10-8-2014

Empirical Law and Economics

Jonah B. Gelbach
University of Pennsylvania Law School, jgelbach@law.upenn.edu

Jonathan Klick
University of Pennsylvania Law School, jklick@law.upenn.edu

Empirical Law and Economics

Jonah B. Gelbach
Jonathan Klick

I. Introduction

"[M]ost law professors regard empirical work as a form of drudgery not worthy of first-class minds[1]," Bill Landes lamented in 2003. He made this observation as part of an explanation for the dearth of empirical law and economics being done in law schools at the time. Although he noted that much of the work done by people in economics departments that could be considered law and economics[2] was empirical in nature, factors affecting both supply and demand were to continue to stunt the development of empirical work in law schools, where law and economics (at least in its theoretical flavor) arguably has had its largest impact.[3]

In hindsight, it appears as though Landes was wrong about the future of empirical work in law and economics. The top journals in the field virtually all saw an increase in the amount of empirical work published,[4] most of the top law schools hired faculty whose main methodological approach is econometric, and even legal scholars outside of law and economics began focusing on empirical work.[5]

The move toward more empirical work is likely the result of a number of different causes. As fields mature, it is only natural for there to be an increase in the desire to test earlier hypotheses. Beyond this fairly standard development, as indicated by the growth of empirical work in legal scholarship even outside the field of law and economics, other factors, including increased computing power, the proliferation of electronic datasets, and more empirical training among the younger scholars in this period[6] probably contribute to the rise of empirical work generally.

[1] William M. Landes (2003). "The Empirical Side of Law & Economics." University of Chicago Law Review, 70(1): 167-180, 180.

[2] Interestingly, none of the three individuals Landes uses as examples, Ed Glaeser, Steve Levitt, and Andrei Shleifer, self identifies law and economics as one of his fields and none regularly teach a law and economics course. While law and economics subjects can be found throughout (at least micro) economic scholarship, there is little focus on developing any deep understanding of the way legal institutions actually function.

[3] Anthony Kronman, later to become the dean of the Yale Law School, declared that "In the years since 1965, no other approach to the study of law has had a comparable effect on the way that academic lawyers write and teach." Anthony T. Kronman, The Lost Lawyer: Failing Ideals of the Legal Profession (1993), p. 166.

[4] See Jonathan Klick, The Empirical Revolution in Law and Economics: Inaugural Lecture for Erasmus Chair in Empirical Legal Studies (2011).

[5] See Daniel E. Ho and Larry Kramer (2013), Introduction: The Empirical Revolution in Law, Stanford Law Review, 65(6): 1195-1202.

[6] This is likely somewhat related to the increasing tendency of top law schools to hire individuals with doctoral degrees in non-law fields. See Joni Hersch and W. Kip Viscusi (2012). "Law and Economics as a Pillar of Legal Education," Review of Law and Economics, 8(2): 487-510.

It is at least a happy accident that the success of the empirical movement in law and economics coincided with what has been dubbed the “credibility revolution”[7] in empirical economics. The hallmark of this revolution has been a focus on research designs that helped to overcome some of the impediments to empirical work in law schools identified by Landes.

In Part II of this essay, we provide a stylized discussion of some trends over the last two or three decades, linking the credibility revolution to the ascendancy of empirical work in law and economics. Then, in Part III, we provide some methodological observations about a number of commonly used approaches to estimating policy effects. In Part IV, we use the literature on the economics of crime and criminal procedure to illustrate the ways in which many of these techniques have been used. In Part V, we give examples of fields—corporate law and economics and civil procedure—that we believe would benefit from increased attention to modern empirical analysis and methods. We then offer some concluding observations.

II. A Stylized History of Empirical Work in Law & Economics

The central problem in much empirical work is omitted variables bias. Sometimes this problem can be solved by controlling for more covariates—if the problem is omission, then inclusion should be a good solution. But this solution is often not feasible, because many omitted variables will be unknown to the researcher, and still others that theory suggests should be included will be unavailable or unquantifiable. Despite these issues, simply adding more control variables was standard operating procedure in empirical law and economics before the mid-1990s.

Another approach was to admit the existence of the bias but to assert that the bias necessarily is in a given direction or to speculate about its probable magnitude. Such an approach may be plausible when there is a single omitted variable, if information is available about its correlation with the policy variable and outcome are known. But if there are multiple omitted variables, this approach is more problematic, because the sign and magnitude of the bias from excluding omitted variables then depends on the relationship between the policy variable of interest and all the omitted variables, as well as the signs and magnitudes of the coefficients on those omitted variables’ coefficients.[8] Except in unusual cases, such an approach may amount to little more than guesswork.

Perhaps because these approaches are contestable, many people viewed empirical work in law and economics (indeed, in the social sciences more generally), as unreliable, if not downright dishonest.[9] Ronald Coase more pithily noted the endogeneity of empirical results to the researcher’s motivating intuitions when he stated, “[I]f you torture the data enough, nature will always confess.”[10]

[7] Joshua D. Angrist and Jörn-Steffen Pischke (2010) "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." Journal of Economic Perspectives, 24(2): 3-30.

[8] On omitted variables bias with multiple omitted variables, *see* William H. Greene, ECONOMETRIC ANALYSIS, 7th ed.; for an approach to the omitted variables bias formula that views omitted variables bias in terms of the joint heterogeneity due to all omitted variables bias simultaneously, *see* Jonah B. Gelbach, “When Do Covariates Matter? And Which Ones, and How Much?”, Journal of Labor Economics (forthcoming, 2016),

[9] For the classic example of this view, see Edward E. Leamer (1983). “Let’s Take the Con Out of Econometrics,” American Economic Review, 73(1): 31-43.

[10] R.H. Coase (2012). Essays on Economics and Economists, p. 27.

In the mid-1990s, many empirical micro-economists began to shift focus to research designs they motivated in terms linked to the method of randomized controlled experiments. Omitted variable bias is not a concern in such experiments since the "treatment" is assigned randomly, so that assignment is statistically independent of any otherwise important omitted variables. In a random assignment experiment, average treatment effects can then be measured simply, using the average change in the outcome of interest for the experimental treatment group, minus the average change in the experimental control group.[11,12]

Empirical law and economics embraced this approach, implementing so-called difference-in-differences research designs to examine a host of legal changes. In general, this approach compares the change in outcomes in jurisdictions adopting a given policy with any contemporaneous change in non-adopting jurisdictions.[13] Under some assumptions that we shall discuss in Part III below, the resulting estimate is consistent (i.e., asymptotically unbiased) for the average effect of the "treatment"—change in law or policy—in the jurisdictions where the change occurred.[14]

With the arguable exception of Florida, jurisdictions most likely do not adopt policy changes at random. Unless one has a strong and believable model of the determinants of policy change—and researchers rarely do—this means that policy changes are endogenous.[15] In that setting, difference in differences approaches are problematic, since the line separating the treatment and comparison group jurisdictions is crossed by deliberate choice, rather than some external and unchosen factors. Some studies bearing the "natural experiments" moniker therefore use instrumental variables to purge their estimates of endogenous policy choice. A valid instrumental variable in this context is one that is correlated with the adoption of a policy change, but not otherwise correlated with the outcome of interest. The first requirement is easy to demonstrate empirically, if it holds. But the second requirement which is an "exactly identifying assumption," cannot be tested and therefore is adopted only because it appears reasonable in context; intuition may be the only real guide to whether the second condition holds.[16]

[11] Additional covariates can be included to improve the precision of the estimate of the estimate, but doing so is not necessary to obtain consistent estimates.

[12] Average effects are not the only type of treatment effects that can be estimated. For examples of studies that consider distributional effects, *see* James J. Heckman, Jeffrey Smith and Nancy Clements (1997), *Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts*, Review of Economic Studies, 64(4): 487-535; and Marianne P. Bitler, Jonah B. Gelbach, and Hilary W. Hoynes (2006), *What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments*, American Economic Review, 96(4): 988-1012.

[13] Mechanically, such analyses are performed by including jurisdiction-specific dummies to account for jurisdiction-level heterogeneity and time period dummies that control for non-linear trends that are common to all jurisdictions represented in the analysis. Other covariates are also often included in hopes that they will pick up any systematic residual differences; note that this approach is another example of the add-more-variables approach, which we above suggested might provide only cold comfort.

[14] That is, it is a consistent estimate of the effect of "treatment on the treated."

[15] *See, e.g.,* Timothy Besley and Anne Case, "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," *Economic Journal* 110(467), pp. F672-F694 (2000).

[16] When there is more than one candidate instrumental variable, the instrumental variables estimator is "overidentified," and it is possible to test the null hypothesis that *all* instruments are *jointly* valid. But it is well known that this test must be conducted conditional on the assumption that at least one instrument is valid. Thus, failure to reject the null hypothesis of joint instrument validity is not evidence of unconditional instrument validity (to oversimplify a bit, such statistical evidence is equally consistent with the possibilities that (i) all instruments are valid, and (ii) all instruments are equally invalid).

Obtaining causal estimates from non-experimental data always requires a judgment that omitted variables bias can be eliminated, so that treatment and comparison jurisdictions can be made comparable. This might be done by adding covariates, by using difference in differences, by using instrumental variables, or by using some other approach (in Part III we discuss additional ones). Ultimately, it is an unavoidable if uncomfortable fact of empirical life that untestable assumptions, and untestably good judgment about them, are indispensable to the measurement of causal effects of real-world policies.

A further point, and perhaps the most important limitation on the usefulness of natural experiments-motivated work, involves the degree of generalizability, or "external validity." The most plausibly exogenous natural experiments may be the ones in which the "shocks" inducing identifying variation are the most limited in terms of what they can tell us about the effects of policy change in other settings. That is, precisely the oddity that gives rise to the shock may make the effects we can estimate from the shock least relevant to other circumstances of interest. This problem has contributed both to Angus Deaton's criticism of the natural experiment methodology[17] and to other authors' arguments in favor of structural econometric methods to generate estimates that can be more policy relevant than those provided by quasi-experimental methods.[18]

Even regarding internal validity, the credibility of a quasi-experimental research design depends crucially on untestable assumptions concerning which treatment and comparison groups are sufficiently comparable.[19] Although this issue is often treated only implicitly in the empirical law and economics literature, it plays a central role in methodological work on non-experimental studies.[20] Some natural experiment designs also generate problems with respect to statistical inference, to the degree that the policy shocks are sticky over time, necessitating careful attention to hypothesis testing and covariance estimation.[21] That said, our view—which is perhaps even shared by some critics of the approaches just cataloged—is that work in empirical microeconomics, including law and economics, has improved substantially since the quasi-experimental perspective has been adopted.

[17] Angus Deaton (2010). "Instruments, Randomization, and Learning about Development," Journal of Economic Literature, 48(2): 424-455.

[18] See, for example, Aviv Nevo and Michael D. Whinston (2010). "Taking the Dogma Out of Econometrics: Structural Modeling and Credible Inference," Journal of Economic Perspectives, 24(2): 69-82 and James J. Heckman and Sergio Urzúa (2010). "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," Journal of Econometrics, 156(1): 27-37. For a response, see Guido W. Imbens (2010). "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua," Journal of Economic Literature, 48(2): 399-423.

[19] This is nicely pointed out in Edward E. Leamer (2010). "Tantalus on the Road to Asymptopia," Journal of Economic Perspectives, 24(2): 31-46 where Leamer once again stresses the value of sensitivity analyses and humility in empirical work.

[20] See, for example, Alberto Abadie, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," Journal of the American Statistical Association, 105(490): 493-505. More generally, see Paul R. Rosenbaum (2010). Observational Studies (Springer Series in Statistics), 2nd ed.

[21] See Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan (2004). "How Much Should We Trust Differences-in-Differences Estimates?" Quarterly Journal of Economics, 119(1): 249-275; A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller, (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors," Review of Economics and Statistics, 90(3): 414-427; and A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller, "Robust Inference with Multi-way Clustering" Journal of Business and Economic Statistics, 29(2): 238-249 (2011).

III. Empirical Challenges and Credibility

Econometric studies come in two basic flavors: structural and reduced form. Structural modelling involves writing down an explicit mathematical and statistical representation of the determinants of individual, firm, or organizational behavior, such that these relationships can be captured with a finite collection of parameter estimates. For example, it is a consequence of Roy's identity that any parametric specification of individual demand for a good can be converted into a parametric utility function.[22] Thus, if one estimates a parametric demand equation, one is estimating parameters of individual utility functions, which are structural parameters that can be used to estimate the effects of future changes in policies. Structural modelling is an approach that generally has not been used in empirical law-and-economics.[23]

Structural modelling has largely fallen out of favor in many fields of empirical economics, because it requires explicit functional form assumptions concerning preferences (of individuals) or cost functions (of firms). The foot soldiers of the credibility revolution alluded to above have tended not to favor making such explicit assumptions; coincidentally, then, empirical work became more common in law and economics just as the larger trend in empirical economics pointed away from *structural* empirical work. Thus, most recent work in empirical law and economics has been *reduced form* in nature. For present purposes, this means such work has not been focused on directly estimating individuals' preferences or firms' cost determinants. Reduced form work instead involves attempting to estimate more generally defined contextual objects such as the average treatment effect of past implementations of policy changes. While reduced form work can involve fewer functional form assumptions, it is certainly not assumption-free; further, it is possible that one doesn't learn as much from reduced form estimation as from valid structural estimation. Thus, the choice between structural and reduced form approaches can involve trading off the need to make stronger assumptions (structural work) against the prospect of learning less information (reduced form work) that could prove to be valuable.

Because so much of recent empirical work on topics connected to law and economics has been reduced form, we focus the limited methodological discussion we offer here on several approaches that are generally used in reduced form studies.

The fundamental challenge in this context is omitted variable bias. That is, when attempting to isolate the causal effect of policy P on outcome Y through, say, the use of multiple regression analysis, it is necessary to rule out the possibility that any estimated effect is not driven by unobserved (or at least

[22] *See, e.g.,* Jerry A. Hausman (1981), *Exact Consumer's Surplus and Deadweight Loss*, American Economic Review, 71(4): 662-76.
TAXES AND LABOR SUPPLY, Handbook of Public Economics, vol. I, edited by A.J. Auerbach and M. Feldstein, Elsevier Science Publishers B. V. (North-Holland).

[23] One exception is in the field of industrial organization ("IO"), the micro-economic field that focuses on understanding how market structure affects consumer and producer welfare. Structural modelling has flourished in IO; *see, e.g.,* Steven Berry, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), pp 841-890 (July 1995); Gautam Gowrisankaran, Aviv Nevo and Robert Town, "Mergers When Prices Are Negotiated: Evidence from the Hospital Industry," (Forthcoming, American Economic Review), http://www.u.arizona.edu/~gowrisan/pdf_papers/hospital_merger_negotiated_prices.pdf; James W. Roberts and Andrew Sweeting, "Airline Mergers and the Potential Entry Defense," http://public.econ.duke.edu/~jr139/pedef_airline_merger.pdf?new_window=1 (August 29, 2013). For a non-IO structural example, *see* Joshua C. Teitelbaum, Levon Barseghyan & Jeffrey Prince, *Are Risk Preferences Stable Across Contexts? Evidence from Insurance Data*, 101 Am. Econ. Rev. 591-631 (2011).

uncontrolled for) variables that happen to be correlated with P. Estimates of the effect of P that do not purge such omitted variables will include any actual causal effect of P on Y, as well as some of the relationship between omitted variable(s) operating through their mutual association with P and Y. This general omitted variable bias problem goes by many names (e.g., endogeneity, selection effects, reverse causality, simultaneity, etc.), all of which boil down to effectively the same problem. In this section, we provide a relatively general analysis of omitted variables bias and strategies for eliminating it in empirical work.

Suppose we are interested in how changes in a policy P affect some continuous outcome variable Y.[24] A traditional way to model the relationship between these variables was to assume that there is a parametric function F that relates them structurally, through a combination of assumptions on individual behavior, organizations' cost functions, and market forces (or other aggregating forces) relating them to each other, such that $Y=F(P;\tau,\varepsilon)$, where τ is a parameter and ε is an unobserved term. The causal effect of a policy change from P1 to P2 is thus to shift Y from $F(P1;\tau,\varepsilon)$ to $F(P2;\tau,\varepsilon)$. If we assume that F is linear in P and ε, then the structural relationship between Y and P is captured by the equation $Y=P\tau+\varepsilon$ together with the claim that when ε is held fixed, a change in P's value from P1 to P2 will induce a change of τ units in Y's value. On this account, the parameter τ measures the causal effect on Y of a one-unit change in P. If P and ε are uncorrelated, then the OLS estimator is consistent for this causal effect.[25] On the other hand, if P and ε are correlated, then the OLS estimator will differ from τ even in large samples.[26]

Researchers are not always willing to specify all the details of a structural model. The more typical modern approach is to regard a linear relationship between P and Y as an approximation. This approach has some intuitive attraction, but it also has its limits if causal effects are the object of estimation. Suppose we have

$$Y_{it} = P_{it}\tau + \varepsilon_{it}, \text{ where } E[\varepsilon_{it}|P_{it}] = 0. \qquad (1)$$

Here the cross-section unit is indexed by i and time is indexed by t, ε_{it} is an unobserved term, and τ is a parameter—something that is fixed and true about the relationship between Y and P. We take both Y and P to have mean zero in order to eliminate the need for a constant in equation (1) (and equation (2) to follow); this can always be achieved by demeaning both variables. The assumption that ε is mean-independent of P, i.e., that $E[\varepsilon_{it}|P_{it}]=0$, is sufficient for τ to be interpretable as the causal effect of the policy P on the conditional average value of outcome Y, because then $E[Y_{it}|P_{it}]=P_{it}\tau+E[\varepsilon_{it}|P_{it}]=P_{it}\tau$, so that $dE[Y_{it}|P_{it}]/dP_{it}=\tau$.[27]

If instead $E[\varepsilon_{it}|P_{it}]$ is non-zero, the relationship between Y, on the one hand, and P and X, on the other, is more complicated to describe. It is important to understand that one can *always* write, as a matter of definition of the parameter τ and the vector u, the following relationship:

[24] Our discussion could be expanded to address applications with categorical outcome variables (including binary ones), as well as to address situations in which a continuous outcome variable is limited due to censoring or truncation. For brevity, though, we focus our analytical discussion on the canonical case of a continuous and unlimited outcome variable.

[25] By "consistency" we mean that an estimator's bias converges to zero in probability as the sample size grows without bound. We ignore technical details related to alternative notions of such convergence.

[26] In that event, the OLS estimator will converge to $\tau+Cov(P,\varepsilon)/Var(P)$, so that OLS overstates the causal parameter τ when above-mean values of the policy and above-mean values of the unobserved term ε tend to go together.

[27] *See* Jeffrey Wooldridge (2002), Econometric Analysis of Cross-Section and Panel Data, Chapter 2.

$$Y_{it} = P_{it}\tau + \varepsilon_{it}, \text{ where } E[P_{it}\varepsilon_{it}] = 0. \tag{2}$$

Equations (1) and (2) look very similar, and since independent variables are uncorrelated, equation (1) does imply equation (2). However, in equation (2) the residual ε_{it} is only uncorrelated with P, not necessarily mean-independent, so equation (2) needn't imply equation (1). From (2), we have $E[Y_{it}|P_{it}] = P_{it}\tau + E[\varepsilon_{it}|P_{it}]$. So if the unobserved term ε is not mean-zero given the policy value, then τ cannot necessarily be viewed as the causal effect of P on Y, since $dE[Y_{it}|P_{it}]/dP=\tau+dE[\varepsilon_{it}|P_{it}]/dP$. The right hand side of this expression will not equal τ unless ε_{it} is independent of P_{it}. This means we cannot interpret the parameter τ as telling us anything causal about the relationship between the policy and the conditional expectation of the outcome.

Suppose that we somehow know that the structural relationship between Y, P, and ε is linear, such that in general,

$$Y = F(P; \tau, \varepsilon) = P\tau + \varepsilon. \tag{3}$$

In that event, we have $F(P_0; \tau, \varepsilon) = P_0\tau + \varepsilon$ and $F(P_0 + 1; \tau, \varepsilon) = (P_0 + 1)\tau + \varepsilon$, from which it follows that $F(P_0 + 1; \tau, \varepsilon) - F(P_0; \tau, \varepsilon) = \tau$. That is, the parameter τ is the but-for causal effect on the outcome of increasing the policy one unit, holding constant all unobserved determinants of the outcome that are captured in ε.

This analysis is closely connected to the analysis of what is estimated by the ordinary least squares ("OLS") estimator. In our simple example, the OLS estimator of τ takes the form

$$\begin{aligned} \hat{\tau} &\equiv (P'P)^{-1}P'Y \\ &= \tau + (P'P)^{-1}P'\varepsilon. \end{aligned} \tag{4}$$

Under standard large-sample conditions, it can be shown that $(P'P)^{-1}P'\varepsilon$ converges to zero provided that P and ε are uncorrelated. It follows that the OLS estimator is a consistent estimator of the parameter τ. We can sum up this discussion with the following observations:

(A) When equation (1) holds, so that $Y = P\tau + \varepsilon$, with the residual ε being mean-independent of the policy P, the parameter τ equals the but-for causal effect of the policy on the conditional expectation of the outcome Y, at each given policy level P.

(B) When the outcome, policy, and residual have a linear structural relationship of the form in equation (3), and when the residual is uncorrelated with the policy, so that equation (2) holds, then the parameter τ equals the but-for causal effect on the outcome of a one-unit increase in the policy when the residual is held constant.

(C) The OLS estimator is consistent for τ in both situation (A) and situation (B).

Thus, OLS generally measure a causal effect of interest in large samples when either (A) the residual ε is mean-independent of the policy variable, or (B) there is both a linear structural relationship between the outcome and the policy variable, and a zero correlation between the policy and residual factors determining the outcome. The latter conclusion adds an important qualifier to the discussion often given in empirical papers. Authors often state that a causal effect of interest can be learned provided simply that our equation (2) holds: i.e., provided only that P_{it} and ε_{it} are uncorrelated. In general, equation (2) is necessary but not sufficient for that result; without an additional assumption such as equation (3)'s linear structural relationship, the parameter estimated by OLS will not generally have a useful causal interpretation.

To understand this point, observe that we can *always* use OLS to consistently estimate the linear projection parameter for the linear projection of Y on P. Define the parameter[28]

$$\theta \equiv E[PP]^{-1}E[PY]. \tag{5}$$

There will always exist a random variable u with the property that $Y = P\theta + u$, which can be verified by multiplying both sides of this equation by P and taking expectations. The result is $E[PY] = E[PP]\theta + E[Pu]$. Using the definition of θ, the right hand side of this result is $E[PY] + E[Pu]$, and since the left hand side is $E[PY]$, it follows that $E[Pu] = 0$ must hold. Consequently, we have the result that

$$Y_{it} = P_{it}\theta + u_{it}, \text{ where } E[P_{it}u_{it}] = 0. \tag{6}$$

The only difference between equations (2) and (6) is the labels used: where we have τ and ε in (2), we have θ and u in equation (6). This observation has two important implications. First, OLS will consistently estimate θ. Indeed, OLS is *always* consistent for the appropriately defined linear projection parameter. Second, though, the parameter θ represents an interesting causal effect only under additional assumptions. Under the additional assumption that the residual u in equation (6) is mean-independent of the policy, we have situation (A) from above. Under the additional assumption that the linear structural relationship in equation (3) holds with θ replacing τ and u replacing ε, we have situation (B) from above. Without one of these assumptions, or some other assumption that is demonstrated to be useful for this purpose, the linear projection parameter θ cannot be interpreted as an interesting causal effect.[29]

The key to policy-relevant empirical work, then, involves two questions. First, is it reasonable to assume that ε and P are mean-independent, or that there is a linear structural relationship between Y and P, with P and ε uncorrelated? Often, if not usually, the answers to these questions will be no. Suppose Y is a measure of crime and P is a measure of policing. Supposing that policymakers believe policing reduces crime, areas that have more crime for reasons unrelated to the intensivity of policing can be expected to do more policing than areas that have less crime, so that, *ceteris paribus*, we can expect ε to be greater in areas with more policing. Consequently, we should not expect bivariate OLS estimates of the relationship between policing and crime to reveal the causal effect of policing on crime.

[28] The parameter ρ will be well defined provided that E[PP] is nonzero, which holds as long as there is variation in the policy in the population; if there is no such variation, we have bigger measurement problems to worry about than what estimator to use.

[29] To understand how θ might not be a causally interesting parameter, suppose the structural relationship between Y and P is quadratic, so that $Y = F(P; \alpha_1, \alpha_2, \varepsilon) = P\alpha_1 + P^2\alpha_2 + \varepsilon$. Then it can be shown that the linear projection parameter is $\theta = \alpha_1 + \left(\frac{E[P^3]}{E[P^2]}\right)\alpha_2 + \left(\frac{E[P\varepsilon]}{E[P^2]}\right)$, which generally doesn't have an interesting causal-effect interpretation (even if P and ε are uncorrelated and P is symmetrically distributed, so that the last two terms are zero, we obtain an estimate of only α_1, which is insufficient to learn anything of general interest).

The second key question is how to estimate causal effects when it is not reasonable to assume that either situation (A) or (B) holds. An enormous amount of modern empirical work is focused on answering this question. We consider a number of approaches in turn.

A. Random Assignment

One approach to solving the problem of dependence between ε and P is to assign policy levels to units randomly. This approach, common in studies involving the effects of medical and psychological interventions, is frequently used in empirical economics, where important studies have been done using random assignment dating at least to the 1970s: examples include the negative income tax; health insurance policy parameters; audit studies designed to test for discrimination; welfare reform; tax preparation; retirement savings; and a variety of settings involving the study of economic development in developing nations.

The advantage of random assignment is that it directly imposes the mean independence of ε and P, so that τ may be regarded as the causal effect of the policy, at least within the particular population studied experimentally. For this reason, it is common in the empirical economics literature to consider randomized controlled trials (RCTs) the conceptual benchmark against which other study types are measured. Indeed, it is often said that randomized controlled trials are the "gold standard" against which all other types of studies should be measured. This is surely too strong a claim, as Heckman and Smith (1995)[30] and Deaton (2010)[31] have ably discussed, because RCTs do have potentially important drawbacks.

One drawbacks is that not all questions are susceptible to study using RCTs. RCTs cannot measure what are sometimes called "general equilibrium effects," i.e., effects that a policy change has to behavior outside the study's domain of impact. For example, suppose we want to measure the effect of liberal discovery rules on civil litigation costs in the federal courts. It is possible to randomly assign non-standard discovery rules to civil actions once they are filed in federal court. But it is not feasible to randomly assign rules regimes to the parties to future disputes. Thus random assignment can be used to measure only some determinants of civil litigation costs; RCTs simply cannot measure the effects of discovery rules on primary behavior—that is, people's and firms' behavior in contract formation, care-taking, and other walks of life that determine whether disputes emerge in the first place. Similarly, a randomized trial whose study population includes only already-filed cases cannot measure any effects of a new federal discovery regime that operate by causing plaintiffs to join in-state defendants in cases that otherwise would be removable to federal court.

Other concerns related to randomization seem insurmountable in at least some contexts relevant for law and economics. An obvious example is the effects of prison sentence length on criminal recidivism. Even if policymakers were willing to give it a go, it is difficult to imagine that randomly assigning sentence length would meet constitutional muster in the U.S. Still other questions would require random assignment of non-manipulable variables, such as race or skin color.[32]

[30] James J. Heckman and Jeffrey A. Smith (1995), *Assessing the Case for Social Experiments*, Journal of Economic Perspectives, 9(2): 85-110.

[31] See note 17 above.

[32] Certain aspects of the U.S. justice system involve what might be termed non-experimental randomization. For example, in their study of whether the quality of legal representation matters in criminal cases, David Abrams and

B. Control function

One approach to dealing with omitted variables bias is to allow the residual from equations (1) or (2) to be dependent on and/or correlated with P while decomposing ε into the sum of problematic and unproblematic components. We can always decompose the residual into a component that is mean-independent of the policy variable and a component that might not be; that is, we can always write ε_{it}= μ_{it}+v_{it}, where $E[v_{it}|P_{it}]=0$.[33] The "control function" approach involves assuming that we have access to some vector of covariates X such that the functional form of $E[\mu_{it}|X_{it}]$ is known. Under the leading example of this approach, one assumes that the relationship is linear, so that $E[\mu_{it}|X_{it}]=X_{it}\beta$.[34] Typically one further assumes that v_{it} is mean-independent of *both* the policy and the covariates. In this event, we have

$$\varepsilon_{it} = X_{it}\beta + v_{it}, \text{ where } E[v_{it}|X_{it}, P_{it}] = 0. \tag{7}$$

Combining this equation with equation (2) then yields

$$Y_{it} = P_{it}\tau + X_{it}\beta + v_{it}, \text{ where } E[v_{it}|X_{it}, P_{it}] = 0. \tag{8}$$

Because we then have $E[Y_{it}|P_{it},X_{it}]=P_{it}\tau+X_{it}\beta$, it follows that $dE[Y_{it}|P_{it},X_{it}]/dP_{it}=\tau$. In other words, τ can be interpreted as the causal effect of the policy on the conditional mean of the outcome, where the conditioning is done on both the policy variable and the covariates. This approach is known as the control function approach because it involves assuming a functional form for ε_{it} that allows one to "control" for enough variables to estimate a causal impact of the policy.

Whether a control function approach is adequate in any particular application obviously depends on both the nature of the omitted variables bias problem in that application, as well as on the quality and extent of covariates that are available. For example, suppose the outcome is a measure of crime and the policy variable increases with the extent of policing, and suppose that the suspected source of omitted variables bias is that local governments increase policing when crime is higher for reasons unrelated to the extent of policing. In this situation, the component μ of residual ε would capture all the

Albert Yoon use the fact that Nevada's Clark County Public Defender office randomly assigns attorneys to incoming felony cases. This makes attorney characteristics exogenous to case quality, so that simple comparisons of case outcomes across attorney characteristics yield valid estimates. This approach solves the problem that attorney-client pairings generally can be expected to be the result of deliberate selection by both attorneys and clients. Abrams and Yoon find that attorney performance varies substantially along multiple characteristics, suggesting that attorney characteristics matter, at least in the public defender context. *See* Abrams, David and Yoon, Albert (2007). "The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability," University of Chicago Law Review, Vol. 74, No. 4. *See also* Shamena Anwar, Patrick Bayer, and Randi Hjalmarsson (2012), *The Impact of Jury Race in Criminal Trials*, Quarterly Journal of Economics, 127(2), 1017-1055.

[33] Let $\mu_{it}=E[\varepsilon_{it}|P_{it}]$, and let $v_{it}=\varepsilon_{it}-\mu_{it}$. Then $E[v_{it}|P_{it}]=E[\varepsilon_{it}|P_{it}]-\mu_{it}=0$ by construction.

[34] This "linear-in-controls" assumption barely nicks the surface of what can be done to account for observable covariates. For example, an enormous literature concerns the use of propensity-based methods that allow covariates to matter in flexible parametric, or even nonparametric, ways. This literature is beyond the scope of the present discussion; for an interesting recent application in the criminal procedure context, *see* M. Marit Rehavi and Sonja B. Starr (2012). "Racial Disparity in Federal Criminal Charging and Its Sentencing Consequences," University of Michigan Law & Econ, Empirical Legal Studies Center Paper No. 12-002, http://ssrn.com/abstract=1985377.

heterogeneity in crime levels that is causally driven by factors other than policing. If all such heterogeneity—all variation in μ—is driven by, say, economic conditions, local-area demographics, and political variables, then including controls for all these variables might reasonably be expected to eliminate the omitted variables bias problem. On the other hand, if covariates that account for such variation are not available, then the control function approach will not yield causal effects. Even with a rich set of controls, the control function approach will generally fail if the true functional form of $E[\varepsilon_{it}|X_{it}]$, i.e., the true control function, is not linear in X_{it}.

While it is not always recognized, one example of the control function approach is the widely used difference in differences strategy for eliminating omitted variables bias. Suppose that we have data on a large number of areas that switch from policy P=0 to policy P=1 between time t=0 and time t=1. One measure of the effect of this policy change is the average difference in the outcome variable across the two periods among areas that made the switch, i.e., using familiar notation for the mean, $\Delta = \bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 0}$. To see how this example fits in our basic framework, we drop the assumption that Y and P are mean zero and allow for a constant, α, in the outcome equation:

$$Y_{it} = \alpha + P_{it}\tau + \varepsilon_{it}. \tag{9}$$

Since $P_{i1}=1$ and $P_{i0}=0$ for all areas making the policy change, we have $Y_{i1}=\alpha+\tau+\varepsilon_{i1}$ and $Y_{i0}=\alpha+\varepsilon_{i0}$. Thus the difference Δ equals $\tau+E[\varepsilon_{i1}-\varepsilon_{i0}]$. This difference can be interpreted as a causal effect of the change in the policy if ε_{i1} and ε_{i0} are each assumed to have mean 0. Notice that this is equivalent to the assumption that $E[\varepsilon_{it}|P_{it}]=0$, since, by construction, $P_{it}=1$ in period 1 and $P_{it}=0$ in period 0.

Of course, a common concern is that the policy might not be the only determinant of the outcome that changed. Again writing $\varepsilon_{it}=\mu_{it}+v_{it}$, with $E[v_{it}|P_{it}]=0$, we can view μ_{it} as the determinants of the outcome that are systematic and unmeasured, and thus potentially problematic. Econometrically, to say that the policy isn't the only determinant of the outcome that changed is to say that $E[\mu_{i1}|P_{it}=1]$ and $E[\mu_{i0}|P_{it}=0]$ are not both zero.

A common approach to dealing with this concern is to find a comparison group of areas that (i) did not implement the policy change from 0 to 1, and (ii) are otherwise similar to the areas that did implement the change. Let Δ_{CG} be the difference across periods in average outcomes for these comparison group areas, and let Δ_{TG} be the difference across periods in average outcomes for the "treatment group" areas, which changed the policy. The difference in differences approach is to focus on $\Delta\Delta=\Delta_{TG}-\Delta_{CG}$.

The logic of using ΔΔ is simple and intuitive: if the only systematic difference between the treated and comparison group areas is that the treated areas changed the policy and the comparison group ones didn't, then it would be reasonable to think that non-policy driven changes in the outcome should have been similar in the two areas. As we shall show momentarily, we can think of Δ_c as a measure of the bias caused when we ignore the possibility that factors other than the policy changed in the treated areas. Subtracting this bias from Δ_t then leaves only the causal effect of the policy change.

This discussion shows that the difference in differences approach is often viewed as being analogous to running a random assignment experiment. This analogy is useful to the extent that it helps one understand the conditions for the difference in differences approach to work. But it is also important to recognize that there are substantive differences between real and so-called natural experiments. In real experiments, one uses random assignment to create a treated group and a control group, which by the nature of randomization may be treated as comparable. In "natural" experiments, a policy change is typically deliberately, rather than randomly, implemented in treated areas, while the policy is deliberately, rather than randomly, kept the same in comparison (*not* control) group areas. In random assignment experiments, the assumptions necessary to justify treating the treated-control group difference in outcomes as causal are trivially mild: the randomization must have been done properly, and the study size must be large enough such that incidental differences between treated and control observations wash out in the averaging. With “natural” experiments, by contrast, one must be willing to believe that the researcher-selected treated and comparison group areas would have had the same changes in outcomes had the treated areas not implemented the policy change.

To make all this precise, observe that condition (i) for the comparison group areas implies that for them, Δ_{CG}=E[Y_{i1}-Y_{i0}|Comparison]=E[μ_{i1}-μ_{i0}+v_{i1}-v_{i0}|Comparison], whereas for the treatment group we have Δ_{TG}=E[Y_{i1}-Y_{i0}|Treatment]=τ+E[μ_{i1}-μ_{i0}+v_{i1}-v_{i0}|Treatment]. Thus, ΔΔ=τ+ E[Δμ +Δv|Treatment] - E[Δμ +Δv |Comparison]. Since v is mean zero for both groups in all periods, by construction, we can simplify this to ΔΔ=τ+ E[$\Delta\mu_{TG}$] - E[$\Delta\mu_{CG}$]. The difference in differences approach yields the causal effect of the policy change in treated areas, then, whenever E[$\Delta\mu_{TG}$]-E[$\Delta\mu_{CG}$]=0. This is where condition (ii)’s “otherwise similar” requirement comes in. If average unobserved systematic changes in the outcome across time periods would have been the same in the treated and comparison group areas in the absence of the treated group’s policy change, then the average value of $\Delta\mu_{CG}$ will be equal to $\Delta\mu_{TG}$, the corresponding average for the treatment group.

Thus, the precise condition for the difference in differences approach to yield the causal effect of the policy change in treated areas is that, on average, the trend in the outcome must be the same in treated and comparison group areas over the period of time during which the policy change occurred. The art of picking comparison groups, then, is of picking units that would have had the same unobservable-driven changes in outcomes but for the policy change. This is not necessarily easy to do in practice, and with a poor choice, the difference in differences estimator ΔΔ might well be a more biased estimator for the treatment effect than the simple before-and-after difference Δ_t that is constructed without using a comparison group.[35] There are no silver bullets in this business; ultimately, one must be willing to believe that—because one cannot test whether—variables chosen as the basis for differencing really do yield appropriately comparable groups.[36]

C. Instrumental variables

Instrumental variables estimation has been extremely influential in applied microeconomics over the last quarter century. The traditional way to define a valid instrumental variable is to assume a linear structural relationship of the form in equation (3) and then argue that some variable Z exists that is correlated with the policy variable P, but is not otherwise correlated with the outcome Y. To put it

[35] Researchers sometimes deal with this challenge by adding a third dimension of difference and then constructing the “difference in the (differences in differences)”. That is, one defines ΔΔΔ= $\Delta\Delta_1$ - $\Delta\Delta_2$, where $\Delta\Delta_1$ includes the treatment group and $\Delta\Delta_2$ includes some other groups for which the trends left standing after the second difference are supposed to mirror the “bad” (i.e., non-treatment effect) trends in $\Delta\Delta_1$. Whether this approach improves on the two-level difference in differences approach again depends entirely on the degree to which the third dimension differences out bad variation without introducing new sources of it.

[36] Even an 8-dimension differencing estimator (ΔΔΔΔΔΔΔΔ, or what might be called “the Buckaroo Banzai estimator”) will still require the assumption that the “trends in trends in trends …” that are left after the 8th difference are not systematically problematic.

differently, Z must be correlated with the outcome variable, but only through its correlation with the policy variable P, rather than through any correlation with the unobserved term ε. With such a Z in hand, we have

$$\begin{aligned} E[Z'Y] &= E[Z'P]\tau + E[Z'\varepsilon] \\ &= E[Z'P]\tau, \end{aligned} \tag{10}$$

it follows that

$$E[Z'P]^{-1}E[Z'Y] = \tau. \tag{11}$$

The condition that Z and P be correlated is necessary to ensure that $E[Z'P]^{-1}$ is finite. Without the condition that Z and ε are uncorrelated, the second equality in (4) would not hold, and $E[Z'P]^{-1}E[Z'Y]$ would differ from τ. Notice that this discussion leans heavily on the assumption that τ is a causal parameter. An alternative approach is to posit a modified form of equation (1):

$$Y_{it} = P_{it}\tau + \varepsilon_{it}, \text{ where } E[\varepsilon_{it}|Z_{it}] = 0. \tag{1'}$$

The assumption that the instrument Z and the residual ε are mean-independent takes the place of the assumption in equation (1) that X and ε are mean-independent. When assumption (1') holds, $E[Y_{it}|Z_{it}] = E[P_{it}|Z_{it}]\tau$, and so τ may be regarded as a type of causal effect of changing P on Y: it tells us the causal change in Y induced by changing P via a change in the instrument Z.[37] When we are unwilling to maintain either the assumption that $E[\varepsilon_{it}|Z_{it}] = 0$ or the linear structural relationship, we have a similar result to that discussed in conjunction with equation (2) above, in that τ no longer can be interpreted as a causal effect.[38]

The IV estimator is calculated by taking the ratio of the sample covariance between Y and Z to the sample covariance of P and Z; this yields what is known as the indirect least squares estimator, which can be shown to be consistent for τ in equation (11). It is also easy to show that this estimator equals the ratio of the OLS estimator of Y on Z to the OLS estimator of Y on P.[39] When there are more than one valid instrumental variables, they can be used together with the two-stage least squares estimator. This estimator can be calculated in various ways, but the easiest to explain (and the source of its moniker) is the procedure of first using OLS to project P on all the instruments, then using the resulting coefficient vector to calculate the fitted values of P, and then using OLS to project Y on these fitted values of P.[40]

[37] Observe $\frac{dE[Y_{it}|Z_{it}]}{dZ_{it}} = \frac{dE[P_{it}|Z_{it}]}{dZ_{it}}\tau$, so that a change in Z changes Y by an amount equal to the product of τ and the change in P induced by changing Z. For a more abstract treatment of some of these issues in the case in which the policy P is binary, *see* Angrist and Imbens (1994), *Identification and Estimation of Local Average Treatment Effects*, Econometrica 62(2): 467-475.

[38] Instead, it is, roughly speaking, the ratio of the parameters from a linear projection of Y on Z and a similar projection of Z on P.

[39] This fact echoes the interpretation of τ given in footnote 37. Suppose E[P|Z]=Zα. Then $\frac{dE[P_{it}|Z_{it}]}{dZ_{it}} = \alpha$ and $\frac{dE[Y_{it}|Z_{it}]}{dZ_{it}} = \alpha\tau$, so that $\frac{\frac{dE[Y_{it}|Z_{it}]}{dZ_{it}}}{\frac{dE[P_{it}|Z_{it}]}{dZ_{it}}} = \frac{\alpha\tau}{\alpha} = \tau$. With mean independence, the numerator can be regarded as the estimand of the OLS projection of Y on Z, while the denominator can be similarly regarded with P replacing Y.

[40] We will ignore the issue of estimating consistent standard errors, which needs to be handled carefully here; many statistical software packages have canned routines to handle two-stage least squares.

When there is exactly one instrumental variable, the methods just discussed yield numerically identical estimates of τ.

A simple example of a valid instrumental variable involves any random assignment experiment in which intended assignments are all fulfilled. Suppose half the units in a population are assigned to receive P=1 and the other half are assigned to receive P=0, with received policy values equal to these assignments. Writing Z for the intended assignment (1 or 0), we see that Z is perfectly correlated with P; since Z is chosen randomly, it is also independent of ε, and therefore mean-independent as well. The IV estimate of τ is then consistent for the causal effect of P on the conditional mean of Y. Of course, it is easy to show that the identical estimate would be obtained by subtracting the average value of Y among those randomly assigned P=1 and P=0. IV does differ from this simple difference in the situation in which compliance with intended assignment is less than perfect, so that some units assigned P=0 instead receive P=1 and some assigned P=1 instead receive P=0. Provided that such noncompliance is functionally random, the IV estimator of τ will yield the causal effect of P on Y that operates through the instrumental variable Z.

Importantly, IV is applicable outside the narrow realm of RCTs with imperfect compliance, and IV has been extremely influential. But it has also been very controversial. One issue sometimes cited is that the assumptions that an instrumental variable must meet are fundamentally untestable. Of course, the same is true of the assumptions necessary to use OLS; indeed, *any* method of measuring causal effects will involve some untestable assumptions somewhere.[41] A more compelling concern is that when researchers are not willing to specify a structural model of the Y=F(P;τ,ε) form, the population whose indirect causal effect is estimated using IV also may be unobservable; *see* Heckman (1997) on this point, vis-à-vis Angrist & Imbens's (1994) Local Average Treatment Effect ("LATE") parameter.[42] Thus the flexibility of avoiding fully specifying structural relationships is effectively purchased by making implicit assumptions on the form of these relationships. Such assumptions may be difficult even to understand in the absence of a clear statement of precisely the specification that is sought to be avoided.

D. Regression Discontinuity Approaches

Another approach is to restrict the set of units under consideration. The idea behind this "data restriction" approach is that equation (1)'s mean independence assumption might be reasonably thought to hold for certain observable subsets of the overall population under study. The classic example of such approaches is regression discontinuity ("RD").[43] RD designs exploit the fact that in practice, assignment of units to interesting statuses is often done using a binary function applied to variables that may be thought of in roughly continuous terms. Thus, binariness creates a discontinuity around some threshold, which can allow one to cleanly measure the effect of the status of interest on important outcomes.

[41] For a discussion of this fact in the context of empirical civil procedure scholarship, see Jonah B. Gelbach, "Can The Dark Arts Of The Dismal Science Shed Light On The Empirical Reality Of Civil Procedure?", Stanford Journal of Complex Litigation, 2(2): 223-296.

[42] James J. Heckman (1997), *Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations*, Journal of Human Resources, 32(3): 441-462.

[43] For a detailed discussion, *see* Guido W. Imbens and Thomas Lemieux, "Regression discontinuity designs: A guide to practice," *Journal of Econometrics* 142(2): 615-635.

To illustrate, suppose that one is interested in measuring the earnings effect of attending a fancy law school. Comparing earnings of those who do and don't attend elite law schools is a problematic way to do this, since (a) only those who are admitted to elite law schools may attend them, and (b) such applicants might be expected to have greater average earnings than those who aren't admitted, regardless of whether the admittees did attend elite law schools. Suppose, though, that Fancy Law School has a policy of accepting every applicant with an LSAT score equal to or greater than 170 and rejecting every applicant with a lower score. Suppose in addition that anyone accepted to Fancy Law School attends and graduates (this assumption is just for exposition).

Now suppose one had data on the LSAT scores and future earnings of all applicants to Fancy Law School in a particular year—before anyone outside the School knew of this policy. It would likely be problematic to compare *all* those who attended Fancy Law School to *all* those who did not, since these two categories can be expected to include people with widely divergent LSAT scores, who we have already decided might be expected to differ in wage-earning ability, quite apart from their law school attendance. But what if we compare earnings of Fancy Law School applicants with an LSAT score of 170 to those with a score of 169? Such applicants are plausibly very similar to each other. Among them, the sets who ended up with a 170 and a 169 might be as much due to seemingly random factors—e.g., who had coffee and cigarettes for breakfast versus those who ate Wheaties—as to anything related to general wage-earning ability.[44] Accordingly, any difference in earnings of Fancy Law School applicants with LSAT scores of 170 and 169 is unlikely to be due to differences in underlying ability. Such earnings differences would therefore be a good candidate for measuring the earnings effect of attending Fancy Law School, at least among those with LSAT scores of 169 or 170.

This example illustrates the simplest RD design, in which researchers compare the outcome, Y, for units observed just above and below a threshold level of some other variable, known as the "running variable," which determines the attribute whose effect one is trying to measure. In our Fancy Law School example, the outcome variable, Y, is earnings, and the "policy" variable, P, is whether a person attends Fancy Law School; the running variable is the person's LSAT score, and the threshold is 170.

RD designs can account for "imperfect compliance," i.e., the situation in which crossing the threshold doesn't perfectly determine the attribute captured by P. For example, suppose we knew that *only some* admitted applicants actually attend Fancy Law School, rather than *all*. Then the average earnings of Fancy Law School applicants with scores of 170 would include some people who attended the School and some who didn't. A common approach to dealing with such imperfect compliance is to "scale up", dividing the difference in mean earnings by the difference in the Fancy Law School attendance rate among those who have a score of 170 and those with a score of 169. Since no one in the latter category is admitted, its attendance rate is 0. Thus, this "scaling-up" approach involves dividing the earnings difference by the share of 170-scoring applicants who actually attend Fancy Law School.[45]

To illustrate scaling up, suppose subsequent annual earnings are $1,000 among applicants with a score of 170, and suppose that half of those with 170-scoring applicants attend Fancy Law School. Then the

[44] According to LSAC data, for the period between 2010-2013, an LSAT score of 170 corresponded to the 97.4th percentile, while a score of 169 corresponded to the 96.6th percentile—a very small difference in relative rank; *see* the LSAT Percentiles Table available at http://www.cambridgelsat.com/resources/data/lsat-percentiles-table/.

[45] This is actually a special case of instrumental variables, with the policy variable being a dummy variable indicating attendance at Fancy Law School, and the instrument being a dummy indicating whether an applicant scored 170.

impact of Fancy Law School attendance is estimated to be \$1,000÷0.50, or \$2,000. Intuitively, those 170-scorers who don't attend Fancy Law School aren't supposed to have different earnings from 169-scorers. Thus, we can think of the observed \$1,000 difference as equaling the simple average of zero and \$2,000—corresponding to the earnings difference for those 170-scorers who don't attend Fancy Law School and those who *do*.

Notice that this example highlights an important assumption of the scaling-up approach. Suppose there's exactly one other elite law school out there, Justasfancy Law School, which follows the same policy as Fancy Law School, and suppose that anyone who applies to Fancy Law School also applies to Justasfancy Law School. Finally, suppose that everyone who's admitted to these law schools will choose to attend one of them. It follows that 170-scorers who applied to Fancy Law School and don't attend must have gone to Justasfancy Law School. But now we have a problem, at least if (a) fanciness is what makes elite law school attendance pay, and (b) Justasfancy Law School is just as fancy as Fancy Law School. Under these conditions, 170-scorers who don't attend Fancy Law School *get the same "treatment effect" as those who do attend Fancy Law School*. Instead of having an earnings difference of \$1,000 relative to those with 169 scores, those who attend Justasfancy Law School get the same \$1,000 treatment effect as those who attend Fancy Law School. Validity of the scaling-up approach to RD designs with imperfect compliance is wrecked by such a phenomenon; such validity requires that it is attendance at Fancy Law School, not receiving a score of 170, that drives any observed earnings difference.

With perfect compliance, or enough faith in scaling up with imperfect compliance, the RD approach provides clean measures of treatment effects. It's important to realize, though, that *the external validity of these effects is confined to units that fall within the local range where the discontinuity occurs*. Our original RD study, the one with perfect compliance, is great for learning the effects of attending Fancy Law School for those applicants who score 169 or 170. But if we want to know what the effects of going to Fancy Law School would be for someone with an LSAT score of 158, or 175, our RD study is of limited use, because the binary rule that operates on LSAT scores does so *only* at a score of 170. To use our study's results requires one to believe the true earnings effect of Fancy Law School attendance is the same for the {158,175} LSAT population as for the {169,170} population. In other words, letting P indicate attendance at Fancy Law School and S_i be person i's LSAT score, one must believe

$$\begin{aligned} \tau &= E[Y_i|P_i = 1, S_i = 170] - E[Y_i|P_i = 0, S_i = 169] \\ &= E[Y_i|P_i = 1, S_i = 175] - E[Y_i|P_i = 0, S_i = 158] \end{aligned} \tag{12}$$

In turn, this implies that for all those with (P_i, S_i) values in the set $L \equiv \{(0,158), (0,169), (1,170), (1,175)\}$, the relationship between earnings and Fancy Law School attendance satisfies

$$Y_i = \alpha + P_i\tau + \varepsilon_i, \text{ where } E[\varepsilon_i|(P_i, S_i) \in L] = 0, \tag{13}$$

where τ is the (constant) treatment effect of attending Fancy Law School, P_i is a dummy variable indicating attendance at Fancy Law School, α is the average earnings of one who does not attend Fancy Law School, and ε_i captures other earnings determinants. Importantly, ε_i is mean-independent of Fancy Law School attendance, for all those with LSAT scores 158, 169, 170, and 175. But if that is true, then we wouldn't need to use the admissions-policy discontinuity in the first place: we could have just calculated average earnings for those who attend Fancy Law School and have LSATs 170 or 175, and then subtracted the corresponding average for those with LSATs 158 or 169.[46]

We can generalize this example. Let S_i be *i*'s LSAT score, and suppose the following model is true:

$$Y_i = \alpha + P_i\tau + S_i\beta + u_i, \text{ where } E[u_i|P_i] = 0. \tag{14}$$

Notice that this is very similar to equation (13). Defining $\varepsilon_i = S_i\beta + u_i$ links the equation for Y_i in the two equations. The only difference is that $E[\varepsilon_i|P_i] = S_i\beta + E[u_i|P_i] = S_i\beta$, which is zero only if β is. If this model is true, then a simple multivariate least squares approach applied to (14) will yield valid estimates of the treatment effect τ, provided that LSAT score is included as an additional covariate.

This discussion suggests that the further one's interest is from the RD threshold, the more one is relying for identification on functional form assumptions, and the less one is relying on the "clean" variation generated by the binary rule. As Imbens and Lemieux (2008, p. 622) put it, "Without strong assumptions justifying extrapolation[, such as…] homogeneity of the treatment effect[, … RD] designs never allow the researcher to estimate the overall average effect of the treatment. In that sense the design has fundamentally only a limited degree of external validity, although the specific average effect that is identified may well be of special interest, for example in cases where the policy question concerns changing the location of the threshold." Thus, RD studies necessarily involve either accepting a tradeoff between internal and external validity concerns, or the good luck of finding a discontinuity in just the part of the population of interest.[47]

IV. Exemplar of the "Credibility Revolution": The Law & Economics of Crime

In this Part we discuss one subfield, broadly construed, in which many techniques popularized in the credibility revolution have been deployed: the law & economics of crime, which offers examples of both missteps and plausible successes.[48] We shall focus on the literature on the effect on crime of the number of police in local jurisdictions.

The effect of policing on crime has been studied extensively for many decades, but many early studies were tough to swallow. One literature review in the late 1980s noted that the majority of empirical studies found either no relationship or even a positive relationship between police and crime.[49] While it is possible that there is no causal relationship between policing and crime, even the credulous might

[46] Further, there are efficiency reasons why we'd want to use all possible observations: we can obtain more precise estimates of the treatment effect τ with more observations than we can with fewer.

[47] Studies using RD designs to study law-relevant topics include Miguel de Figueiredo (2014), *Throw Away the Jail or Throw Away the Key? The Effect of Punishment on Recidivism and Social Cost*, working paper; David S. Lee and Justin McCrary, "Crime, Punishment, and Myopia," NBER Working Paper (July 2005).

[48] For a striking discussion of an entire literature that can be charitably called a misstep, see John J. Donohue and Justin Wolfers (2005). "Uses and Abuses of Empirical Evidence in the Death Penalty Debate," Stanford Law Review, 58(3): 791-846; *see also* our discussion of errors in Steven Levitt's work, *infra*.

[49] Samuel Cameron (1988). "The Economics of Crime Deterrence: A Survey of Theory and Evidence," Kyklos, 41(2): 301-323.

have a hard time imagining how police add measurably to the kind of crime generally studied.[50] It is much easier to imagine that these studies suffer from omitted variable bias, with police having been added as crime rose. To deal with this omitted variables bias, many of these studies used either the "just add more variables" control-function approach or used only time series data with no credible comparison group; those studies using panel data didn't generally defend the implicit assumption that police staffing levels must not be driven by omitted variables related to crime.

Against this backdrop, Steven Levitt's influential 1997 paper hypothesized that, in order to seek election, mayors and governors attempt to credibly signal that they are tough on crime by hiring more police prior to Election Day.[51] Because elections in U.S. jurisdictions virtually always follow a pre-set schedule, their timing is not affected by crime or anything related to crime. Thus, election timing might generate observable shocks to police hiring, allowing IV estimation to measure the causal effect of police on crime. Levitt found that the number of police had a large negative effect on crime after addressing the omitted variable problem through his election instrument. Levitt's design was ingenious, but unfortunately, it was also wrong: Justin McCrary later showed that Levitt's results were driven by coding errors related to how Levitt weighted his data. [52]

In his note discussing Levitt's error, McCrary concluded, "In the absence of stronger research designs, or perhaps heroic data collection, a precise estimate of the causal effect of police on crime will remain at large (p. 1242)." In a published reply to McCrary,[53] Levitt himself offered an alternative instrumental variables-based approach, using the number of municipal firefighters per capita to instrument for the size of police forces. If municipalities with more firefighters are more willing to hire public sector workers generally, then they should also hire more police, all else equal; Levitt's estimates suggest this is true empirically, which establishes the first condition for instrument validity. As long as deployment of fire fighters isn't either causally or coincidentally related to crime trends, the number of firefighters won't be associated with crime outside of its association with police force size; that would establish the second condition. Levitt's newer instrumental variables estimates suggest an elasticity of crime with respect to police force size of between -0.4 and -0.5, though these estimates are relatively imprecise.

Shortly after the exchange between McCrary and Levitt, two sets of authors exploited variation in policing following terrorism events in two very different jurisdictions, using this variation to isolate effects of police on crime. In 1994, the government in Buenos Aires reacted to a terrorist attack on the main Jewish center there by providing additional police protection to all Jewish institutions in the city. Rafael Di Tella and Ernesto Schargrodsky examined the change in automobile thefts in those blocks receiving extra police protection relative to blocks where police coverage remained constant, yielding an elasticity between crime and police of -0.3.[54] Jonathan Klick and Alexander Tabarrok use changes in the status of the U.S. Department of Homeland Security ("DHS") Terror Alert System in 2002-2003 as a

[50] Of course, police brutality and corruption also involve crime. Whatever the frequency of such incidents, we doubt they meaningfully affect the sorts of crime statistics typically examined in the literature on policing and crime.

[51] Steven D. Levitt (1997). "Using Electoral Cycles in Police Hiring To Estimate the Effects of Police on Crime," American Economic Review, 87(3): 270-290.

[52] See Justin McCrary (2002). "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Comment," American Economic Review, 92(4): 1236-1243.

[53] Steven D. Levitt (2002), "Using Electoral Cycles in Police Hiring to Estimate the Effects of Police on Crime: Reply," American Economic Review, 92(4): 1244-1250.

[54] Rafael Di Tella and Ernesto Schargrodsky (2004). "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack." American Economic Review, 94(1): 115–133.

shock to police coverage in a specific area of Washington D.C., where the police policy was to have officers work overtime during periods of heightened terrorism threat.[55] Because DHS decisions were not based upon mine-run crime conditions in Washington, terror alert status changes were plausibly exogenous with respect to crime. Klick and Tabarrok estimate a police-on-crime elasticity of about -0.3—the same as Di Tella and Schargrodsky—with the effect concentrated among opportunistic crimes (such as auto theft). Subsequent work by others yielded similar results using police redeployments after the 2005 London terror attacks.[56] These three studies use essentially the same research design: difference-in-differences, using one or more shocks to local policing needs that are likely unrelated to run of the mill crimes.

Additional research approaches exploiting variation in policing driven by sources other than terrorism concerns have generated estimates comparable to those just discussed. William Evans and Emily Owens exploit the Violent Crime Control and Law Enforcement Act of 1994, which funded hiring of additional police officers through the Community Oriented Policing Services program.[57] While applications for the grants and the awarding of the grants could potentially be endogenous to unobservable variables related to crime, they do not find any relationship between ultimate awards and pre-funding crime trends, ruling out the most direct source of endogeneity. Evans and Owens find elasticities that are similar to those found in the terrorism-related studies.

John MacDonald, Jonathan Klick, and Ben Grunwald use the largely arbitrary boundary separating high and low policing areas around the University of Pennsylvania campus as the basis for an RD design.[58] They observe that the Penn police force does not patrol beyond the historical campus boundary, while areas inside and outside that boundary are otherwise quite similar. Because the university police force provides many more officers per unit of area than does the surrounding Philadelphia police force, an RD study can be based on comparing city blocks just beyond the campus to those just within it. These authors find a statistically significant negative effect of police on crime, of a magnitude nearly identical to that found in the terrorism-related studies.

Lastly, Aaron Chalfin and Justin McCrary provide estimates from a standard panel data analysis of crime and police in U.S. cities based on accounting for measurement error concerning police force size.[59] Measurement error, which has largely been ignored in the literature, creates omitted variables bias. Like other sources of omitted variables bias, then, measurement error can be dealt with using instrumental variables. With two alternative variables measuring police force size, Chalfin and McCrary are able to use one of these variables to instrument for the other, an approach that yields valid estimators so long as the measurement error in each variable satisfies certain "classical" conditions. They find results broadly comparable to the quasi-experimental designs recounted above. Further, after performing a

[55] Jonathan Klick and Alexander Tabarrok (2005). ""Using Terror Alert Levels to Estimate the Effect of Police on Crime," Journal of Law and Economics, 48(1): 267-279.

[56] Mirko Draca, Stephen Machin, and Robert Witt (2011). "Panic on the Streets of London: Police, Crime, and the July 2005 Terror Attacks." American Economic Review, 101(5): 2157-2181.

[57] William N. Evans and Emily G. Owens (2007). "COPS and Crime," Journal of Public Economics, 91(1-2): 181-201.

[58] John MacDonald, Jonathan Klick, and Ben Grunwald (2012). "The Effect of Privately Provided Police Services on Crime," University of Pennsylvania Institute for Law and Economics Research Paper No. 12-36.

[59] Aaron Chalfin and Justin McCrary (2013). "The Effect of Police on Crime: New Evidence from U.S. Cities, 1960-2010," NBER Working Paper No.: 18815.

cost-benefit analysis, they find that a dollar spent on police leads to substantially more than one dollar in crime reduction benefits, which suggests that cities should hire more police than they currently do.[60]

While each of these papers necessarily entails exactly identifying assumptions, it is also true that each uses a research design that plausibly mitigates at least some concerns about omitted variable bias. Further, even though the weaknesses across the papers are different, they reach generally similar qualitative conclusions. Broadly speaking, then, Chalfin and McCrary's estimates of the effect of policing on crime are relatively insensitive to the empirical identification strategy taken and the location or time period examined. These conclusions suggest that the weaknesses in each study have similar biasing effects. One would like to believe that this similar amount of bias is zero. If so, then the police-and-crime literature should be viewed as a real success story of empirical work in law and economics: through the application of contemporary empirical methods largely motivated by the "credibility revolution," researchers in this literature have uncovered compelling evidence on an issue of great policy import.

V. Two Literatures that Could Use Improvement

In this Part we discuss two areas where the ideas discussed above have had less purchase: corporate law and economics, and civil procedure.

A. Corporate Law & Economics

One area of law and economics where empirical work has been slow to adopt the modern quasi-experimental approach is corporate law and economics. In this area, the standard approach to empirical inquiry involves the use of event studies to deduce the effect of various public policies or internal corporate governance mechanisms on firm value.[61]

Event studies in this literature generally follow a fairly simple recipe. A researcher identifies the timing of the event of interest and assumes that the effect of this event will be capitalized into the value of the asset in question, appealing to the assumption of semi-strong market efficiency. The researcher then estimates a time-series model of the asset's returns based upon the historical relationship between the asset's returns and other covariates, usually including some proxy for the market return. Based upon this model, the researcher predicts of the asset's return at the time of the event. This prediction yields the researcher's best guess about what the asset's return would have been in the counterfactual

[60] However, it is important to note the general problem of how far results can be extrapolated from studies such as this. When Washington, D.C. significantly increased its police force in 1989-1990, it ended up hiring gang members and other candidates who might not have been optimally placed as police officers. *See* Keith A. Harriston and Mary Pat Flaherty, *D.C. Police Paying for Hiring Binge*, The Washington Post, August 28, 1994, page A1. A similar, example of this phenomenon was portrayed in the 1984 Steve Gutenberg movie *Police Academy*.

[61] Eugene Fama, Lawrence Fisher, Michael C. Jensen, and Richard Roll (1969). "The Adjustment of Stock Prices to New Information," International Economic Review, 10(1): 1-21 is often credited as the first event study. However, as noted by Joseph Michael Newhard (2014). "The Stock Market Speaks: How Dr. Alchian Learned to Build the Bomb," Journal of Corporate Finance, 27: 116-132, Armen Alchian had performed one in 1954 while he was working at Rand to deduce the fusion fuel being used in the newly developed hydrogen bomb using stock prices of firms providing the candidate fuels. This paper was never published since it was deemed a threat to national security.

absence of the event in question. The researcher then calculates the excess return for the event date, which is the difference between the actual return at the time of the event and the predicted counterfactual return. This excess return is then standardized to account for the volatility of the asset (i.e., the excess return is normalized by dividing by some measure of volatility of the asset's return such as the standard deviation of non-event day excess returns), and statistical inferences are made.[62]

Although it is not generally presented this way, the recipe relayed above is equivalent to estimating the coefficient on a dummy variable indicating the date of the event using OLS estimation of a model that relates daily returns to this event dummy and a market-return variable (and perhaps other covariates). If the event in question takes place exactly on a single day, then the coefficient on the event dummy will exactly equal the residual for that day calculated using the more detailed method described in the previous paragraph.[63] If the event occurs on multiple dates, the estimated coefficient will roughly equal the average value of the residuals on those dates.

Expressed this way, it is easy to see that event studies are functionally equivalent to estimating a simple before-and-after time series model with no comparison group. Thus, their validity hinges on one's faith in the proposition that no other important things happened at the same time as the event in question. That assumption has been the basis of event studies' use in securities litigation, for instance.[64]

It has also been used by Jonathan Klick and Robert Sitkoff to study Henry Manne's market-for-corporate-control hypothesis.[65] This hypothesis states that for publicly held firms, the possibility of a hostile takeover will discipline a firm's managers, forcing them to maximize firm value even if the firm's board is less than perfect in its monitoring. Klick and Sitkoff exploit the political economy dynamics that surrounded an attempt to sell the controlling interest in the Hershey Company in 2002. Since the majority of shareholder voting power was held by the Milton Hershey School Trust, the company's management was not subject to whatever beneficial takeover pressures there are, because the outsiders would not be able to acquire Hershey via the takeover mechanism central to Manne's hypothesis. With the trustees having little to personally gain from a sale, even to an entity that could generate more value from the company's assets, management might not be disciplined by the fear of an acquisition.

But out of the blue, a staffer from the Pennsylvania Attorney General's Office, which supervises charitable trusts in the state, suggested that the trust needed to diversify its portfolio. This triggered the expectation of a sale, presumably to a party having more highly powered incentives. Such a sale's consummation would provide the conditions for Manne's hypothesis to hold. These events were followed by an increase in the price of Hershey's shares that were publicly traded. These shares, which would not have been involved in any sale then expected by market participants, experienced an excess return more than 20 standard deviations above zero. Later, the Attorney General stopped the sale for

[62] For a useful review, see Sanjai Bhagat and Robert Romano (2002). "Event Studies and the Law: Part I: Technique and Corporate Litigation," American Law and Economics Review, 4(1): 141-168.

[63] *See* Jonah Gelbach, Eric Helland, and Jonathan Klick (2013). "Valid Inference in Single-Firm Single-Event Studies," American Law and Economics Review, 15(2): 495-541.

[64] *See* Gelbach, Helland & Klick (2013) for a discussion of event studies in the securities litigation context; *see also Halliburton Co. v. Erica P. John Fund, Inc.*, 134 S. Ct. 2398, 2415 (2014), wherein the Supreme Court blesses event studies as a way for parties to "introduce evidence of the existence of price impact" following "pertinent publicly reported events".

[65] Jonathan Klick and Robert Sitkoff (2008). "Agency Costs, Charitable Trusts, and Corporate Control: Evidence from Hershey's Kiss-Off" Columbia Law Review, 108(4): 749-838.

what appeared to be only political reasons, and there was a corresponding decline in the value of the publicly traded stock—just as Manne's theory predicts should happen when a company no longer faces the threat of outside acquisition.

Results from Klick and Sitkoff's study offer credible support for the Manne hypothesis, even though they amount to standard before-and-after analysis, because the events in question did not appear to be confounded by any other systematic changes. When such granular analysis of the events in question cannot be conducted, though, traditional event studies are less credible. As Vladimir Atanasov and Bernard Black discuss in a recent literature review, it is rare for empirical research in corporate finance, including that done by law and economics scholars, to focus on the kinds of natural experiments discussed above.[66]

One might buttress the credibility of such studies by using plausible comparison companies. While comparison groups are sometimes used indirectly for so-called falsification tests, or as general control variables, it is rare that they are used to generate difference-in-differences estimates. One notable exception is provided by Michael Greenstone, Paul Oyer, and Annette Vissing-Jorgensen who reexamine the issue of mandatory disclosure[67] that had previously been studied at least as far back as George Stigler's famous 1964 study.[68] Contrary to findings in much of the previous literature that had not employed counterfactual comparison groups, these authors found that mandatory disclosure had significant effects on the returns of the affected firms.

Finally, we flag another problem that arises in event studies with few events, such as Klick and Sitkoff's Hershey study, as well as many event studies used in securities litigation. Such studies examine relatively few independent events—perhaps as few as one. In such a context, standard approaches to statistical inference are problematic, since they rely either on (i) the assumption that securities' excess returns follow a normal distribution, or (ii) appropriate application of a central limit theorem that can be justified only when there are a large number of event dates. The normality assumption has been roundly rejected empirically, and a single event is obviously not an appropriate basis for application of asymptotic results such as a central limit theorem. Non-parametric solutions to this problem have been suggested,[69] but, so far, they have not been widely employed in the literature.

A. Civil Procedure

Empirical evidence can play an important role in civil procedure. For example, the Rules Enabling Act process[70] through which the Federal Rules of Civil Procedure are periodically amended, relies heavily on an Advisory Committee process in which empirical contentions are frequently made and advanced. Further, the Federal Judicial Center ("FJC"), which is the research arm of the federal judiciary, regularly releases studies related to the performance of federal courts. In recent years, such FJC activity has

[66] Vladimir Atanasov and Bernard Black (2014). "Shock-Based Causal Inference in Corporate Finance Research," Northwestern Law and Economics Research Paper 11-08.
[67] Michael Greenstone, Paul Oyer, and Annette Vissing-Jorgenson (2006). "Mandated Disclosure, Stock Returns, and the 1964 Securities Acts Amendments," Quarterly Journal of Economics, 121(2): 399-460.
[68] George J. Stigler (1964). "Public Regulation of the Securities Markets," Journal of Business, 37(2): 117-142.
[69] See Gelbach, Helland & Klick (2013), note 64, above.
[70] See 28 U.S.C. §§ 2071-2077.

included studies[71] prepared as a result of a request by the Advisory Committee on Civil Rules in response to the Supreme Court's controversial *Twombly*[72] and *Iqbal*[73] decisions related to the pleading standard in federal civil actions.

In *Twombly* and *Iqbal* together, the Supreme Court announced a new "plausibility" standard for pleading claims in federal civil actions, regardless of their subject matter, replacing the logical-possibility test implicit in the *Conley v. Gibson*[74] no-set-of-facts standard. When a complaint fails the plausibility standard as to a claim, the claim will be dismissed in the face of a defendant's Rule 12(b)(6) motion to dismiss. These developments have spawned a cottage industry seeking to determine something quantitative about how things have changed as a result of those cases. The bulk of this empirical literature has focused on comparing one or another measure of how frequently district court grants occur in Rule 12(b)(6) motions to dismiss for failure to state a claim.[75] Authors finding an increase in their chosen measure of the grant rate have concluded that the pleading standard has been tightened, especially in constitutional cases.[76] Those finding no statistically significant increase have concluded that not much about judicial behavior has changed.[77]

These studies are plagued by an important problem: as many observers observed following *Twombly* and *Iqbal*, the apparent change in the pleading standard can be expected to change parties' behavior.[78] Other things equal, defendants can be expected to file more Rule 12(b)(6) motions; plaintiffs can be expected to file fewer cases in the first place; and the set of cases that settle—whether before or after the plaintiff files suit—can be expected to change in complicated ways. Gelbach has shown that, without strong assumptions, one cannot conclude *anything* about the change in judicial behavior even if one knows, simultaneously, the direction of change in (i) the Rule 12(b)(6) grant rate, (ii) the share of filed cases in which Rule 12(b)(6) motions are filed, and (iii) the number of cases filed.[79]

The basic problem here can be thought of in terms of omitted variables bias. A simple before-after comparison in the grant rate is equivalent to estimating a univariate regression model in which the only variable besides the constant is a dummy variable indicating whether a case's Rule 12(b)(6) motion was adjudicated before or after *Twombly* and *Iqbal*. The changes in party behavior plausibly induced by *Twombly* and *Iqbal* might well change the quality distribution of those cases that are observed actually facing Rule 12(b)(6) motions. This quality-change effect winds up in the residual of the aforementioned

[71] JOE S. CECIL ET AL., FED. JUDICIAL CTR., MOTIONS TO DISMISS FOR FAILURE TO STATE A CLAIM AFTER IQBAL: REPORT TO THE JUDICIAL CONF. ADVISORY COMM. ON CIVIL RULES (2011) (the "initial Cecil report"); JOE S. CECIL ET AL., FED. JUDICIAL CTR., UPDATE ON RES. OF RULE 12(B)(6) MOTIONS GRANTED WITH LEAVE TO AMEND: REPORT TO THE JUDICIAL CONF. ADVISORY COMM. ON CIVIL RULES (2011) (the "updated Cecil report").

[72] *Bell Atlantic Corp. v. Twombly*, 550 U.S. 544 (2007).

[73] *Ashcroft v. Iqbal*, 556 U.S. _, 129 S. Ct. 1937 (2009).

[74] 355 U.S. 41 (1957).

[75] For an enumeration and discussion of many of these papers, see David Freeman Engstrom, "The Twiqbal Puzzle and Empirical Study of Civil Procedure," 65 STAN. L. REV. 1203, 1214-17 (2013).

[76] *See, e.g.,* Patricia Hatamyar Moore, An Updated Quantitative Study of Iqbal's Impact on 12(b)(6) Motions, 46 U. RICH. L. REV. 603 (2012).

[77] *See, e.g.*, the initial and updated Cecil reports in note 71, above, as well as William H. J. Hubbard, "Testing for Change in Procedural Standards, with Application to Bell Atlantic v. Twombly," 42 J. LEGAL STUD. 35 (2013).

[78] See Jonah B. Gelbach, "Locking the Doors of Discovery?" 121 YALE L.J. 2270 (2012) and Gelbach (2014), note 41, above.

[79] Jonah B. Gelbach, "Selection in Motion: A Formal Model of Rule 12(b)(6) and the Twombly-Iqbal Shift in Pleading Policy" (2012), available at http://ssrn.com/abstract =2138428.

univariate regression model, and since it is correlated with the time dummy, the resulting estimate of the time dummy's coefficient cannot be interpreted as a consistent estimate of the causal effect of *Twombly* and *Iqbal* on judicial behavior.

Some authors have tried to address this issue by including covariates. For example, Joe Cecil and coauthors include district dummies and a dummy indicating whether the challenged complaint had been amended. But the amended complaint dummy is itself endogenous to pleading standard-related changes in judicial behavior, because numerous studies find evidence that the rate at which Rule 12(b)(6) motions were granted with leave to amend changed differently from the rate of grants without such leave. Nor is it plausible that district dummies control for all relevant changes in quality composition.[80]

In another study, William Hubbard tried to deal with selection by restricting attention to cases that were filed just before *Twombly* was announced, and comparing Rule 12(b)(6) outcomes in those cases to cases filed on the same calendar dates a year earlier.[81] This approach might eliminate selection related to plaintiffs' choice not to file some cases under a more demanding pleading standard than they would under the *Conley* standard. But it does not address the problem of defendants' incentive to file additional Rule 12(b)(6) motions, which comes into play only after cases are filed, and thus possibly after *Twombly*'s announcement, even for Hubbard's "treatment" group. It also does not control for changes in settlement that might occur among treatment cases after *Twombly*'s announcement.

Thus, the studies just described are characterized by problematic strategies for measuring the change in judicial behavior. But that isn't their only limitation. Gelbach also points out that *Twombly* and *Iqbal* would affect parties in ways not captured by changes in judicial behavior. Changes in party litigation behavior can be expected to affect case outcomes over and above any change in judicial behavior. For example, consider disputes in which plaintiffs would file suit under either pleading standard, but in which defendants file Rule 12(b)(6) motions only under the *Twombly*/*Iqbal* standard. Some of these motions no doubt will be granted under *Twombly*/*Iqbal*, in which case *Twombly* and *Iqbal* will have but-for caused the plaintiffs not to get past the answer/Rule 12(b)(6) stage.

The tendency of studies to ignore this kind of effect is troubling.[82] This observation points to a larger issue, and one on which it is fitting to close our substantive discussion. As Barry Friedman wrote not too

[80] For discussion of some other issues related to controls in the Cecil et al reports, see Jonah B. Gelbach, "Can We Learn Anything About Pleading Changes From Existing Data?," International Review of Law & Economics (forthcoming), responding to various claims made in Joe Cecil, "Of Waves and Water: A Response to Comments on the FJC Study Motions to Dismiss for Failure to State a Claim after Iqbal" (2012), available at http://ssrn.com/abstract=2026103.

[81] Hubbard (2013), "Testing for Change".

[82] As Gelbach (2012) discusses, it is impossible to point-identify even the combined impact of this type of effect and judicial behavior effects. However, Gelbach does show how a lower bound can be constructed on the combined impact. In such situations, constructing bounds is a potentially fruitful avenue for future exploration. On the other hand, Gelbach's (2012) approach is identified purely from time series variation, and it is possible that the composition of cases filed changed after *Twombly* and *Iqbal* for reasons other than the change in the pleading standard (the long economic downturn that began in late 2007 is an obvious potential explanation). One study has used data from Nebraska, which adopted the plausibility standard for civil pleading in state courts, together with data from comparison states to try to deal with such concerns; *see* Roger Michalski and Abby K. Wood, *Twombly and Iqbal at the State Level* (July 31, 2014). USC Law Legal Studies Paper No. 14-30, http://ssrn.com/abstract=2468864. This study uses cutting edge techniques to define the comparison group and

long ago, “Normative bite ought to define the problem, not be an afterthought. Falsifiable hypotheses should be about something of consequence.”[83] One important lesson of the *Twombly* and *Iqbal* literature, then, is that empirical studies must clearly state and explain why their object of empirical study is worth studying.[84]

VI. Looking Forward

Empirical work in law and economics has come a long way over the last quarter century or so. Just as credibility revolution brought additional attention to identification concerns across a range of fields in microeconomics generally, empirical researchers in law and economics have begun to focus seriously on the nature and quality of the variation that identifies would-be estimates of policy effects. This trend has been especially strong in certain law and economics subfields, with the economic analysis of crime being a particular success story. As we discussed in Part IV, we now have multiple studies of the effects of police on crime, using a variety of identifying and estimation approaches.

Beyond our observation in Part V that some topic areas have plenty of room for improvement, we have not dwelled much on developments in the future of empirical law and economics. To do so at length would be beyond the scope of this chapter, but we finish by pointing to two methodological developments in empirical microeconomics generally that might profitably be imported into law and economics research. First, there has been a substantial increase in attention to distributional effects, i.e., effects that cannot be expressed simply as operating on conditional means. Measuring such effects gives a more complete, and sometimes very different, understanding of policy effects.[85] It would be good for law and economics researchers to do more on this front. Second, there has been an explosion in sophisticated structural econometric modeling in empirical industrial organization.[86] Structural methods can deliver valuable policy-measuring information, and it would be good to see them taken seriously in law and economics scholarship.

does a very nice job of implementing its approach. On the other hand, the approach itself is vulnerable to a number of the selection-related critiques described above. Perhaps the most important of these is the observation from Gelbach’s (2012) *Selection in Motion* paper (*see* note 79, above) that empirical findings concerning the number of cases filed, the number of Rule 12(b)(6) motions filed, and the grant rate among cases facing these motions are by themselves uninformative concerning whether judicial behavior has changed (*see* text above). Thus the Michalski and Wood study does a nice job of dealing with potential confounders in measuring the relationship between changes in Nebraska’s pleading standard and a number of outcomes, but Gelbach’s result suggests that these measurements may tell us nothing about the underlying question of interest.

[83] Barry Friedman, Taking Law Seriously, 4 PERSP. ON POL. 261, 263 (2006).

[84] See also Joshua B. Fischman, Reuniting ‘Is’ and ‘Ought’ in Empirical Legal Scholarship, 162 U. PA. L. REV. 117 (2013).

[85] *See, e.g.,* Bitler, Gelbach & Hoynes (2006), note 12 above.

[86] *See* note 23.