x*-axis (on a log scale). (The bins are in 1-, 5-, or 10-minute increments as presented by Moore.) The dark green outlined histogram presents parking durations for cars parked from 5:30–6:30 PM, when the 15-minute limit applied. The light green filled histogram presents parking durations for cars parked from 7:00–8:00 PM, when no limit applied. Parking durations shifted considerably. Roughly 36% of cars parked for 15 minutes or less when the limit was inapplicable, compared with 57% when the limit applied. On average, the effect of the time limit was to decrease the time in the space by 40 minutes, although a large number of drivers still failed to comply with the time limit in place.

Of course, from a modern perspective, Moore's methodology is lacking in certain respects. Demand for parking may differ sharply after 7:00 PM. The time limit may still affect parking behavior after 7:00 PM. And differences

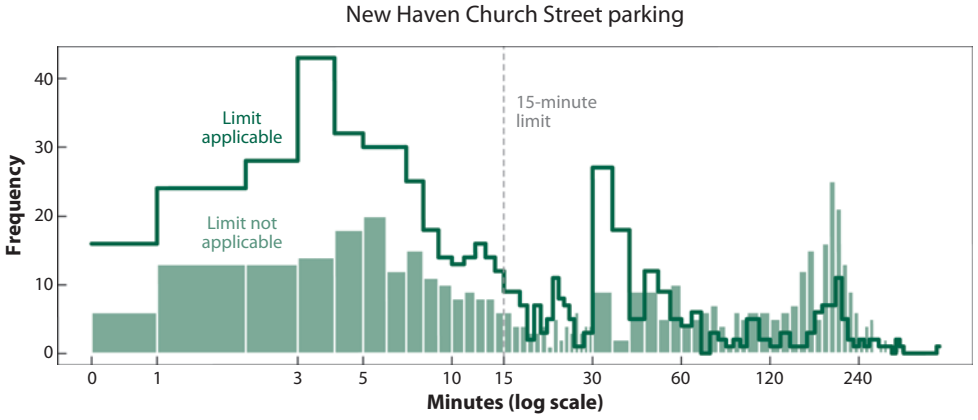

Figure 1

New Haven parking on Church Street for 35 days in 1936. The light green histogram represents duration of parking starting from 7:00–8:00 PM when the 15-minute parking limit (*gray dashed line*) was inapplicable. The dark green outlined histogram represents duration of parking starting from 5:30–6:30 PM when the 15-minute parking limit was applicable. The *x*-axis is on a log scale and bin widths are as Moore & Callahan reported (1 minute up to 30 minutes, 5 minutes from 30–100 minutes, and 10 minutes from 100–450 minutes). Source: Moore & Callahan (1943, pp. 104–6).

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

could be due to chance alone.[1] But in the crucial respect of *research design*, Moore's study was pioneering. Indeed, it may be the first informal application in law of what we would now call "regression discontinuity" design—using the discontinuity at 7:00 PM to assess the causal effect of regulation on parking—appearing some 30 years before the technique was formalized. Moore may be to regression discontinuity what the Trial of the Pyx is to hypothesis testing: legal pioneering of what statistics would later formalize (Stigler 1977).

In this article, we review modern developments in the statistics of causal inference, focusing in particular on matching methods and regression discontinuity. What unifies such approaches is the prioritization of research design to create—without reference to any data on outcomes—subsets of comparable units. In what might be considered a vindication (or even "kindred soul") of Moore, modern approaches emphasize design over methods of analysis.

Our article proceeds as follows. Section 2 discusses the broad shift toward credible, design-oriented inference in social science. Section 3 explains the widely used potential outcomes framework that clarifies the central issues of causal inference. We use as a running example data first analyzed in an important study by Berk & de Leeuw (1999) (BdL) of the causal effect of maximum-security incarceration on prison misconduct, which we detail in Section 4. Section 5 uses this data set to illustrate the chief problem of "model sensitivity" that plagues much conventional regression-based practice. Section 6 details what we mean by a focus on research design: collecting, organizing, measuring, and preparing the data without reference to outcome data. Sections 7 and 8 apply matching methods and regression discontinuity to BdL's prison data, which we compare to experimental results in Section 9. Both approaches provide estimates much closer to experimental findings than do naive regression-based approaches. Section 10 concludes.

[1] In modern terminology, these defects would refer to issues of continuity/smoothness of outcomes with respect to the forcing variable, the "stable unit treatment value assumption," and sampling variability.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

2. CAUSAL INFERENCE IN EMPIRICAL LEGAL STUDIES

Causal inference has always been central to the enterprise of empirical legal studies. How does no-fault insurance law affect auto injury compensation? Do defendants with court-appointed counsel fare worse than those with retained counsel? How does discretionary jurisdiction affect the business of the Supreme Court? All these were questions that led the likes of Roscoe Pound, Felix Frankfurter, and James Landis to turn to quantitative data collection in the 1920s and 1930s (Kritzer 2010). Yet their efforts met with frustration. Said William O. Douglas at the conclusion of a project on the causes of bankruptcy: "All the facts which we worked so hard to get don't seem to help a hell of a lot" (Schlegel 1995, p. 230).

More recently, a similar frustration has surfaced in cognate disciplines about the limits of conventional (regression-based) causal inference. Clearer conceptualization of causal inference has led to an increasing skepticism about the "age of regression" (Morgan & Winship 2007; see also Berk 2004; Donohue & Wolfers 2006; Gelman & Meng 2004; Leamer 1978, 1983; Manski 1995; Pfaff 2010; Sobel 2000; Strnad 2007). "Without . . . strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive" (Sekhon 2009). Or, as Douglas might note: "All the [regressions] we worked so hard to get don't seem to help a hell of a lot."

Yet something else is afoot. Ayres (2008) dubbed the use of large-scale microdata and field experiments the age of "super crunching." Two leading economists have coined it the "credibility revolution" (Angrist & Pischke 2010). And one researcher forecasts "dramatic transformation" in social science with a deeper understanding of causal inference (Sobel 2000). The unifying feature of this movement is the attempt to hew as closely as possible to an experiment. The law has not remained untouched by this movement. Experimental approaches have reinvigorated our understanding of discrimination (Ayres 1991, Pager 2003), corporate governance (Guttentag et al. 2008), the legal profession (Abrams & Yoon 2007), and health care (King et al. 2007), to name just a few (for others, see, e.g., Angrist 1990, Gerber & Green 2000, Gibson 2008, Green & Winik 2010, Ho & Imai 2006). Even when there is no randomized intervention (when the study is "observational"), approaches directly appealing to an experimental template have crystallized the key issues for empirical inference. Matching methods, for example, have been applied to race and sex (well defined only in certain contexts) (Boyd et al. 2010, Greiner 2008, Greiner & Rubin 2010, Ridgeway 2006), criminal law (Berk & Newton 1985, Helland & Tabarrok 2004, Mocan & Tekin 2006, Papachristos et al. 2007, Petersilia et al. 1986), intellectual property (Qian 2007), corporate governance (Litvak 2007), labor and employment (Dehejia & Wahba 2002, Morantz 2010), environment (List et al. 2006), regulation (Galiani et al. 2005), constitutional law (Persson & Tabellini 2002), election law (Brady & McNulty 2007), civil rights (Epstein et al. 2005), and education (Ho 2005a,b). Regression discontinuity has similarly touched on numerous areas of the law including education (Angrist & Lavy 1999, Kane et al. 2006, Ludwig & Miller 2007, Thistlethwaite & Campbell 1960, van der Klaauw 2002), antidiscrimination (Grogger & Ridgeway 2006, Hahn et al. 1999), corporate governance (Black et al. 2008; Listokin 2008, 2009), crime (Chen & Shapiro 2007; Hjalmarsson 2009a,b; Lee & McCrary 2005), labor and employment (DiNardo & Lee 2004, Lalive 2008, Lemieux & Milligan 2008), health (Card et al. 2008), environment (Chay & Greenstone 2005), property (Bubb 2009), housing (Berry & Lee 2007), and elections (Eggers & Hainmueller 2009, Hopkins 2009, Lee 2008; see also Gerber et al. 2008; for more examples, see Lee & Lemieux 2010, pp. 339–42, and table 5 therein). For top economics, political science, sociology, and statistics journals, **Figure 2** reveals a dramatic impact in the past decade, measured by articles mentioning matching and regression discontinuity.

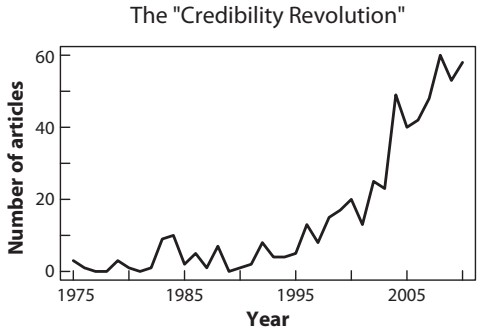

Figure 2

The credibility revolution: number of articles discussing matching and regression discontinuity in top 23 economics, political science, sociology, statistics & probability, and social science (mathematical methods) journals. All top 10 journals based on 2009 impact factor in Journal Citation Reports categories for which full-text searches were available from 1975–2010 were chosen, with duplicated journals omitted. Search strings were for "matching methods," "regression discontinuity," "propensity score," "Thistlethwaite/5 Campbell," and "matching and 'potential outcomes'" in JSTOR, ProQuest, and journal-specific Web sites.

Yet scholars not following these developments may be baffled. How should legal scholars understand and assess these approaches? What principles can legal empiricists incorporate from this rapidly growing literature? Are they in fact more credible? We provide a first guide and review to begin to answer these questions for a general legal audience.

3. POTENTIAL OUTCOMES

We begin by articulating a widely used framework for causal inference, often called the "Rubin Causal Model" (Holland 1986) owing to a series of seminal papers by Rubin (Rubin 1974, 1976, 1978, 1979; Rosenbaum & Rubin 1983b, 1984). The idea is deceptively simple, yet it clarifies the key conceptual issues of causal inference and can be explained without math. Specifically, we are interested in the effect of a single intervention, which we refer to as the "treatment," compared with the baseline of "control." For example, one crucial question for prison administration is the causal effect of maximum-security imprisonment (treatment) versus minimum-security imprisonment (control) on the outcome of misconduct.

Each unit then has two "potential outcomes," one under treatment and one under control. The fundamental problem of causal inference is that we never observe both (Epstein & King 2002, Holland 1986, Rubin 1978). If a prisoner is sent to a maximum-security prison, we cannot observe the "counterfactual" outcome of how she might have fared in a minimum-security prison. Implicit in this framework is that (*a*) there are no hidden versions of the treatment and (*b*) treatment of one unit does not affect the potential outcomes of another unit (sometimes referred to as the stable unit treatment value assumption, SUTVA). The former would be violated, for example, if two types of maximum-security prisons were available to one prisoner, each with different effects on that prisoner's behavioral misconduct. The latter would be violated, for example, if the prison assignment of one gang member affected the behavior of a member of an opposite gang.

Figure 3 visualizes how this framework applies to a data set. The box can be considered the data set, with units in rows, pretreatment covariates in the left columns, potential outcomes in the middle two columns, and an

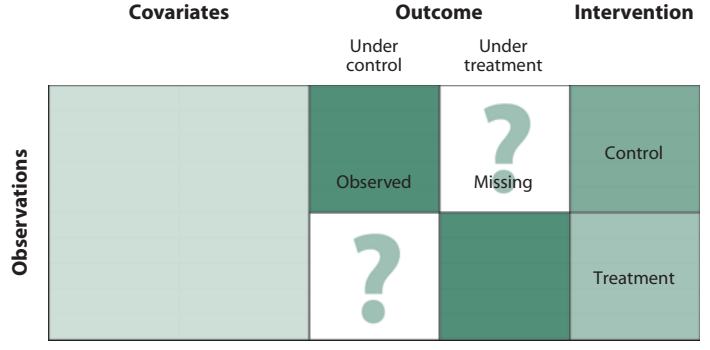

Figure 3

Potential outcomes framework for causal inference. This figure plots a hypothetical data set, with the left columns representing pretreatment covariates, the next two columns representing the potential outcomes, and the last column representing the intervention of treatment or control. Potential outcomes are never jointly observed, and causal inference can therefore be conceived of as a missing data problem.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

indicator for treatment in the last column. Green cells indicate that data are observed, and white cells represent "missing data." For example, we observe the outcomes under control for the upper half of the data set and the outcomes under treatment for the lower half of the data set.

This framework highlights several points. First, causal inference is a matter of inferring missing data. Because we never observe counterfactual outcomes, causal inference is inherently uncertain (and hence a probabilistic venture). Second, causal inference is difficult to conceive of without an intervention. More succinctly, "No causation without manipulation" (Rubin 1975, Holland 1986). This poses a particular challenge for empirical inference in areas such as antidiscrimination law, where immutable characteristics per se cannot be manipulated in a clearly defined way (Greiner & Rubin 2010). Lastly, the framework highlights the importance of an experimental template for the research. If resources were no constraint, researchers should be able to articulate how one might design an experiment to study the question of interest.

Why is the intellectual idea of an experiment so crucial, even in observational research? The key feature of an experiment is that treatment is *randomly* assigned to units. Randomization over a large number of units ensures that treatment and control units are comparable in all respects other than the treatment. "Balance" exists along all possible covariates. Randomizing prisoners to security levels would ensure that maximum-security prisoners, for example, are similar in age, sex, and criminal history to non-maximum-security prisoners. We can thereby properly infer the missing potential outcomes and hence estimate the "treatment effect."

In observational settings, differences in the outcomes may be "confounded" by other factors. Prison authorities, for example, may intentionally sort prisoners by risk profile into facilities of different security levels. Thus, differences in behavior between maximum- and non-maximum-security prisoners would be confounded by risk profile. Observational research can be seen as replicating the hypothetical experiment by achieving balance on these confounding (pretreatment) covariates. The crucial assumption in most observational studies then boils down to unconfoundedness (alternatively known as exogeneity, conditional exogeneity, ignorability, or selection on observables): that, given covariates, the treatment is random, so researchers can attribute differences to the treatment. The credibility of unconfoundedness, as discussed below, is a qualitative judgment that depends crucially on substantive knowledge.

4. APPLICATION: MAXIMUM-SECURITY PRISONS AND MISCONDUCT

As a running example to fix ideas, we use a prison data set first analyzed by BdL. The data set contains information for 3,918 California inmates admitted to prison in 1994. We (and BdL) are interested in the causal effect of maximum-security imprisonment on misconduct. We use maximum security as shorthand for facilities that have "inside or outside cell construction with a secure perimeter, and both internal and perimeter armed coverage" (CDC 2000, ch. 6, art. 5, § 62,010.6). (Variation within such Level IV facilities still exists, but we follow BdL and focus only on the impact of Level IV.)

To assess the credibility of a causal inference, understanding the treatment assignment process is crucial. California's procedure in 1994 for assigning inmates into prison worked in three steps (BdL, CDC 2000, Petersilia 2008). First, after a defendant was sentenced, the California Department of Corrections (CDC) classified inmates by security risk.[2] CDC used inmate background, prior escape, and prior incarceration information to

[2] Of course, CDC here does not refer to the Centers for Disease Control and Prevention, but we use the acronym to be consistent with BdL.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Table 1 Example of inmate classification. For example, a sentence length of 10 years would result in 27 points [= (10–1) × 3] added to the classification score and no high school degree adds 2 points. Given this background and prior incarceration behavior, the inmate would be assigned a score of 53. The example is only illustrative, as other factors (favorable prior behavior, undocumented prior behavior) are taken into account [CDC 2000, ch. 6, art. I, § 61010.11.2 (Form 839)]

		Example	
Factors	**Calculation**	**Value**	**Score**
Background factors			
Sentence length (x)	$(x - 1) \times 3$	10	27
Under age 26	+ 2	Yes	2
Not married	+ 2	Yes	2
No high school degree	+ 2	Yes	2
Unemployed	+ 2	Yes	2
No military service	+ 2	Yes	2
Number of escapes from minimum custody	× 4	1	4
Number of escapes from medium custody	× 8	0	0
Number of escapes with force	× 16	0	0
Prior incarceration behavior			
Number of serious disciplinaries	× 4	1	4
Number of assaults on staff	× 8	1	8
Number of assaults on inmates	× 4	0	0
Number of possessions of deadly weapon	× 4–8	0	0
Number of inciting disturbances	× 4	0	0
Number of assaults causing serious injury	× 16	0	0
		Total score:	53

calculate a "classification score" ranging from 1–80. **Table 1** sketches how major factors were incorporated for a hypothetical inmate, resulting in an overall score of 53. Sentence length was the primary factor, with each additional year resulting in 3 more points, but age, marital status, high school degree, employment, military service, and prior escape attempts were also included. Any prior physical assault on prison staff would add 8 points to the score. Across the sample, the mean classification score was 31 (SD = 12).

Second, in most cases CDC exclusively used the classification score to assign inmates to facilities of given security levels. The CDC *Operations Manual* provided that a classification score of 52 or higher would lead to maximum-security confinement (CDC 2000, ch. 6, art. 1, § 61010.11.4). The first row of **Table 2** shows that inmates in non-maximum-security facilities had an average score of 24 (SD = 13), compared with 66 (SD = 12) in maximum-security facilities.

Third, in certain "administrative placements," CDC deviated from the score based on other attributes. Sex offenders, for example, were more likely to be placed in maximum-security prisons (BdL). Overcrowding could result in alternate placement. Other special case factors included (*a*) whether behavior indicated that an inmate was "capable of successful placement" at a lower level facility, (*b*) the existence of documented "enemies" at institutions, (*c*) family ties, (*d*) medical conditions, and (*e*) work skills [CDC 2000, ch. 6, art. I, § 61010.11.2 (Form 839)]. Because the CDC score was the primary determinant of inmate placement, we also refer to it as the "forcing" variable, namely the variable that "forces" the treatment of maximum-security confinement. Administrative placements, however, mean that the classification score only

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Table 2 Summary statistics of prison incarceration data. The first two columns present statistics for prisoners in non-maximum-security prisons ("controls"); the third and fourth columns present statistics for maximum-security prisoners ("treatment"); the fifth and sixth columns present statistics for all subjects in the data set. The last column presents the *p*-value testing for the difference in means or proportions between treatment and control groups. For the inmate classification score (an ordinal measure), the statistics are means and SDs, while for strike-three offense and behavioral misconduct (binary measures), the statistics are counts and proportions (of the subgroup)

	Non–maximum security		Maximum security		All		Difference
	Mean/ count	SD/ prop.	Mean/ count	SD/ prop.	Mean/ count	SD/ prop.	*p*-value
Inmate classification score	24	13	66	12	31	12	0.000
Strike-three offense	208	0.07	523	0.72	731	0.19	0.000
Behavioral misconduct	890	0.28	246	0.34	1136	0.29	0.002
Total number	3188	0.81	730	0.19	3918	1.00	

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

probabilistically forced treatment.[3] The last row of **Table 2** shows that roughly 81% of inmates were placed in non-maximum-security prisons (control), whereas 19% were placed in maximum-security prisons (treatment).

Table 2 presents two further key variables. One key pretreatment covariate was whether an inmate was sentenced under California's "three strikes" law. (The full set of covariates used in the intake procedure was, unfortunately, not available to us.) Under California law, a third serious felony led to sentence enhancements. These, in turn, increased the probability of maximum-security-level assignment. In the BdL data, roughly 72% of maximum-security inmates were three-strike inmates, compared with only 7% of non-maximum-security inmates. The outcome of interest is whether the inmate was cited for any instance of behavioral misconduct while imprisoned (e.g., failure to obey an order, drug trafficking, or assault). Roughly 29% of inmates—34% of maximum-security and 28% of non-maximum-security inmates—engaged in misconduct.

The last column reports results from tests of differences between the treatment and control groups in the classification score (the forcing variable), strike-three offense (the covariate), and behavioral misconduct (the outcome). We provide these tests only for expositional purposes; as we emphasize below, the design phase should not examine final outcome data. Although the raw difference in the outcome is statistically significant (p-value $= 0.002$), treated inmates also had statistically significantly higher classification scores and third strikes. Three-strike inmates were likely prone to more dangerous conduct, hence confounding the raw difference.

Figure 4 plots the classification score on the x-axis against the probability of misconduct on the y-axis. Each dot represents the proportion of prisoners that engaged in behavioral misconduct at a given classification score and security level. The solid gray and hollow green dots represent inmates in non-maximum- and maximum-security prisons, respectively, and are proportional to sample size. For example, the large gray dot in the bottom left represents 125 non-maximum-security inmates with a classification score of 1, 25% of whom engaged in misconduct. The vertical line represents the threshold of the classification score used to assign inmates to maximum-security prison. Ninety-two percent of inmates with scores of 52 or above were assigned to maximum-security prison, while 98% with scores below 52 were assigned to non-maximum-security prison.

[3] Some use the terms "sharp" and "fuzzy" regression discontinuity to distinguish whether the threshold of the forcing variable deterministically or probabilistically assigns treatment.

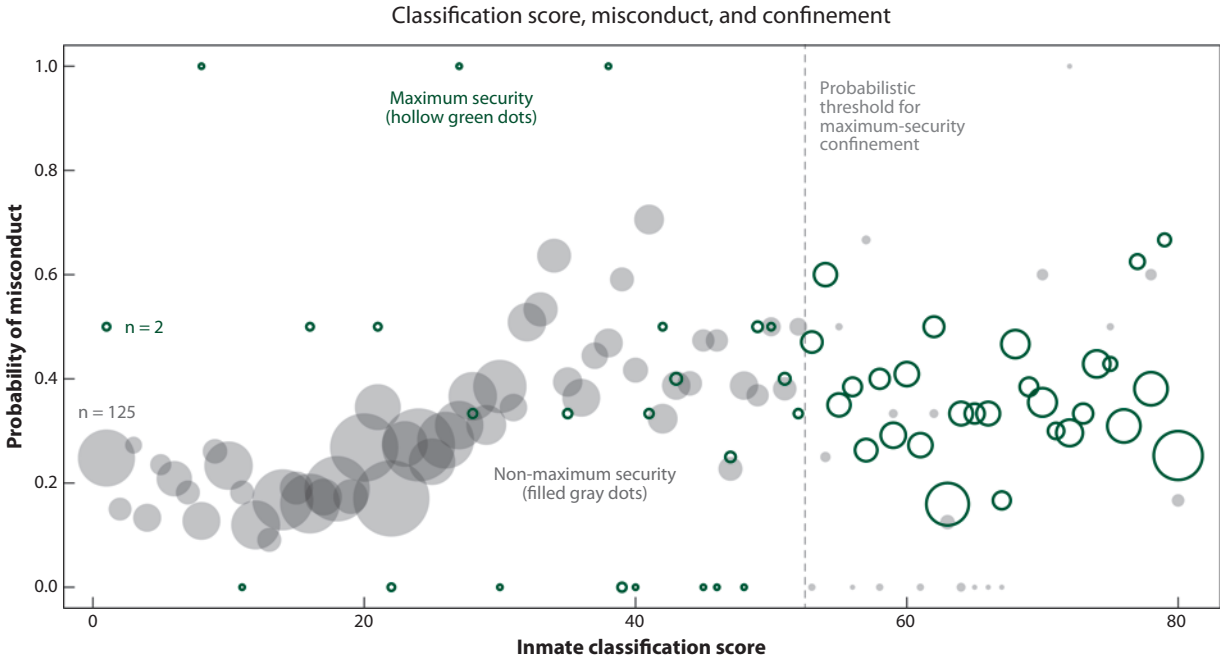

Figure 4

Outcome of behavioral misconduct in prison against the inmate classification score. The *x*-axis presents the inmate classification score, and the *y*-axis presents the proportion of prisoners at each score engaging in behavioral misconduct while imprisoned. Filled gray dots indicate prisoners in non-maximum-security prisons, and hollow green dots indicate prisoners in maximum-security prisons. Dots are proportional to sample size, so the large dot in the bottom left represents 125 prisoners in non-maximum-security prisons, 25% of whom engaged in misconduct. The vertical gray dashed line represents the threshold of the classification score, which substantially increases the probability of maximum-security confinement.

Because the classification score is only used probabilistically, inmates could still be assigned to either level of security across the entire range of the classification score. For example, the small green dot at the left represents two inmates with a score of 1 confined to maximum-security prison, one of whom engaged in misconduct. Overall, 67 inmates with scores of 52 or above were placed in non-maximum-security prisons, and 51 inmates with scores below 52 were placed maximum-security prisons. As discussed below, the visualization in **Figure 4** can help considerably in grasping how a causal effect is identified by regression, matching, and regression discontinuity approaches.

BdL originally used the data to study (*a*) the effectiveness of CDC risk sorting and (*b*) the causal effect of maximum-security imprisonment on prison misconduct. Logically, even the direction of the causal effect is unclear. Stronger security measures may deter misconduct (Zimring & Hawkins 1973), or such facilities may induce marginal inmates to acquire deviance from the worst inmates, thereby increasing misconduct (Bayer et al. 2009). Based on a (logit) regression model that capitalized on the discontinuity in the assignment process and laudable sensitivity analyses, BdL concluded that "the balance of evidence supports an interpretation in which assignment to [maximum security] reduces the odds of misconduct" (BdL, p. 1052). Although the analysis below diverges from these findings, these approaches have been rapidly developing in the past decade and so are recondite to most researchers. We use BdL because it is a landmark study that deserves recognition not only for its central insights in research design, but also for applying them into

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

pioneering field experiments that enable us to assess the validity of observational approaches.

5. INCREDIBLE INFERENCE: CONVENTIONAL REGRESSION-BASED PRACTICE

For causal inference, the overwhelming recognition in applied statistics is that regression alone is fragile (Angrist & Krueger 1999; Berk 2004; Dehejia & Wahba 1999; Ho et al. 2007; King & Zeng 2007; Lalonde 1986; Leamer 1978, 1983; Manski 1995; Rubin 1973, 1975, 2006; Strnad 2007). Even under unconfoundedness, results are highly sensitive.

To illustrate this fact, we apply naive regression-based approaches to the prison data. Each of the panels in **Figure 5** overlays model-based (pointwise) 95% confidence intervals to summarize the results from a range of regression models against the prison data. For example, the top left panel presents the (logit) model reported by BdL. The gray band plots the predicted probability of misconduct for non-maximum-security prisoners, and the green band plots the predicted probability of misconduct for maximum-security prisoners. These curves, if correctly specified, allow us to impute counterfactual outcomes. The difference between the two is the estimated average treatment effect: Maximum security decreases misconduct by 13%, plus or minus 4%.

But the model imposes two strong and unwarranted assumptions. First, it assumes that the probability effectively has a *linear* relationship with the classification score (more precisely, linearity in the log odds). Second, it assumes that the relationship between the classification score and misconduct is *homogeneous* across treatment and control groups. The data in **Figure 4** immediately show why these assumptions are not only heroic, but also largely unverifiable by the data. Because there are very few control units with scores above 52 and very few treated units with scores below 52, the gray bands extrapolate considerably from the data. Few data exist in those regions, so the predictions are highly sensitive to linearity and homogeneity assumptions.

Figure 5*b* relaxes the homogeneity assumption, allowing the slopes of the two curves to differ between treatment and control groups. The model now predicts that maximum-security prison (*a*) reduces misconduct at scores above the threshold of 52, but (*b*) *increases* misconduct at low ranges. Some might interpret this as evidence of how prison inculcates bad behavior (Bayer et al. 2009). But the answer is not really found in the data. Only six maximum-security prisoners have a score below 20.

Figure 5*c* instead allows for nonlinear smooth trends with a constant shift for the level of security. Here the curves are indistinguishable, showing no evidence of an effect. **Figure 5*d–f*** allow for heterogeneous smooth trends with varying degrees of smoothness. **Figure 5*d*** might suggest that the effect is only positive just below the threshold, directly contradicting the model of **Figure 5*b***. **Figure 5*e,f*** simply show that the statistical uncertainty dwarfs any evidence of a treatment effect, with the confidence bands overlapping entirely across the entire range of the classification score.

How would a researcher determine the "best" model? How much smoothness should be assumed? Should we impose homogeneity? Linearity? Even with just one covariate, a staggering set of specification choices presents itself. One "nonparametric" way forward would be to estimate the probability at each classification score for treatment and control groups, resulting in 160 parameters. But as other covariates are added, the number of parameters grows exponentially. Adding 360 possible months of sentence length, the number of parameters becomes 57,600 (360×160); adding ages of 15–64 years, the number becomes 2,880,000 ($50 \times 57{,}600$); adding employment status, sex, prior strikes, and marital status leads to over 69 million parameters. Conventional results—which often impose strong and unwarranted functional form assumptions—can be fragile. When groups differ sharply, regression may not credibly "control" for confounding factors.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

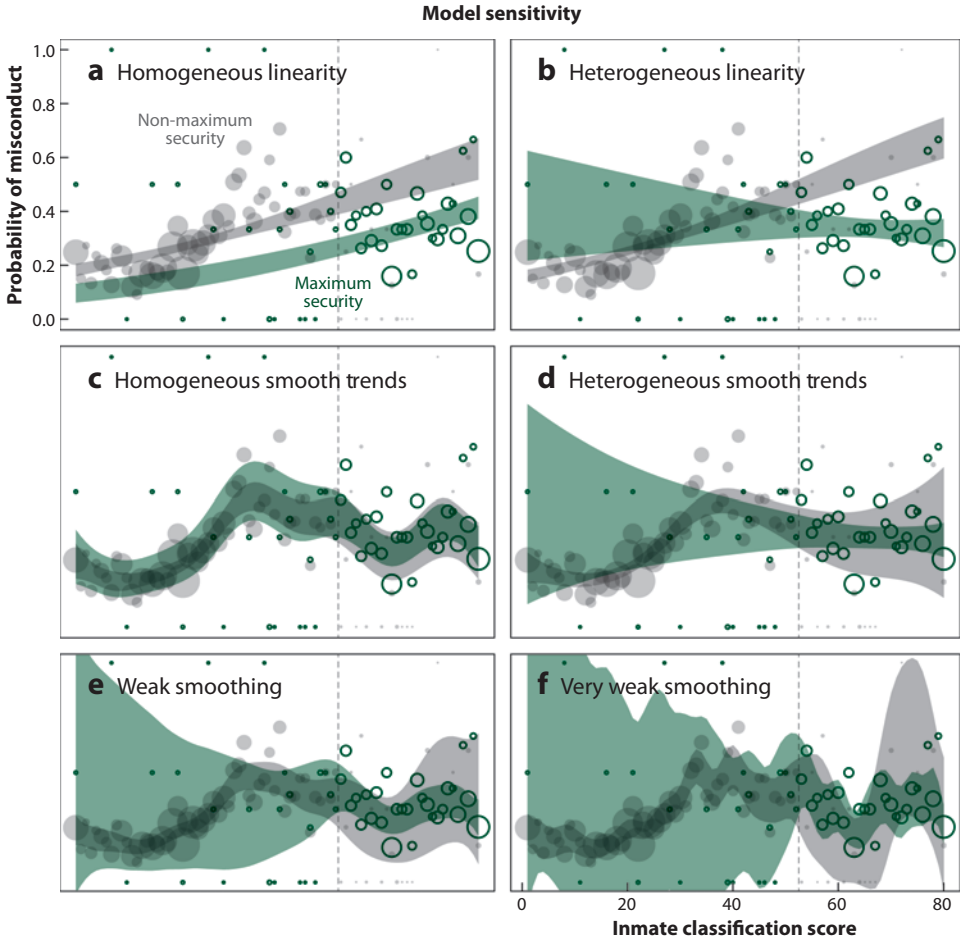

Figure 5

Model sensitivity of regression approaches. Each panel presents the 95% pointwise confidence bands from regression models. Gray bands are for non-maximum-security prisoners, and green bands are for maximum-security prisoners. Panel (*a*) presents the (logit) model of BdL, which assumes linearity and homogeneity across treatment and control groups (in the log odds). Panel (*b*) allows for heterogeneous slopes. Panel (*c*) allows for homogeneous smoothed trend [via a generalized additive model (GAM) (Hastie & Tibshirani 1990)]. Panel (*d*) allows for heterogeneous smoothed trends, and panels (*e*) and (*f*) sequentially decrease smoothness assumptions (by decreasing the GAM's bandwidth and increasing the number of knots). Results are substantially similar as polynomial terms are expanded in the logit model. These panels show how the estimated treatment effect is subject to tremendous model sensitivity.

And regression does not amount to research design.

6. CREDIBLE INFERENCE: DESIGN TRUMPS ANALYSIS

Our central message is that research design trumps methods of analysis. By research design we mean "contemplating, collecting, organizing, and analyzing of data that takes place prior to seeing any outcome data" (Rubin 2008). Methods of analysis, in contrast, involve the development of a model for outcomes (e.g., linear regression, generalized linear models, machine learning algorithms). Just as experiments elaborate a procedure

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

without knowing values of the outcome, observational studies can be designed according to key principles.

First, outcome data should be set aside at the design phase. Classical *p*-values from statistical tests are inappropriate when models are fit multiple times. The possibility of inadvertently choosing a model with a particular result threatens credibility.

Second, the crucial element of design is to use all covariate information to achieve balance along all important pretreatment covariates between treatment and control groups. In Section 7, we show how to balance by matching prisoners on exact classification scores. In Section 8, we note how prevailing practice of regression discontinuity has design and analysis reversed.

Third, the researcher must make a qualitative assessment of the substantive credibility of the "identifying" assumptions. What is the so-called "identification strategy"? Do the data contain enough covariates to make matching credible? Are the covariates properly pretreatment covariates? Are subjects able to manipulate treatment assignment?

There is no substitute for substantive knowledge. Consider a compelling study of the effect of classroom size on educational outcomes by Angrist & Lavy (1999). The study capitalized on "Maimonides's rule" in the Israeli public school system that sets a strict cap on classroom size at 40 students. Because it is plausibly random whether the enrollment at the beginning of the school year is just below or above 40, we can credibly assess the impact of class size by comparing class sizes of 20 and 21 resulting from enrollments of 41 to a class size of 39. Although highly credible in the context of Israeli public schools, in other jurisdictions where parents can switch schools upon discovery of a large classroom, the comparison can be contaminated (Angrist & Pischke 2010, p. 14; Urquiola & Verhoogen 2009). Credibility hence depends on deep, substantive knowledge of the legal system being examined.

7. MATCHING

Matching reduces the role of strong and unwarranted functional form assumptions by trimming the data set down to treatment and control groups that are balanced along pretreatment covariates. The key assumption is that, conditional on covariates, treatment is random. The credibility depends entirely on (*a*) whether enough relevant (pretreatment) covariates have been collected and (*b*) whether sufficient balance has been achieved between treatment and control groups.

In the prison data, is the treatment plausibly random given a specific classification score? How administrative placements are made is crucial here. If sex offenders are the only prisoners with scores below 52 placed in maximum-security prison, we may be effectively comparing the propensity for misconduct of sex offenders and non–sex offenders at low scores. Do CDC officials differ systematically in the use of administrative placements? If so, how are officials assigned? If overcrowding at maximum-security-level prisons results in prisoners with high scores being placed in non-maximum-security prisons, is the timing of overcrowding random or might waves of gang violence explain overcrowding shocks (when gang membership may generally lead to more behavioral misconduct)? To ground the assumptions, substantive knowledge and research are required.

Assuming that given a score prison assignment is random, the best practice is to report how much balance has improved after matching. Matching exactly on classification score solves that problem in our data. In other instances, multiple and continuous covariates can make exact matching impossible, in which case "propensity score" matching provides a way forward. For inexact matches, researchers should do everything to achieve the best balance using substantive knowledge. For example, in a study of a drug's impact on birth defects, matching women on age requires scientific knowledge. A two-year difference between a 21- and a 23-year-old may be trivial, but a one year difference between a

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

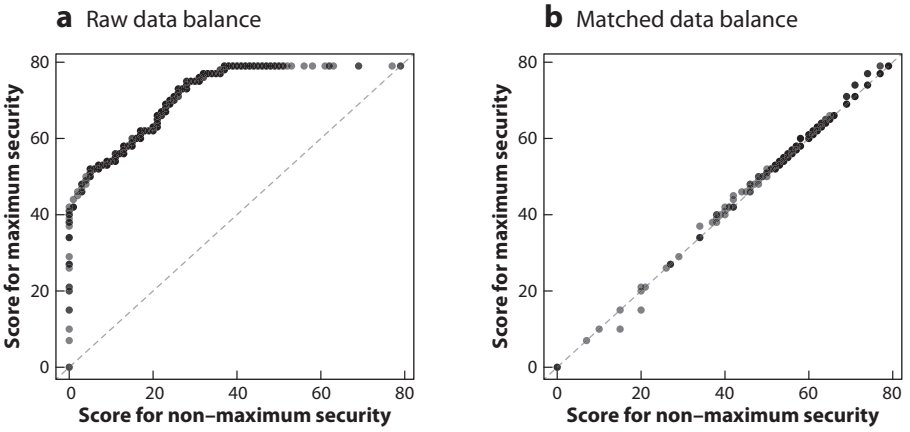

Figure 6

Quantile-quantile plot of balance of inmate classification score between treatment and control groups. Panel (*a*) plots the raw data, showing that maximum-security prisoners have far higher classification scores. Panel (*b*) plots the matched data, showing balance on the classification score (points falling along the 45-degree line). The latter exhibits slight sampling variability by sampling proportional to weights from exact matching.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

41- and a 42-year-old could invalidate the study. Fortunately, legal academics are precisely the ones who harbor the deepest knowledge about the legal system under study and are thus often in the best position to evaluate balance.

Figure 6*a* plots the quantile-quantile plot of the control group on the *x*-axis and the treatment group on the *y*-axis. If there is balance, the dots should line up on the 45-degree line. The raw difference, however, is stark. The right panel presents the same plot for the matched data set. Unsurprisingly, because units are exactly matched, balance is good.

Figure 7 presents the difference in misconduct probability at each classification score where there are both treatment and control units. For example, the leftmost dot represents the 25% difference at score 0 between the 1 of 2 maximum-security prisoners and 25% of 125 non-maximum-security prisoners who engaged in misconduct. The intervals represent 95% confidence intervals, and dots are weighted by sample size of the smallest group, with 1,910 units in the matched sample. These conditional effects show no pattern, and most contain the origin.[4] To calculate an overall effect, we can use a weighted average across these categories (weighted by the number of treated units at each score, represented by the hollow green circles), resulting in an effect estimate of 0.03, plus or minus 0.08 (the gray interval). In other words, comparing inmates with identical classification scores, maximum security causes from a 5% decrease to an 11% increase in misconduct. Although the interval is fairly informative, we cannot reject the null hypothesis that security level has no impact on behavior.

8. REGRESSION DISCONTINUITY

The key assumptions of regression discontinuity (RD) are that (*a*) treatment assignment is *discontinuous* at a threshold of the forcing variable, which cannot be precisely manipulated, and (*b*) all other covariates are *smooth* (or balanced) at the threshold. Under those assumptions, units just below and above the threshold are

[4]The confidence interval is constructed with a χ^2 approximation and, if anything, may be conservative for the small samples at each classification score.

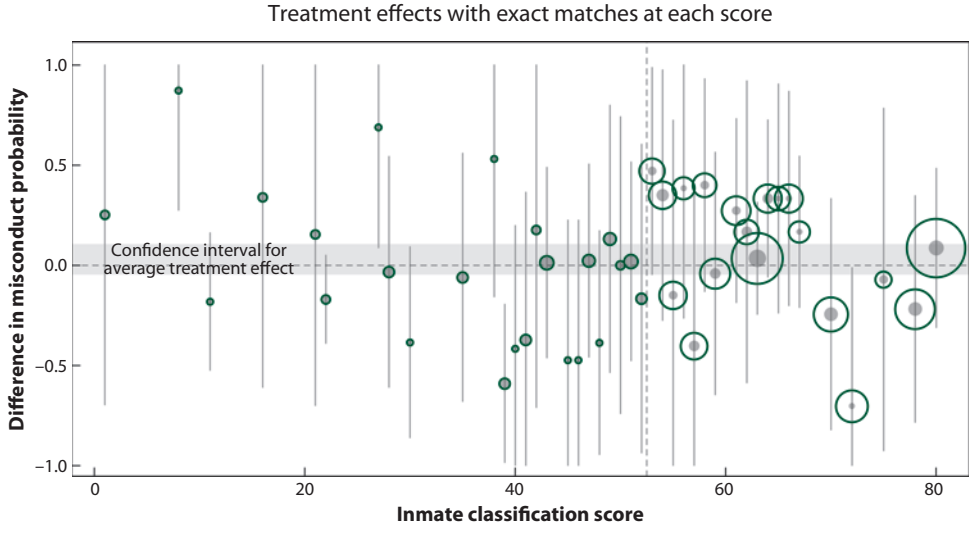

Figure 7

Treatment effects for each classification score where there are both treatment and control units. The gray dots represent the difference in proportions and are proportional to the minimum number of treated or control units at that score. The hollow green circles are proportional to the number of treated units at that score. The vertical gray lines represent 95% confidence intervals. The horizontal light gray band represents the 95% confidence interval of the average treatment effect on the treated, which includes the origin. The vertical gray dashed line represents the treatment threshold.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

plausible comparison groups. Outcome distributions that differ sharply can be attributed to the treatment.

How credible is the discontinuity assumption with the prison data? Prison assignment sharply changes when the classification score reaches 52. The top panel of **Figure 9** (discussed more at length below) shows that at the threshold of 52, the probability of assignment to maximum-security prison jumps from 0.2 to 0.9. Do CDC officials manipulate the intake process to target prisoners based on their potential outcomes? Qualitative assessments of both the intake scoring method and the administrative placements are crucial here. For example, 16 points are added to the score if a prison assault "caused serious injury . . . the extent of which are [*sic*] life threatening in nature and require hospital care or cause disability over an extended period (medical attention beyond first-aid or . . . treatment and release)" (CDC 2000, art. 1, ch. 1, § 61010.11.2). To what degree does this standard permit subjective scoring to place individuals just above or below the threshold based on expected behavior? Similarly, administrative placements may be used to target placement when the score belies expected behavior. Not only are there subjective special case factors (e.g., whether the inmate "has strong family ties to a particular area where other placement would cause an unusual hardship"), but the criteria themselves suggest optimization on potential outcomes (i.e., whether the inmate's "behavior record indicates he or she is capable of successful placement at an institution level lower than that indicated [by the] score") (CDC 2000, art. 1, ch. 1, § 61010.11.3). If so, this precise manipulation invalidates regression discontinuity.

It is important to note here the substantive difference in the identification assumptions between matching and RD. Matching essentially identifies effects using administrative placements. If the classification score were

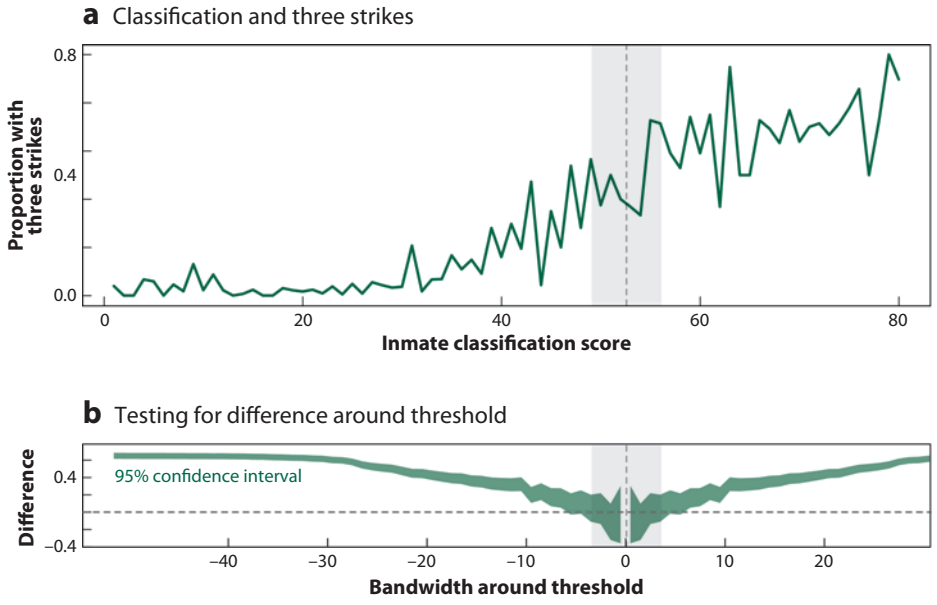

Figure 8

Covariate balance of whether inmate is a three-strike inmate. Panel (*a*) presents the classification score on the *x*-axis and the proportion of inmates with three strikes on the *y*-axis. Panel (*b*) plots the confidence interval of the difference in proportions by varying the bandwidth around the threshold of a score of 52 points (*vertical gray dashed line*). The larger the bandwidth, the sharper the discontinuity of proportion of three-strike inmates. The vertical light gray bands represent the bandwidth range for which there is relative balance of the third strike covariate.

used deterministically, there would be no overlap of maximum- and non-maximum-security prisoners at a given score. On the other hand, RD identifies effects using the arbitrariness of scoring just above or just below 52 points. Both approaches could gain credibility with more covariates (such as those in **Table 1**).

Assuming that the treatment cannot be manipulated precisely, conventional RD practice might be to fit numerous regressions to the outcome data (varying polynomial terms, the covariate set, and bandwidths). This ignores two crucial issues. First, it ignores covariate balance. If the design is right, balance should be verified (and the appropriate bandwidth chosen) *prior* to examination of any outcome data. Second, it imports all the problems of model sensitivity before implementing research design (Rubin 1977).

Conventional practice, in that sense, has the process reversed. Research design (and covariate balance) should be implemented before any analysis. The primary goal in design is to determine the bandwidth around the threshold that results in comparable groups, without resorting to outcomes (cf. Lee & Lemieux 2010).

To illustrate the design phase, **Figure 8*a*** shows that the proportion of three-strike inmates increases with the classification score. An increase in the score from 40 to 50 is associated on average with a roughly 12% increase in the probability of a third strike. Moreover, there is a sharp spike from 33% to 72% third strikes at scores 54 to 55. Designs that fail to account for this discontinuity may falsely attribute differences in misconduct to security facility. At the design phase, all substantive knowledge should be used to determine the appropriate bandwidth just below and above the

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

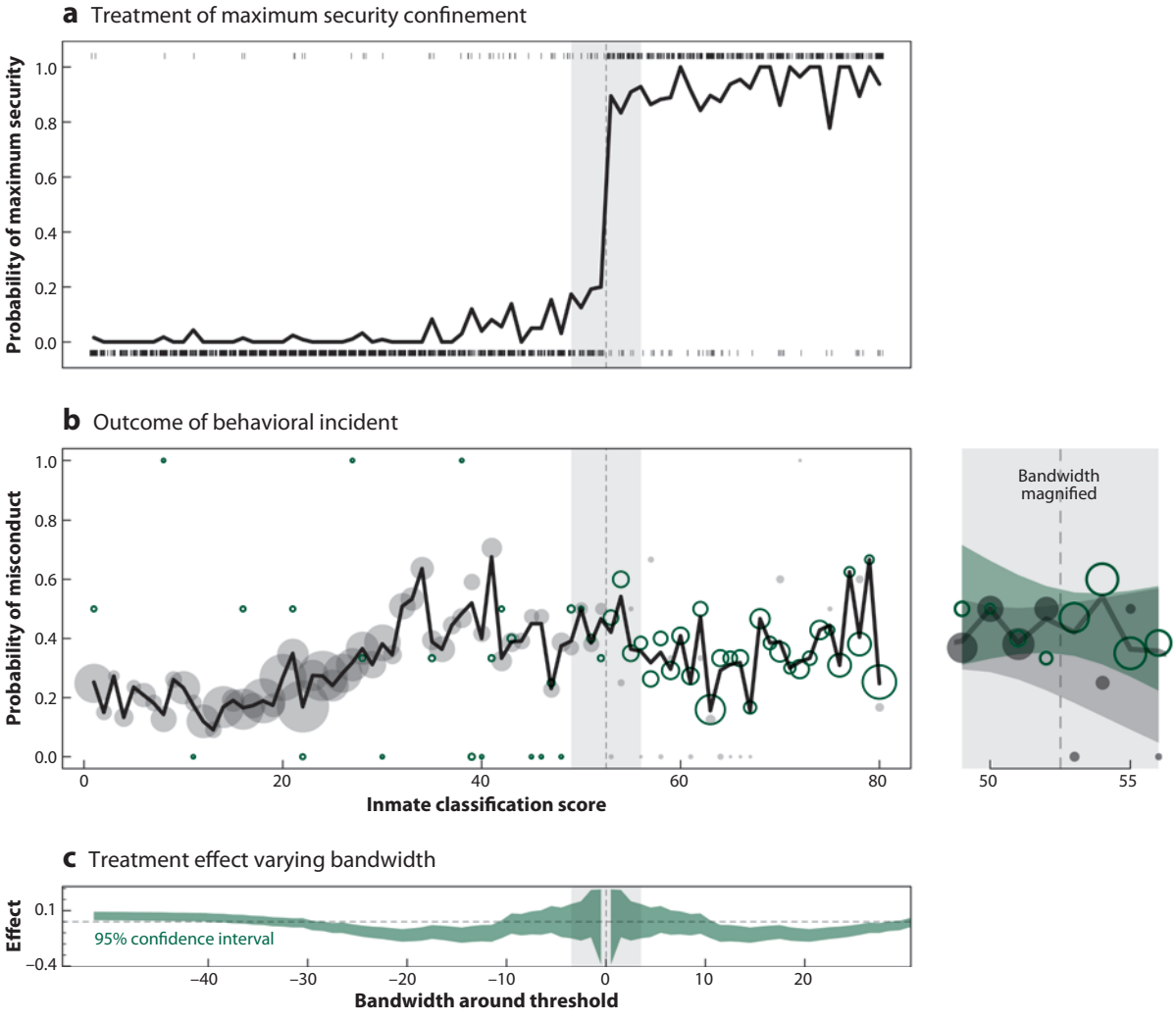

Figure 9

Treatment discontinuity and outcome continuity. Panel (*a*) plots the probability of maximum-security confinement (treatment) on the *y*-axis against inmate classification score on the *x*-axis. Short vertical lines represent each data point (randomly jittered for visibility). Panel (*b*) presents the probability of misconduct (outcome) on the *y*-axis against inmate classification score on the *x*-axis. Panel (*c*) plots the 95% confidence interval of the treatment effect (more precisely, the "intention to treat" effect), varying the bandwidth around the threshold of 52 points.

threshold. In that sense, matching and regression discontinuity are comparable.[5] **Figure 8*b*** plots pointwise 95% confidence intervals as the bandwidth is expanded around the threshold (hence symmetric around the threshold). Just as in matching, a classic bias-variance trade-off exists: the narrower the bandwidth, the lower the bias, but the higher the variability due to sample

[5]Compare Heckman et al. (1999, p. 1969) (noting that "[r]egression discontinuity estimators constitute a special case of 'selection on observables'") with Lee & Lemieux (2010, p. 291) (positing that "RD design is more closely related to randomized experiments than to . . . matching"). We think the relative credibility depends on the application.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

size. The vertical gray bands choose one plausible bandwidth from scores of 49 to 56, where the confidence interval includes the origin. In practice, the design phase should develop this bandwidth by examining balance across all important covariates using substantive knowledge.

Having achieved balance on the key covariates, we may examine the outcome data. Our analysis is straightforward. **Figure 9*b*** plots the score against the proportion of behavioral incidents. Contrast the discontinuity of the treatment with the continuity of the outcome at the threshold. A sharp jump in the outcome would have been evidence of a treatment effect, but no perceptible change occurs at the threshold. Based on this estimate, the causal effect is indistinguishable from zero, with a wide 95% interval from −17% to 27%.

The vertical gray bands overlay the bandwidth chosen based on the third strike covariate. Within the bandwidth, a (local logistic) regression can be used to adjust for remaining imbalance. The **Figure 9*b*** inset magnifies the bandwidth range and overlays pointwise 95% confidence bands, showing that there is little evidence of any treatment effect. Based on this regression adjustment, the overall 95% interval of the causal effect contains the origin, ranging from −10% to 29%. Lastly, **Figure 9*c*** plots the 95% confidence intervals of treatment effects varying the bandwidth. As the bandwidth increases, the interval converges to the simple raw difference in means reported in **Table 2**. In sum, the RD design reveals no evidence of a treatment effect.

9. COMPARISON TO EXPERIMENTAL RESULTS

How do these methods compare to a randomized experiment? For most social science applications, few such validations exist (but cf. Dehejia & Wahba 1999; Lalonde 1986; Heckman et al. 1998a,b). Fortunately, Berk et al. (2003) performed a valuable field experiment in cooperation with the CDC, allowing for potential validation of observational approaches. Berk et al. (2003) randomized inmates to the existing intake procedure and to a new proposed one, in which the latter increased the security level for some inmates. This randomization thus allows us to identify the causal effect of security level on the subgroup of inmates whose assignment was affected by the difference in protocols.

To be sure, the field experiment was limited in certain respects. First, the randomization was over the intake procedure, not the treatment of security level. Second, the experimental intake procedure generally increased security from low to medium levels, with little effect on maximum-security confinement. The experiment is hence uninformative about the effect of maximum security per se. Therefore, we focus on (*a*) the overall (so-called "intention-to-treat") effect of the experimental intake procedure, essentially comparing misconduct rates between all inmates scored on old and new intake procedures, and (*b*) the subgroup effect on inmates whose assignment was in fact affected by randomization (Camp & Gaes 2005). These effects may, of course, diverge from the population treatment effect.[6]

Figure 10 presents point estimates and 95% confidence intervals from the various observational approaches and the experiment. The first line presents naive (logit) regression estimates from BdL and Section 5 in green. The second line presents estimates from exact matching in Section 7. The third line presents estimates from regression discontinuity, either using the simple difference in means within the bandwidth or a model-based adjustment within the bandwidth from Section 8. The overall experimental estimate in the fifth line is −1%, plus or minus 1%, and the subgroup effect (of facilities with individual cells versus open dormitories) in the sixth line is −4%, plus or minus 8%, both

[6]If there are heterogeneous treatment effects, then matching, regression discontinuity, and the experiment may properly identify effects for the subsets of (*a*) inmates affected by administrative placements and population overrides, (*b*) inmates just above and below the classification score of 52, and (*c*) inmates for whom the experimental procedure changed ultimate placement, respectively, but these effects could nonetheless diverge.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

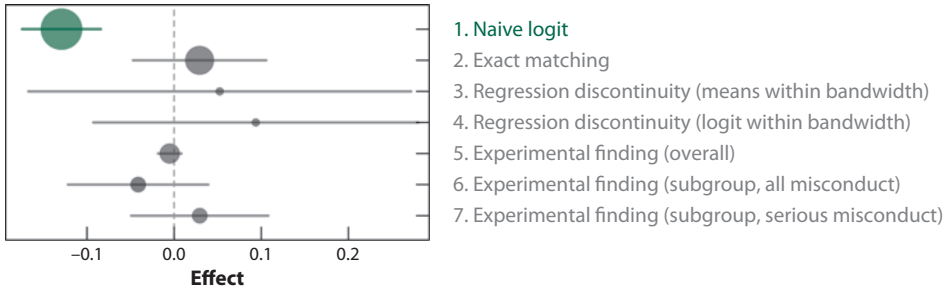

Figure 10

Comparison of observational approaches to experimental findings. Dots plot point estimates (weighted by effective sample size), and lines represent 95% confidence intervals. The naive logit (*green*) is the overall average treatment effect (based on asymptotic posterior simulation) of the regression reported in Berk & de Leeuw (1999, p. 1048, table 1, model 1) and Section 5. Exact matching is the average treatment effect on the treated, based on the weighted difference of exact matches on the classification score. Regression discontinuity is the intention-to-treat effect, either (*a*) the mean difference between subjects above and below the threshold within the bandwidth or (*b*) the treatment effect (based on asymptotic posterior simulation) of the logit regression within the bandwidth. The overall experimental findings are calculated based on sample sizes and effects from Berk et al. (2003, p. 228, table 1, p. 232). The subgroup experimental findings are from Camp & Gaes (2005, pp. 434, 436, tables 1 and 2 therein).

indistinguishable from 0. The last line presents experimental subgroup estimates of an increase of 3%, plus or minus 8%, on serious misconduct. Naive regression-based estimates, finding a reduction of 13% (plus or minus 4%), deviate considerably from experimental estimates.[7] Intervals from matching, regression discontinuity, and the experiment, on the other hand, all contain the origin.

As a last comparison, Bench & Allen (2003) randomized 200 inmates to maximum- or medium-security prisons in Utah. Although the measurement of misbehavior differs, the study is perhaps closest to the ideal experiment. It found that "there is no meaningful difference in the number of . . . disciplinaries" between treated and control groups (Bench & Allen 2003, p. 377).

In the end, neither the California nor the Utah study is a gold-standard experiment. In Utah, 10% of the inmates changed security assignments midstream. In California, randomization was not over the treatment of interest. The distinction between experiments and observational studies is one of degree, not of kind. A well-designed observational study, in which fluctuations in the availability of prison beds, for example, affect inmate placement, can be more informative than a broken experiment. The comparisons of **Figure 10** provide considerable evidence that quasi-experimental approaches, by reducing the role of unwarranted functional form assumptions, are more likely to recover the true causal effect.

10. CONCLUSION: A RETURN TO MOORE

Causal inference is hard. As we have reviewed, recent advances should allow researchers to more *credibly* assess the impact of legal institutions. Once designs are stripped of technical garb, research should empower the broader legal academic community—precisely the community with the comparative advantage—to assess the credibility of the inference.

We reiterate the important lessons, if in pithy format:

[7]To address the causal effect in the experiment, Berk et al. (2003) also estimate analogous (logit) regressions within treatment and control arms (pp. 234–35, tables 4 and 5 therein), acknowledging that "any such analysis must be interpreted with caution [as] there was no random assignment to security level" (p. 235).

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

1. Conceptualize the experimental template.
2. Design research with outcomes last.
3. Collect and balance covariates.
4. Visualize the data.

Although formalization of the approaches we have discussed is relatively recent, in one way the emphasis on research design calls for a return to the early legal empiricist Underhill Moore. True, his parking studies had no standard errors, failed to assess pre- and post-time trends around the threshold, and did not employ conventional models for analysis. But these are second order. Moore got design. Credible design occurs prior to outcomes and, as the BdL data show, can require deep substantive knowledge of the law itself. In that sense, Moore's law of parking was ahead of the curve, if not the curb.

APPENDIX: WHERE TO GO FROM HERE

Our exposition here is only an informal review of a vast, rapidly growing, and sophisticated literature. In this Appendix, we provide some guidance on where researchers can go to study these approaches in greater depth.

Angrist & Krueger (1999), Angrist & Pischke (2008), Morgan & Winship (2007), and Rosenbaum (2002) provide general overviews of program evaluation and causal inference. For a complementary approach that relies on graphical models, see Pearl (2000).

Heckman et al. (1998b), Ho et al. (2007), Imbens (2004), and Stuart & Rubin (2008) provide overviews of matching methods (see also Rubin 2006). Software implementations can be found in R (Iacus et al. 2009b, Hansen & Fredrickson 2010, Ho et al. 2004, Sekhon 2011) and Stata (Abadie et al. 2001, Becker & Ichino 2002). Imbens (2000) and Joffe & Rosenbaum (1999) develop extensions of matching for categorical treatments, and Imai & van Dyk (2004) and Hirano & Imbens (2004) discuss generalizations for continuous treatments. For alternative matching approaches, see Iacus et al. (2009a), Rosenbaum (2002), Abadie & Gardeazabal (2003), and Hansen (2004).

Imbens & Lemieux (2008) and Lee & Lemieux (2010) provide general overviews of regression discontinuity (RD) (see also Thistlethwaite & Campbell 1960). Hahn et al. (2001) formalize the conditions under which RD provides an unbiased estimate of causal effects. McCrary (2008) develops a useful test for manipulation of the forcing variable.

For developments of sensitivity and bounds analyses, see Manski (1990, 1995) and Rosenbaum & Rubin (1983a). For randomization inference, see Imbens & Rosenbaum (2005), Ho & Imai (2006), and Donohue & Ho (2007).

There are, of course, many examples of panel approaches to assessing the impact of law. For examples of difference-in-differences, see Card & Krueger (1994), Autor et al. (2004), and Rubinfeld (2010). Bertrand et al. (2004) make a crucial point about variance estimation in assessing one-time policy interventions.

For an interpretation of instrumental variables from a potential outcomes perspective, see Angrist et al. (1996). A generalized framework is that of principal stratification (Barnard et al. 2003, Frangakis & Rubin 2002, Hirano et al. 2000).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Shira Oyserman, Xiangnong Wang, and Olga Zverovich for terrific research assistance, Gail Long at the California Department of Corrections and Rehabilitation for help with the

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

history of inmate classification, George Wilson at the Stanford Law Library for help in tracking down CDC operations manuals, Stephen Galoob for comments, and Richard Berk and Jan de Leeuw for sharing data.

LITERATURE CITED

Abadie A, Drukker D, Herr JL, Imbens GW. 2001. Matching estimators. *Statistical Software.* **http://www.economics.harvard.edu/faculty/imbens/software_imbens**

Abadie A, Gardeazabal J. 2003. The economic costs of conflict: a case study of the Basque country. *Am. Econ. Rev.* 93:113–32

Abrams DS, Yoon AH. 2007. The luck of the draw: using random case assignment to investigate attorney ability. *Univ. Chicago Law Rev.* 74:1145–77

Angrist JD. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *Am. Econ. Rev.* 80:1284–86

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables (with discussion). *J. Am. Stat. Assoc.* 91:444–55

Angrist JD, Krueger AB. 1999. Empirical strategies in labor economics. In *Handbook of Labor Economics*, ed. O Ashenfelter, D Card, 3A:1277–366. Amsterdam: Elsevier

Angrist JD, Lavy V. 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Q. J. Econ.* 114:533–75

Angrist JD, Pischke J-S. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ. Press

Angrist JD, Pischke J-S. 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *J. Econ. Perspect.* 24:3–30

Autor DH, Donohue JJ III, Schwab SJ. 2004. The costs of wrongful-discharge laws: large, small, or none at all? *Am. Econ. Rev.* 94:440–46

Ayres I. 1991. Fair driving: gender and race discrimination in retail car negotiations. *Harvard Law Rev.* 104:817–72

Ayres I. 2008. *Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart*. New York: Bantam

Barnard J, Frangakis CE, Hill JL, Rubin DB. 2003. Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York (with discussion). *J. Am. Stat. Assoc.* 98:299–311

Bayer P, Hjalmarsson R, Pozen D. 2009. Building criminal capital behind bars: peer effects in juvenile corrections. *Q. J. Econ.* 124:105–47

Becker SO, Ichino A. 2002. Stata programs for ATT estimation based on propensity score matching. *Statistical Software*. **http://www.lrz.de/~sobecker/pscore.html**

Bench LL, Allen TD. 2003. Investigating the stigma of prison classification: an experimental design. *Prison J.* 83:367–82

Berk RA. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage

Berk RA, de Leeuw J. 1999. An evaluation of California's inmate classification system using a generalized regression discontinuity design. *J. Am. Stat. Assoc.* 94:1045–52

Berk RA, Ladd H, Graziano H, Baek JH. 2003. A randomized experiment testing inmate classification systems. *Criminol. Public Policy* 2:215–42

Berk RA, Newton PJ. 1985. Does arrest really deter wife battery? An effort to replicate the findings of the Minneapolis Spouse Abuse Experiment. *Am. Sociol. Rev.* 50:253–62

Berry CR, Lee SL. 2007. *The Community Reinvestment Act: a regression discontinuity analysis*. Harris Sch. Work. Pap. Ser. 07.04, Univ. Chicago, Chicago, IL

Bertrand M, Duflo E, Mullainathan S. 2004. How much should we trust differences-in-differences estimates? *Q. J. Econ.* 119:249–75

Black BS, Kim W, Jang H, Park KS. 2008. *How corporate governance affects firm value: evidence on channels from Korea*. ECGI Fin. Work. Pap. No. 103/2005. **http://ssrn.com/abstract=844744**

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Boyd CL, Epstein L, Martin AD. 2010. Untangling the causal effects of sex on judging. *Am. J. Polit. Sci.* 54:389–411

Brady HE, McNulty JE. 2007. *The costs of voting: disruption and transportation effects*. Presented at Midwest Polit. Sci. Assoc., Chicago, IL, April 12. **http://www.can-so.org/vote/polling-locations-and-transportation-effects.pdf**

Bubb R. 2009. *States, law, and property rights in West Africa*. Work. Pap., Dep. Econ., Harvard Univ. **http://isites.harvard.edu/fs/docs/icb.topic637140.files/Bubb_StatesLawProperty.pdf**

Calif. Dep. Correct. (CDC). 2000. *Operations Manual*. Sacramento: State Calif.

Camp SD, Gaes GG. 2005. Criminogenic effects of the prison environment on inmate behavior: some experimental evidence. *Crime Delinquency* 51:425–42

Card D, Dobkin C, Maestas N. 2008. The impact of nearly universal insurance coverage on health care: evidence from Medicare. *Am. Econ. Rev.* 98:2242–58

Card DE, Krueger AB. 1994. Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania. *Am. Econ. Rev.* 84:772–84

Chay KY, Greenstone M. 2005. Does air quality matter? Evidence from the housing market. *J. Polit. Econ.* 113:376–424

Chen MK, Shapiro JM. 2007. Do harsher prison conditions reduce recidivism? A discontinuity-based approach. *Am. Law Econ. Rev.* 9:1–29

Clark W, Douglas WO, Thomas DS. 1930. The business failures project—a problem in methodology. *Yale Law J.* 39:1013–24

Dehejia RH, Wahba S. 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J. Am. Stat. Assoc.* 94:1053–62

Dehejia RH, Wahba S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* 84:151–61

DiNardo J, Lee DS. 2004. Economic impacts of new unionization on private sector employers: 1984–2001. *Q. J. Econ.* 119:1383–441

Donohue JJ III, Ho DE. 2007. The impact of damage caps on malpractice claims: randomization inference with difference-in-differences. *J. Empir. Legal Stud.* 4:69–102

Donohue JJ III, Wolfers J. 2006. Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Rev.* 58:791–845

Douglas WO. 1950. Underhill Moore. *Yale Law J.* 59:187–88

Eggers AC, Hainmueller J. 2009. Mps for sale? Returns to office in postwar British politics. *Am. Polit. Sci. Rev.* 103:513–33

Epstein L, Ho DE, King G, Segal JA. 2005. The Supreme Court during crisis: how war affects only non-war cases. *N. Y. Univ. Law Rev.* 80:1–116

Epstein L, King G. 2002. The rules of inference. *Univ. Chicago Law Rev.* 69:1–133

Fisher RA. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd

Fisher RA. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd

Frangakis CE, Rubin DB. 2002. Principal stratification in causal inference. *Biometrics* 58:21–29

Galiani S, Gertler P, Schargrodsky E. 2005. Water for life: the impact of the privatization of water services on child mortality. *J. Polit. Econ.* 113:83–120

Gelman A, Meng X-L, eds. 2004. *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives*. Hoboken, NJ: Wiley

Gerber A, Kessler DP, Meredith MN. 2008. *The persuasive effects of direct mail: a regression discontinuity approach*. NBER Work. Pap. 14206, Natl. Bur. Econ. Res., Cambridge, MA. **http://www.nber.org/papers/w14206.pdf**

Gerber AS, Green DP. 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment. *Am. Polit. Sci. Rev.* 94:653–63

Gibson JL. 2008. Challenges to the impartiality of state supreme courts: legitimacy theory and 'new-style' judicial campaigns. *Am. Polit. Sci. Rev.* 102:59–75

Green DP, Winik D. 2010. Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. *Criminology* 48:357–87

Greiner DJ. 2008. Causal inference in civil rights litigation. *Harvard Law Rev.* 122:533–98
Greiner DJ, Rubin DB. 2010. Causal effects of perceived immutable characteristics. *Rev. Econ. Stat.* In press
Grogger J, Ridgeway G. 2006. Testing for racial profiling in traffic stops from behind a veil of darkness. *J. Am. Stat. Assoc.* 101:878–87
Guttentag MD, Porath CL, Fraidin SN. 2008. Brandeis' policeman: results from a laboratory experiment on how to prevent corporate fraud. *J. Empir. Legal Stud.* 5:239–73
Hahn J, Todd P, Kasarda JD. 1999. *Evaluating the effect of an antidiscrimination law using a regression-discontinuity design*. NBER Work. Pap. 7131, Natl. Bur. Econ. Res., Cambridge, MA. **http://www.nber.org/papers/w7131.pdf**
Hahn J, Todd P, van der Klaauw W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69:201–209
Hansen BB. 2004. Full matching in an observational study of coaching for the SAT. *J. Am. Stat. Assoc.* 99:609–18
Hansen BB, Fredrickson M. 2010. optmatch: functions for optimal matching. *Statistical Software* **http://cran.r-project.org/web/packages/optmatch/index.html**
Hastie TJ, Tibshirani R. 1990. *Generalized Additive Models*. London: Chapman Hall
Heckman J, Ichimura H, Smith J, Todd P. 1998a. Characterizing selection bias using experimental data. *Econometrica* 66:1017–98
Heckman JJ, Ichimura H, Todd P. 1998b. Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* 65:261–94
Heckman JJ, Lalonde RJ, Smith JA. 1999. The economics and econometrics of active labor market programs. In *Handbook of Labor Economics*, ed. O Ashenfelter, D Card, 3A:1865–2097. Amsterdam: Elsevier
Helland E, Tabarrok A. 2004. The fugitive: evidence on public versus private law enforcement from bail jumping. *J. Law Econ.* 47:93–122
Hirano K, Imbens GW. 2004. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. A Gelman, X-L Meng, pp. 73–83. Hoboken, NJ: Wiley
Hirano K, Imbens GW, Rubin DB, Zhou XH. 2000. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 1:69–88
Hjalmarsson R. 2009a. Crime and expected punishment: changes in perceptions at the age of criminal majority. *Am. Law Econ. Rev.* 11:209–48
Hjalmarsson R. 2009b. Juvenile jails: a path to the straight and narrow or to hardened criminality? *J. Law Econ.* 52:779–809
Ho DE. 2005a. Affirmative action's affirmative actions: a reply to Sander. *Yale Law J.* 114:2011–16
Ho DE. 2005b. Why affirmative action does not cause black students to fail the bar. *Yale Law J.* 114:1997–2004
Ho DE, Imai K. 2006. Randomization inference with natural experiments: an analysis of ballot effects in the 2003 California recall election. *J. Am. Stat. Assoc.* 101:888–900
Ho DE, Imai K, King G, Stuart EA. 2004. MatchIt: nonparametric preprocessing for parametric causal inference. *Statistical Software*. **http://gking.harvard.edu/matchit**
Ho DE, Imai K, King G, Stuart EA. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* 15:199–236
Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81:945–60
Hopkins DJ. 2009. *Language access and initiative outcomes: Did the Voting Rights Act influence support for bilingual education?* CELS 2009 4th Annu. Conf. Empir. Legal Stud. Pap. **http://ssrn.com/abstract=1434374**.
Iacus SM, King G, Porro G. 2009a. Causal inference without balance checking: coarsened exact matching. *Polit. Anal.* In press. **http://gking.harvard.edu/gking/files/cem-plus.pdf**
Iacus SM, King G, Porro G. 2009b. CEM: coarsened exact matching software. *Statistical Software*. **http://gking.harvard.edu/cem**
Imai K, van Dyk DA. 2004. Causal inference with general treatment regimes: generalizing the propensity score. *J. Am. Stat. Assoc. Theory Methods* 99:854–66
Imbens GW. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87:706–10

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Imbens GW. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Stat.* 86:4–29
Imbens GW, Lemieux T. 2008. Regression discontinuity designs: a guide to practice. *J. Econ.* 142:615–35
Imbens GW, Rosenbaum PR. 2005. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *J. R. Stat. Soc. Ser. A* 168:109–26
Joffe MM, Rosenbaum PR. 1999. Propensity scores. *Am. J. Epidemiol.* 150:327–33
Kane TJ, Riegg SK, Staiger DO. 2006. School quality, neighborhoods, and housing prices. *Am. Law Econ. Rev.* 8:183–212
King G, Gakidou E, Ravishankar N, Moore RT, Lakin J, et al. 2007. A 'politically robust' experimental design for public policy evaluation, with application to the Mexican universal health insurance program. *J. Policy Anal. Manag.* 26:479–506
King G, Zeng L. 2007. When can history be our guide? The pitfalls of counterfactual inference. *Int. Stud. Q.* 51:183–210
Kritzer HM. 2010. The (nearly) forgotten early empirical legal research. In *Oxford Handbook of Empirical Legal Research*, ed. P Cane, HM Kritzer, pp. 875–900. Oxford: Oxford Univ. Press
Lalive R. 2008. How do extended benefits affect unemployment duration? A regression discontinuity approach. *J. Econom.* 142:785–806
Lalonde R. 1986. Evaluating the econometric evaluations of training programs. *Am. Econ. Rev.* 76:604–20
Leamer EE. 1978. *Specification Searches*. New York: John Wiley & Sons
Leamer EE. 1983. Let's take the con out of econometrics. *Am. Econ. Rev.* 73:31–43
Lee DS. 2008. Randomized experiments from non-random selection in U.S. house elections. *J. Econom.* 142:675–97
Lee DS, Lemieux T. 2010. Regression discontinuity designs in economics. *J. Econ. Lit.* 48:281–355
Lee DS, McCrary J. 2005. *Crime, punishment, and myopia*. NBER Work. Pap. 11491, Natl. Bur. Econ. Res., Cambridge, MA. **http://www.nber.org/papers/w11491**
Lemieux T, Milligan K. 2008. Incentive effects of social assistance: a regression discontinuity approach. *J. Econom.* 142:807–28
List JA, Margolis M, Osgood DE. 2006. *Is the Endangered Species Act endangering species?* NBER Work. Pap. 12777, Natl. Bur. Econ. Res., Cambridge, MA. **http://www.nber.org/papers/w12777**
Listokin Y. 2008. Management always wins the close ones. *Am. Law Econ. Rev.* 10:159–84
Listokin Y. 2009. Corporate voting versus market price setting. *Am. Law Econ. Rev.* 11:608–35
Litvak K. 2007. Sarbanes-Oxley and the cross-listing premium. *Mich. Law Rev.* 105:1857–98
Ludwig J, Miller DL. 2007. Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Q. J. Econ.* 122:159–208
Manski CF. 1990. Nonparametric bounds on treatment effects. *Am. Econ. Rev. Papers Proc.* 80:319–23
Manski CF. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard Univ. Press
McCrary J. 2008. Manipulation of the running variable in the regression discontinuity design: a density test. *J. Econom.* 142:698–714
Mocan NH, Tekin E. 2006. Catholic schools and bad behavior: a propensity score matching analysis. *B.E. J. Econ. Anal. Policy* 5(1):13. **http://www.bepress.com/bejeap/contributions/vol5/iss1/art13**
Moore U, Callahan CC. 1943. Law and learning theory: a study in legal control. *Yale Law J.* 53:1–136
Morantz AD. 2010. *Coal mining safety: Do unions make a difference*? Stanford Law Econ. Olin Work. Pap. No. 413. **http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1846700**
Morgan SL, Winship C. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge Univ. Press
Pager D. 2003. The mark of a criminal record. *Am. J. Sociol.* 108:937–75
Papachristos AV, Meares TL, Fagan J. 2007. Attention felons: evaluating Project Safe Neighborhoods in Chicago. *J. Empir. Legal Stud.* 4:223–72
Pearl J. 2000. *Causality*. New York: Cambridge Univ. Press
Persson T, Tabellini G. 2002. Do constitutions cause large governments? Quasi-experimental evidence. *Eur. Econ. Rev.* 46:908–18
Petersilia J. 2008. California's correctional paradox of excess and deprivation. *Crime Justice* 37:207–78

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Petersilia J, Turner S, Peterson J. 1986. *Prison versus Probation in California*. Santa Monica, CA: RAND
Pfaff JF. 2010. *A plea for more aggregation: the looming threat to empirical legal scholarship*. SSRN Work. Pap., July 16. **http://ssrn.com/abstract=1641435**
Qian Y. 2007. Do national patent laws stimulate domestic innovation in a global patenting environment? A cross-country analysis of pharmaceutical patent protection, 1978–2002. *Rev. Econ. Stat.* 89:436–53
Ridgeway G. 2006. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *J. Quant. Criminol.* 22:1–29
Rosenbaum PR. 2002. *Observational Studies*. New York: Springer. 2nd ed.
Rosenbaum PR, Rubin DB. 1983a. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B Methodol.* 45:212–18
Rosenbaum PR, Rubin DB. 1983b. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79:516–24
Rubin DB. 1973. Matching to remove bias in observational studies. *Biometrics* 29:159–83
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701
Rubin DB. 1975. Bayesian inference for causality: the importance of randomization. *Proc. Soc. Stat. Sect. Am. Stat. Assoc.*, pp. 233–39. Alexandria, VA: Am. Stat. Assoc.
Rubin DB. 1976. Multivariate matching methods that are equal percent bias reducing. I: Some examples. *Biometrics* 32:109–20
Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *J. Educ. Behav. Stat.* 2:1–26
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58
Rubin DB. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Am. Stat. Assoc.* 74:318–28
Rubin DB. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge Univ. Press
Rubin DB. 2008. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* 2:808–40
Rubinfeld DL. 2010. Econometric issues in antitrust analysis. *J. Inst. Theor. Econ.* 166:62–77
Schlegel JH. 1995. *American Legal Realism and Empirical Social Science*. Chapel Hill: Univ. N. C. Press
Sekhon JS. 2009. Opiates for the matches: matching methods for causal inference. *Annu. Rev. Polit. Sci.* 12:487–508
Sekhon JS. 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J. Stat. Softw.* In press. **http://sekhon.berkeley.edu/matching**
Sobel ME. 2000. Causal inference in the social sciences. *J. Am. Stat. Assoc.* 95:647–51
Stigler SM. 1977. Eight centuries of sampling inspection: the trial of the Pyx. *J. Am. Stat. Assoc.* 72:493–500
Strnad J. 2007. Should legal empiricists go Bayesian? *Am. Law Econ. Rev.* 9:195–303
Stuart EA, Rubin DB. 2008. Best practices in quasi-experimental designs. In *Best Practices in Quantitative Methods*, ed. JW Osborne, pp. 155–76. Thousand Oaks, CA: Sage
Thistlethwaite DL, Campbell DT. 1960. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J. Educ. Psychol.* 51:309–17
Urquiola M, Verhoogen E. 2009. Class-size caps, sorting, and the regression-discontinuity design. *Am. Econ. Rev.* 99:179–215
van der Klaauw W. 2002. Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. *Int. Econ. Rev.* 43:1249–87
Zimring FE, Hawkins G. 1973. *Deterrence: The Legal Threat in Crime Control*. Chicago: Univ. Chicago Press

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Annual Review of Law and Social Science

Volume 7, 2011

Contents

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.

Indexes

Errata

An online log of corrections to *Annual Review of Law and Social Science* articles may be found at http://lawsocsci.annualreviews.org

Annu. Rev. Law. Soc. Sci. 2011.7:17-40. Downloaded from www.annualreviews.org
by Stanford University - Main Campus - Lane Medical Library on 12/14/12. For personal use only.