ARTICLE

REUNITING 'IS' AND 'OUGHT' IN EMPIRICAL LEGAL SCHOLARSHIP

JOSHUA B. FISCHMAN†

†Associate Professor, Northwestern University School of Law, joshua.fischman@law.northwestern.edu. I thank David Abrams, Karen Alter, Robert Anderson, Charles Barzun, Bernard Black, Miguel de Figueiredo, Shari Diamond, David Freeman Engstrom, Tim Feddersen, Brandon Garrett, Jonah Gelbach, Michael Gilbert, Jim Greiner, Tara Grove, Mitu Gulati, Toby Heytens, Josh Kleinfeld, Andrew Koppelman, Brian Leiter, Greg Mitchell, Laura Pedraza-Fariña, Nicola Persico, Jim Pfander, J.J. Prescott, Corey Rayburn Yung, Ed Rubin, Max Schanzenbach, David Schwartz, Micah Schwartzman, Scott Shapiro, Greg Sisk, Matthew Stephenson, Sean Sullivan, and audiences at Vanderbilt Law School, Northwestern Law School, the University of Illinois College of Law, and Michigan Law School for their helpful feedback.

INTRODUCTION

A century ago, Roscoe Pound set forth his agenda for a "sociological jurisprudence"[1] that would study the "actual social effects of legal institutions and legal doctrines."[2] Pound sought to use empirical social science to advance normative goals: he "regard[ed] law as a social institution which may be improved by intelligent human effort" and proposed that social science could "discover the best means of furthering and directing such effort."[3] Two decades later, Karl Llewellyn issued his call for a "realistic jurisprudence" that would use empirical social science to study the determinants and consequences of judicial decisions.[4] Llewellyn was also motivated by normative ends, believing that in order to investigate whether the "law does what it ought," one must "first answer what it is doing now."[5] Pound and Llewellyn sparred over their respective visions,[6] but it is important to remember that they shared a common aim: to use empirical social science to *improve* the law.

The early legal empiricists were mindful of the challenges of connecting positive and normative approaches to legal scholarship. In his exchange with Pound, Llewellyn famously called for a "*temporary* divorce of Is and Ought."[7] He believed that the separation of 'is' and 'ought' was necessary for scientific credibility, but that it must be temporary in order to serve the

---

1 *See* Roscoe Pound, *The Scope and Purpose of Sociological Jurisprudence* (pts. 1-3), 24 HARV. L. REV. 591 (1911); 25 HARV. L. REV. 140 (1912); 25 HARV. L. REV. 489 (1912) [hereinafter Pound, *Scope and Purpose*].

2 Pound, *Scope and Purpose* (pt. 3), *supra* note 1, at 513.

3 *Id.* at 516.

4 *See* Karl N. Llewellyn, *A Realistic Jurisprudence—The Next Step*, 30 COLUM. L. REV. 431 (1930).

5 Karl N. Llewellyn, *Some Realism About Realism—Responding to Dean Pound,* 44 HARV. L. REV. 1222, 1223 (1931).

6 *See id.* at 1226-33; Roscoe Pound, *The Call for a Realist Jurisprudence*, 44 HARV. L. REV. 697 (1931); Llewellyn, *supra* note 4, at 433-35.

7 Llewellyn, *supra* note 5, at 1236.

goals of legal reform.[8] In the years that followed, however, legal empiricists struggled to balance the competing demands of social science and legal reform.[9] Some failed to separate 'is' and 'ought,' allowing their normative commitments to influence their factual findings.[10] Others failed to reunite 'is' and 'ought,' producing "a mindless amassing of statistics without reference to any guiding theory whatsoever."[11] Years later, a disillusioned Llewellyn mocked his fellow realists for their pointless empirical projects.[12] He wrote: "I read all the results, but I never dug out what most of the counting was good for."[13]

The early legal empiricists had worthy ambitions, but their accomplishments were meager.[14] There were many reasons for their failure,[15] but prominent among them was their inability to develop any kind of theoretical framework for making their empirical findings relevant to normative

---

8 *See id.* at 1236-37 (arguing that "during the inquiry itself into what Is, the observation, the description, and the establishment of relations between the things described are to remain *as largely as possible* uncontaminated by the desires of the observer," but that for those "who begin with a suspicion that change is needed, a permanent divorce would be impossible").

9 *See, e.g.*, John Henry Schlegel, *American Legal Realism and Empirical Social Science: From the Yale Experience*, 28 BUFF. L. REV. 459, 539-45 (1979) (discussing the struggles faced by legal realists at Yale Law School in balancing the methodological imperatives of social science with the demands of progressive reform).

10 *See id.* at 540-45 (describing how several of the realists, most notably William O. Douglas, abandoned the scientific method when it conflicted with their reform objectives).

11 S.N. Verdun-Jones, *Cook, Oliphant, and Yntema: The Scientific Wing of American Legal Realism*, 5 DALHOUSIE L.J. 3, 43 (1979).

12 *See* Karl N. Llewellyn, *On What Makes Legal Research Worth While*, 8 J. LEGAL EDUC. 399, 401 (1956) (describing studies by Walter Wheeler Cook and Herman Oliphant as "hastily considered, ill-planned, mal-prepared . . . so-called research" and a study by Underhill Moore as "the nadir of idiocy").

13 *Id.* at 403.

14 *See* NEIL DUXBURY, PATTERNS OF AMERICAN JURISPRUDENCE 158 (1995) ("Legal realists made a good deal of fuss about bringing social sciences to the law schools. But they did disappointingly little with such sciences once they had got them there."); MORTON HORWITZ, THE TRANSFORMATION OF AMERICAN LAW 1870-1960 210 (1992) ("Virtually all [legal historians] agree that most of the social science research projects undertaken by Realists were either trivial attempts to prove the obvious through pseudo-scientific methodology or else naive and misconceived efforts at social science research."); Harold D. Lasswell & Myres S. McDougal, *Legal Education and Public Policy: Professional Training in the Public Interest*, 52 YALE L.J. 203, 205 (1943) (describing the realists' empirical scholarship as producing "isolated and trivial results"); Peter H. Schuck, *Why Don't Law Professors Do More Empirical Research?*, 39 J. LEGAL EDUC. 323, 330 (1989) (noting that many of the early empiricists "regarded [their] projects largely as failures").

15 *See* Schlegel, *supra* note 9, at 460 ("[T]he Realists' social scientific research died out because of the impermanence of the institutionalized circumstances in which it was undertaken, the peculiarities of the personalities of the leaders of the undertaking, and the difficulties in matching the impulse to do such research with the social science of the time.").

legal scholarship.[16] Today, empirical legal scholarship is flourishing again,[17] and contemporary empiricists are far more sophisticated than their predecessors. Many law professors now have advanced social science training[18] and employ sophisticated methodologies from other disciplines to analyze and interpret data. Like the early empiricists, however, they are still struggling to balance the methodological imperatives of social science with the desire for legal reform. Often, the quest for scientific credibility leads contemporary empiricists to lose sight of the normative goals of legal scholarship. Some empirical studies make efforts to relate their findings to normative questions about law, and some even offer policy prescriptions, but such studies rarely explain how they derive an 'ought' from an 'is.' Even a cursory examination of the premises underlying such claims often reveals them to be untenable.

Empirical research projects need not generate *immediate* prescriptions, but even positive legal research should address topics that have some importance for legal scholarship. Because the law is a normative practice and exists to serve social purposes, determining what is *important* in legal scholarship requires some reference to the normative goals of law.[19] Thus, any empirical research that purports to be relevant to legal scholarship requires some framework for connecting 'is' and 'ought.'

---

16 *See* RICHARD A. POSNER, OVERCOMING LAW 19 (1995) ("The empirical projects of the legal realists, which not only failed but in failing gave empirical research rather a bad name among legal academics, illustrate the futility of empirical investigation severed from a theoretical framework."); JOHN HENRY SCHLEGEL, AMERICAN LEGAL REALISM AND EMPIRICAL SOCIAL SCIENCE 234 (1995) (describing the lack of "conceptual schema that could explain the results of the Realist's research"); Brian Leiter, *Rethinking Legal Realism: Toward a Naturalized Jurisprudence*, 76 TEX. L. REV. 267, 311-12 (1997) (arguing that the problem with realism was not the lack of a theoretical framework, but "rather adherence to a bad theoretical framework"); Llewellyn, *supra* note 12, at 401 ("There was a rich and reeking failure . . . in finding ideas or words of common ground to translate legal problems or phenomena into meaningfulness to the social disciplines or to interpret social discipline concepts or methods into anything with meaning and appeal to men of law."); Verdun-Jones, *supra* note 11, at 43 (describing the failure of legal realists "to establish even the most rudimentary conceptual framework capable of ordering empirical information into a meaningful form").

17 *See* Shari Seidman Diamond & Pam Mueller, *Empirical Legal Scholarship in Law Reviews*, ANN. REV. L. & SOC. SCI. 581, 590-91 (2010) (noting the growth of empirical scholarship in law reviews between 1998 and 2008); Tracey E. George, *An Empirical Study of Empirical Legal Scholarship: The Top Law Schools,* 81 IND. L.J. 141, 147 (2006) (documenting the increased use of empirical terms in law review articles between 1994 and 2006); Michael Heise, *An Empirical Analysis of Empirical Legal Scholarship Production, 1990-2009*, 2011 U. ILL. L. REV. 1739, 1741-46 (documenting growth in empirical terms in law review titles).

18 *See* Joni Hersch & W. Kip Viscusi, *Law and Economics as a Pillar of Legal Education*, 8 REV. L. & ECON. 487, 489 (2012) (reporting that 20% of faculty members at the 26 highest-ranked law schools have a Ph.D. in a social science discipline).

19 *See infra* Section V.A.

As Barry Friedman has prominently argued, empirical legal scholars should "ask, at the outset of every project, why we . . . might care about what is being studied."[20] Yet it is not enough to admonish legal empiricists to pay more attention to normative implications. In many settings, there are complex relationships between the phenomena that are readily measured and the values that can justify legal reform. Intuition alone cannot suffice to relate observable data to normative claims; legal scholarship needs conceptual frameworks and empirical methods that can bridge the gap between 'is' and 'ought.' Developing such frameworks will require a sustained agenda that integrates empirical methodology with legal theory.

Part I of this Article begins by considering how other disciplines have developed methods for relating quantitative empirical findings to normative claims. Typically, this is accomplished by formulating a normative metric that quantifies the goodness of the results. Using medicine and economics as examples, Part I shows how scholars in these disciplines have developed frameworks and methods for connecting the positive and the normative.

Empirical legal scholars, by contrast, often seek normative relevance by examining measureable phenomena that have some intuitive but only vaguely specified connection to a normative goal. Many studies simply conflate the measureable with the good, justifying policy proposals on the basis of the measureable objects. Parts II–IV provide illustrations of this approach for three commonly discussed judicial statistics. Part II focuses on judicial citation counts, Part III examines reversal rates, and Part IV critiques measures of interjudge disparity. These statistics are often used in empirical legal scholarship to capture conceptions of good judicial decisionmaking, and all three have been used to justify bold policy proposals.

For example, scholars have argued that judicial citation counts should be used to determine a shortlist for Supreme Court nominations,[21] to assess the merits of judicial selection procedures,[22] to determine whether judges are

---

20 Barry Friedman, *Taking Law Seriously*, 4 PERSP. ON POL. 261, 262 (2006).

21 *See* Stephen Choi & Mitu Gulati, *A Tournament of Judges?*, 92 CALIF. L. REV. 299, 300 (2004) [hereinafter Choi & Gulati, *Tournament*] (proposing a citation count–based system for evaluating judges "where the reward to the winner is elevation to the Supreme Court"); Stephen J. Choi & G. Mitu Gulati, *Choosing the Next Supreme Court Justice: An Empirical Ranking of Judge Performance*, 78 S. CAL. L. REV. 23, 34 (2004) [hereinafter Choi & Gulati, *Empirical Ranking*] (explaining how judicial citation counts can be used to select Supreme Court nominees).

22 *See* Stephen J. Choi et al., *Professionals or Politicians: The Uncertain Empirical Case for an Elected Rather than Appointed Judiciary*, 26 J.L. ECON. & ORG. 290, 326 (2008) (using citation counts to compare the performance of appointed and elected state court judges).

overpaid,[23] and even to examine whether men or women make better judges.[24] Studies documenting interjudge disparities played a prominent role in the enactment of the United States Sentencing Guidelines[25] and have also been used to justify reforms in Social Security[26] and immigration adjudication.[27] Reversal rates have been cited in debates about whether to split the Ninth Circuit[28] and used to appraise reforms in immigration adjudication.[29] Because such measures lack intrinsic normative force, however, policy arguments based on these measures alone are untenable. These measures may well have some relevance to normative concerns, but the studies are seldom explicit about their normative goals, how the data relate to these goals, and what premises are needed to justify the conclusions.

Part V discusses ways that legal empiricists can bridge the gap between 'is' and 'ought.' Most fundamentally, legal empiricists need to prioritize normative questions; research should focus on what is important, not what is easily measureable. In addition, empiricists need to rethink some aspects of empirical legal methodology. The choice of methods should be driven by questions, not the other way around. Empiricists should not try to seek objective, assumption-free conclusions, but rather should indicate how findings can be combined with assumptions to generate meaningful conclusions. Finally, due to the nature of the questions that arise in legal scholarship and the limits of experimentation, legal scholars should pay more attention to how their findings can generalize to new settings.

---

23 Stephen J. Choi et al., *Are Judges Overpaid? A Skeptical Response to the Judicial Salary Debate*, 1 J. LEGAL ANALYSIS 47, 67 (2009) (using citation counts to measure the impact of salaries on judicial performance).

24 *See* Stephen J. Choi et al., *Judging Women*, 8 J. EMPIRICAL LEGAL STUD. 504, 508 (2011) (using citation counts to assess whether men or women make better judges).

25 *See* KATE STITH & JOSÉ A. CABRANES, FEAR OF JUDGING: SENTENCING GUIDELINES IN THE FEDERAL COURTS 104-12 (1998) (describing how empirical studies of sentencing disparity played a prominent role in the enactment of the United States Sentencing Guidelines).

26 *See* JERRY L. MASHAW, BUREAUCRATIC JUSTICE: MANAGING SOCIAL SECURITY DISABILITY CLAIMS 22 (1983) (discussing reports that advocated reforms to address the failure of the Social Security Administration "to manage the adjudication of claims in ways that produce predictable and consistent outcomes").

27 *See* Jaya Ramji-Nogales et al., *Refugee Roulette: Disparities in Asylum Adjudication*, 60 STAN. L. REV. 295, 325-49 (2007) (documenting wide disparities among immigration judges in asylum grant rates); *id.* at 378-89 (weighing various reforms for reducing the disparities).

28 See *infra* note 115 and accompanying text.

29 See *infra* notes 120-123 and accompanying text.

I. RELATING THE MEASURABLE TO THE GOOD

Empirical research is inherently descriptive, yet legal scholarship is predominantly normative.[30] Bridging the gap between 'is' and 'ought' therefore requires some form of normative premise. When empirical legal scholars seek to relate their empirical findings to normative claims about the law or legal institutions, however, their claims often have vague, unstated foundations. There is frequently a striking contrast between the effort devoted to making credible statistical inferences and the lax attitude toward articulating premises that can connect empirical findings to normative claims about law.

The challenge of relating empirical findings to normative claims is hardly unique to legal scholarship. Many professional disciplines and applied sciences—such as medicine, engineering, education, and environmental studies—harness scientific knowledge in the pursuit of social purposes. Although most empirical research in the social science disciplines is positive, the research questions of these disciplines are similarly motivated by normative ends.

This Part discusses the use of normative metrics in disciplines other than law. In some settings, the relevant metrics are directly measureable, and the results are self-interpreting. This part then examines frameworks for connecting 'is' and 'ought' in medical research and in economics, which use more sophisticated theories and methods to relate empirical findings to normative goals. In contrast to law, scholars in these disciplines are explicit about how empirical findings are used to support normative claims.

A. *Normative Metrics*

In quantitative studies, a normative premise is typically formulated in terms of a metric that maps states of the world into levels of goodness. A function f would constitute an normative metric if $f(A) > f(B)$ whenever state A is preferred over state B. In economics, for example, the function f typically represents economic surplus or some conception of social welfare. Similarly, research on criminal justice might evaluate policing policies in terms of crime rates,[31] medical researchers examine health outcomes and

---

30 Edward L. Rubin, *The Practice and Discourse of Legal Scholarship*, 86 MICH. L. REV. 1835, 1847 (1988) ("[T]he most distinctive feature of standard legal scholarship is its prescriptive voice.").

31 *See, e.g.*, Steven D. Levitt, *Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime*, 87 AM. ECON. REV. 270 (1997) (finding that an increase in the size of police forces reduces violent crime).

survival rates,[32] and education research often examines academic achievement.[33]

Any policy claim derived from an empirical study is only as credible as the normative metric that is employed. Justifying a metric requires two steps. First, one needs a theory of the good. For example, the metrics described above are premised on the desirability of low crime, economic efficiency, good health, or academic achievement. These normative premises are uncontroversial, even if there might be disagreement about how tradeoffs should be made among competing goals.

Second, one needs to relate observable phenomena to the measure of goodness. When good or bad outcomes are directly measurable—such as when the outcomes of a medical trial are "survival" and "death"—the results will be self-interpreting and no deeper theory is needed. If such a trial is well controlled, simple statistical methods may be adequate to assess the impact of a treatment and to justify prescriptive claims.

In many settings, however, the normative metric will not be directly measureable, but rather must be inferred from other observable variables. In these settings, more complex inferential methods and deeper theories are needed to justify normative claims. The following Sections will discuss concepts and methods that other disciplines have developed to relate measureable outcomes to normative claims.

B. *Medical Research*

Medicine is a prominent example of a discipline that is both scientific and prescriptive. Medical research uses scientific methods to examine the effects of various treatments, but the practice of medicine has explicit normative goals: the "diagnosis, treatment, and prevention of disease."[34] Thus, the commonly accepted metrics for evaluating medical treatments are outcomes that represent "how a patient feels, functions, or survives."[35]

---

32 *See, e.g.*, Biomarkers Definitions Working Group, *Biomarkers and Surrogate Endpoints: Preferred Definitions and Conceptual Framework*, 69 CLINICAL PHARMACOLOGY & THERAPEUTICS 89, 91 (2001) (defining a "[c]linical endpoint . . . used in the assessment of the benefits and risks of a therapeutic intervention" as "[a] characteristic or variable that reflects how a patient feels, functions, or survives").

33 *See, e.g.*, Cecilia Elena Rouse, *Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program*, 113 Q.J. ECON. 553 (1998) (examining the impact of assignment to a private school on test scores in the context of a school voucher program).

34 NEW OXFORD AMERICAN DICTIONARY 1087 (3d ed. 2010).

35 Robert Temple, *Are Surrogate Markers Adequate to Assess Cardiovascular Disease Drugs?*, 282 J. AM. MED. ASS'N 790, 790 (1999).

When these outcomes are directly measureable, the normative implications of a medical trial may be obvious. Often, however, these normatively relevant outcomes cannot be readily measured, especially when the effects of a treatment may not accrue until many years after the treatment is administered. In such trials, medical researchers often use "surrogate outcomes" to proxy for the clinically meaningful outcomes. For instance, when the effect of a drug regimen on heart disease and life expectancy might not be observed for many years, a study might examine whether the drug regimen significantly reduces levels of blood cholesterol. Here, blood cholesterol is a surrogate; it has no normative significance beyond its tendency to promote coronary disease.

In this example, a treatment cannot be justified merely on the basis of its estimated effect on cholesterol levels. To justify an intervention, any surrogate measure must be validated by showing that an effect of the treatment on the surrogate will correspond to an effect on a meaningful clinical outcome. Validation of the surrogate measure requires two steps. First, one must specify the clinical outcome that the surrogate is intended to measure, such as survival, comfort, or functional capacity.[36] Second, one must explain the relationship between the surrogate and the clinical outcome and show how inferences about the former can facilitate inferences about the latter. This step requires both a statistical association between the surrogate and the clinical outcome and an understanding of the causal relationship between the two.

Biostatisticians have developed a rich literature on the use of surrogate measures, providing a variety of complex conditions under which surrogates can be used to support inferences about true outcomes.[37] In particular, correlation between the surrogate and the true measure is not sufficient to justify the use of a surrogate in a clinical trial.[38] Often, there may be multiple causal pathways between a disease and a true clinical outcome, only one of which is captured by the surrogate measure. In such a situation,

---

36 *See id.*

37 *See, e.g.*, Marc Buyse & Geert Molenberghs, *Criteria for the Validation of Surrogate Endpoints in Randomized Experiments*, 54 BIOMETRICS 1014, 1014-16 (1998) (discussing criteria for the proper use of surrogates in clinical studies); Thomas R. Fleming & David L. DeMets, *Surrogate End Points in Clinical Trials: Are We Being Misled?*, 125 ANNALS INTERNAL MED. 605, 605-06 (1996) (same); Ross L. Prentice, *Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria*, 8 STAT. MED. 431, 431-32 (1989) (same). *See generally* THE EVALUATION OF SURROGATE ENDPOINTS (Tomasz Burzykowski et al. eds., 2005) (discussing the use and validation of surrogates in a variety of contexts).

38 *See* Stuart G. Baker & Barnett S. Kramer, *A Perfect Correlate Does Not a Surrogate Make*, 3 BMC MED. RES. METHODOLOGY, no. 16, 2003, at 2-3 ("[P]erfect correlation does not guarantee correct inference when a potential surrogate endpoint replaces a true endpoint.").

measuring the impact of a treatment on the surrogate will fail to capture the impact of the treatment on the true outcome.[39]

For example, it would be infeasible to measure the effectiveness of youth anti-smoking programs by examining the proportion of treated youth who die prematurely from lung cancer. Researchers might instead use subsequent smoking behavior as a surrogate for premature lung cancer death.[40] Similarly, reduction in tumor size might be a valid surrogate for survival rates in estimating the impact of chemotherapy regimens on lung cancer patients. Both cigarette smoking and tumor size are highly correlated with lung cancer deaths, and both have a direct causal impact. But smoking rates could not be used as a surrogate for measuring the effectiveness of chemotherapy, and tumor size could not be used as a surrogate for anti-smoking campaigns.[41]

In a number of instances, drugs have been approved on the basis of their effect on surrogate measures but were subsequently discovered to have harmful effects on clinical outcomes.[42] As these experiences show, the relationship between surrogates and meaningful outcomes cannot simply be asserted, but must be carefully scrutinized. Understanding the causal relationship between medical interventions, surrogates, and clinically meaningful outcomes is essential to the validation of any surrogate measure.

C. Economics

Economics, like many of the social sciences, combines positive research with normative goals. Economists study the production, consumption, and distribution of goods and services, but scholarship in economics is not merely motivated by idle curiosity about producers, consumers, and markets. Rather, the study of economics is motivated by an understanding that economic activity serves social purposes and that certain policies may advance or hinder those purposes.

Empirical economists often have access to voluminous data on prices and levels of output for goods in various markets. Such data, however, typically have no intrinsic normative significance; one would not justify a

---

39 *See* Fleming & DeMets, *supra* note 37, at 605-06 (illustrating how a surrogate endpoint might not reflect a true clinical outcome).

40 *See* Prentice, *supra* note 37, at 432-33. Smoking rates also have independent normative validity because many people find smoking to be distasteful, but the health effects of tobacco use are the primary motivation for anti-smoking campaigns.

41 *Id.*

42 *See* Fleming & Demets, *supra* note 37, at 607-08 (discussing how several heart medications, which were approved by the FDA on the basis of their impact on surrogate measures, were subsequently found to increase mortality in clinical trials).

policy merely on the basis of its tendency to affect prices or output levels. Unlike in medicine, there typically are not measureable phenomena that can be used as surrogates for economic wellbeing. To assess the desirability of outcomes, economists have formulated concepts such as consumer and producer surplus, which represent the gains from trade in a market.[43]

Note that surplus is a purely abstract concept, with no analog in the natural world. It is defined by reference to supply and demand curves, which represent the quantities producers would supply and consumers would demand at various counterfactual prices. Surplus, supply curves, and demand curves cannot be physically measured, but rather must be estimated by combining data on prices and output levels at different points in time with theoretical assumptions about consumer and producer behavior.

Because the framework for relating observable data to surplus is so well established,[44] economists do not need to revisit its fundamentals every time they evaluate a proposed policy. Indeed, it can be easy to overlook the assumptions underlying any calculation of surplus.[45] Nevertheless, the concept of surplus allows economists to organize data on prices and output levels into a measure of economic wellbeing that can assess the impact of various policies and provide justification for proposed reforms.

D. *Empirical Legal Scholarship*

In contrast to medicine and economics, legal scholarship lacks frameworks for connecting empirical findings to normative claims. Occasionally, when legal changes can be assessed in terms of outcomes that have direct normative significance, there is no need for sophisticated theory. For example, in studies that examine the impact of tort reforms on medical

---

[43] Economists often use more sophisticated measures of wellbeing as well, in part because measures of surplus do not account for distributional impact. For simplicity, I focus on surplus in the current discussion.

[44] The concepts of consumer and producer surplus were popularized by the economist Alfred Marshall in 1890. *See* ALFRED MARSHALL, PRINCIPLES OF ECONOMICS 175 (1890) (defining consumer surplus); *id.* at 428 (defining producer surplus).

[45] *See* HAL R. VARIAN, INTERMEDIATE ECONOMICS: A MODERN APPROACH 251-52 (7th ed. 2006) (showing that the definition of consumer surplus depends upon the assumption that consumers have quasilinear utility functions); Charles F. Manski, *Monotone Treatment Response*, 65 ECONOMETRICA 1311, 1315-16 (1997) (noting that supply and demand functions are often assumed to be linear as a matter of convenience and that this assumption is not motivated by economic theory).

complications in childbirth[46] or the effect of school desegregation decisions on black dropout rates,[47] the normative significance of the outcomes is clear.

At other times, frameworks from other disciplines are sufficient to evaluate outcomes. For example, studies that evaluate the impact of school finance decisions on academic achievement may use test scores as outcome variables.[48] Similarly, studies may use economic concepts to appraise the welfare impacts of proposed mergers.[49]

The methods of the other disciplines are most likely to be adequate when the outcomes of interest are similar to those that arise in the other disciplines. If one views law purely as a means to achieve policy goals, then the methods of the social sciences can often be used with little adaptation. Studies such as those discussed above only need to comprehend law well enough to understand the timing and expected impact of legal changes. Such studies, however, are conducted from an external point of view; one does not need to take any position on the *validity* of legal events in order to appraise them from a policy perspective.[50]

Some empirical legal research, however, appears to be motivated by values internal to law. Although these values are rarely made explicit, such studies appear to be animated by concerns about deciding cases correctly, treating likes alike, or writing good judicial opinions. Such concepts are not

---

46 *See* Janet Currie & W. Bentley MacLeod, *First Do No Harm? Tort Reform and Birth Outcomes*, 123 Q.J. ECON. 795, 801-04 (2008) (measuring the effects of joint-and-several liability and damage caps on childbirth complications).

47 *See* Jonathan Guryan, *Desegregation and Black Dropout Rates*, 94 AM. ECON. REV. 919 (2004).

48 *See, e.g.*, William J. Glenn, *School Finance Adequacy Litigation and Student Achievement: A Longitudinal Analysis*, 34 J. EDUC. FIN. 247, 249 (2009) ("Generally, measuring student outcomes entails using test scores."); Thomas A. Downes, *Evaluating the Impact of School Finance Reform on the Provision of Public Education: The California Case*, 45 NAT'L TAX J. 405, 414 (1992) ("There is little evidence that outcomes, as measured by test scores, were less unequal after the school finance reforms of the late 1970's.").

49 *See, e.g.*, Orley Ashenfelter et al., *Empirical Methods in Merger Analysis: Econometric Analysis of Pricing in* FTC v. Staples, 13 INT'L J. ECON. BUS. 265, 270-72 (2006) (discussing alternative approaches to measuring lost consumer surplus due to a proposed merger).

50 This is not to deny that these studies are important to policymakers or to legal scholars. Such research is clearly relevant to legislators and administrators, and many contemporary judges and scholars accept that judges do and should consider the policy consequences of their decisions. *See, e.g.*, STEPHEN BREYER, ACTIVE LIBERTY: INTERPRETING OUR DEMOCRATIC CONSTITUTION 18 (2006) (arguing that "judges, in applying a text in light of its purpose, should look to *consequences*, including 'contemporary conditions, social, industrial, and political, of the community to be affected'"); BENJAMIN N. CARDOZO, THE NATURE OF THE JUDICIAL PROCESS 98-141 (1921) (arguing that it is sometimes appropriate for judges to act like legislators); RICHARD A. POSNER, HOW JUDGES THINK 78-91 (2009) (describing judges as "occasional legislators").

directly measurable, however, and the methods of the social sciences are often inadequate for connecting measureable outcomes to these concepts.

Developing methods for evaluating the effects of legal rules and institutions according to criteria internal to law ought to be a priority for legal empiricists. Influential theorists such as Ronald Dworkin,[51] Joseph Raz,[52] and John Rawls[53] have argued that institutions should be evaluated by their tendency to protect rights and promote justice.[54] Many debates about interpretive methods focus on their tendency to promote accurate interpretation, substantive justice, and the rule of law.[55] And the Supreme Court's administrative due process jurisprudence evaluates procedures according to their "capacity for accurate factfinding and appropriate application of substantive legal norms to the facts as found."[56]

The fundamental challenge is that such internal values cannot be directly measured. Instead of developing theories that can relate observable data to these values, as scholars have done in medical research and in economics, empirical studies in law often substitute proxy variables that have some asserted but unspecified connection to the motivating values. Parts II–IV examine three such measures—citation counts, reversal rates, and interjudge

---

51 *See* RONALD DWORKIN, FREEDOM'S LAW: THE MORAL READING OF THE AMERICAN CONSTITUTION 34 (1999) ("I see no alternative but to use a result-driven rather than a procedure-driven standard for deciding [institutional questions]. The best institutional structure is the one best calculated to produce the best answers to the essentially moral question of what the democratic conditions actually are, and to secure stable compliance with those conditions.").

52 *See* Joseph Raz, *Disagreement in Politics*, 43 AM. J. JURIS. 25, 45 (1998) ("A natural way to proceed is to assume that the enforcement of fundamental rights should be entrusted to whichever political decision-procedure is, in the circumstances of the time and place, most likely to enforce them well, with the fewest adverse side effects.").

53 *See* JOHN RAWLS, A THEORY OF JUSTICE 230 (1971) ("[T]he fundamental criterion for judging any procedure is the justice of its likely results.").

54 Others, most notably Jeremy Waldron, reject the view that consequences are dispositive for questions of institutional design. *See* WALDRON, LAW AND DISAGREEMENT 252-54 (2001) (critiquing rights-instrumentalism).

55 *See, e.g.*, ADRIAN VERMEULE, JUDGING UNDER UNCERTAINTY: AN INSTITUTIONAL THEORY OF LEGAL INTERPRETATION 66-67 (2006) (advocating formalism on the ground that it results in fewer errors); William N. Eskridge, Jr., *Norms, Empiricism, and Canons in Statutory Interpretation*, 66 U. CHI. L. REV. 671, 674-84 (1999) (arguing that canons of interpretation should be appraised according to their tendency to promote democratic values, the rule of law, and other substantive normative goals); Cass R. Sunstein & Adrian Vermeule, *Interpretation and Institutions*, 101 MICH. L. REV. 885, 918 (2003) (arguing that formalism should be evaluated by its tendency to avoid "mistakes and injustices"); Cass R. Sunstein, *Must Formalism Be Defended Empirically?*, 66 U. CHI. L. REV. 636, 656-57 (1999) (arguing that formalism should be evaluated according to its tendency to avoid inaccuracy and uncertainty in judicial decisionmaking and to provide good ex ante incentives).

56 Jerry L. Mashaw, *Administrative Due Process: The Quest for a Dignitary Theory*, 61 B.U. L. REV. 885, 895 (1981).

disparities—that empirical scholars commonly use to evaluate judges and legal institutions. Each of these outcome measures has an intuitive relevance to a normative goal, but the relationship is vague and undertheorized. Because scholars are rarely explicit about the relationship between these measures and the intended measure of merit—indeed, the measure of merit is rarely defined—the empirical evidence cannot justify the normative claims.

II. JUDICIAL CITATION COUNTS

In a series of recent articles, Stephen Choi, Mitu Gulati, and various coauthors have advocated the use of quantifiable metrics to address normative questions about judicial appointment, promotion, retention, and compensation. They have argued that nominations to the Supreme Court should be determined on the basis of three empirical indicators: citation counts, the number of opinions authored, and the rate at which judges disagree with colleagues of the same political party.[57] Using these measures in a series of studies, they found evidence that "female judges . . . perform *better* than male judges"[58] and that "elected judges are *superior* to appointed judges."[59] The authors also used these same performance measures to estimate the effects of judicial compensation, finding that "it is as likely that judges are *overpaid* as that they are *underpaid*."[60]

These authors were not the first to apply a quantitative analysis to the study of judicial citations. As early as 1936, one study tabulated the number of citations to each state's courts from other state courts and the U.S. Supreme Court.[61] The goals of the early citation studies, however, were purely descriptive. Scholars typically characterized citation counts as a measure of influence but did not use them to justify prescriptive claims. At times, these measures were given normative interpretations; for example, Judge Richard Posner claimed that Learned Hand's citation counts confirmed that he "was indeed a great judge."[62] But until recently, no one argued that these measures should guide judicial appointments or the design of legal institutions.

---

57 Choi & Gulati, *Tournament*, *supra* note 21, at 305-13; Choi & Gulati, *Empirical Ranking*, *supra* note 21, at 50-67.

58 Choi et al., *supra* note 24, at 505 (emphasis added).

59 Choi et al., *supra* note 22, at 292 (emphasis added).

60 Choi et al., *supra* note 23, at 63 (emphasis added).

61 *See* Rodney L. Mott, *Judicial Influence*, 30 AM. POL. SCI. REV. 295, 308 tbl.VI, 311 tbl.VII (1936).

62 Richard A. Posner, *The Learned Hand Biography and the Question of Judicial Greatness*, 104 YALE L.J. 511, 540 (1994) (book review).

Before we cut judges' pay and jettison judicial independence, however, we should scrutinize how the authors derived their normative claims from their empirical findings. They do not claim that citations themselves are a measure of goodness; in fact, they acknowledge that their measures "do not provide a perfect metric for judging skill"[63] and are merely "rough proxies."[64] Thus, the fact that Judge *A* has more citations than Judge *B* does not directly justify an assertion that that Judge *A* is *better* than Judge *B*.

But once they acknowledge that citations do not actually measure quality, how can they use aggregate comparisons between groups of judges to justify claims about the relative quality of elected judges versus appointed judges, or male judges versus female judges? Citation counts could arguably be viewed as surrogates[65] for some "true" measure of judicial quality, but if so, their use as surrogates must be validated. Choi and Gulati justify the validity of citation counts largely on theoretical grounds, analogizing the body of precedent to a "market" for judicial opinions.[66] Because the "price" of citing opinions is zero, judges will compete on quality. As they put it, "[a]ll judges will cite the best opinions,"[67] and therefore, the best judges will garner the most citations.

Many critics, however, have questioned how well citation counts actually correlate with merit.[68] In addition, there are plausible arguments that citation measures may be correlated with judicial "vices."[69] One claim is that citation counts reward originality, so these measures will reward judges who change the law rather than follow it.[70] Another argument is that unclear opinions may create uncertainty and generate more litigation, thus

---

63 Choi & Gulati, *Empirical Ranking*, *supra* note 21, at 29.

64 *Id.* at 34.

65 *See supra* Section I.B.

66 Choi & Gulati, *Tournament*, *supra* note 21, at 306.

67 *Id.* at 307.

68 *See, e.g.*, Jay S. Bybee & Thomas J. Miles, *Judging the Tournament*, 32 FLA. ST. U. L. REV. 1055, 1058 (2005) ("[W]e question whether the metrics proposed by Professors Choi and Gulati appropriately measure the performance of circuit judges."); Marin K. Levy et al, *The Costs of Judging Judges by the Numbers*, 28 YALE L. & POL'Y REV. 313, 314 (2010) ("We believe that there is now a general consensus that (1) the judicial virtues the legal empiricists set out to measure probably have little bearing on what actually makes for a good judge; and (2) even if they did, the empiricists' chosen variables have not measured those virtues accurately."); Lawrence B. Solum, *A Tournament of Virtue*, 32 FLA. ST. U. L. REV. 1365, 1389 (2005) (characterizing the argument "that citation rate correlates with judicial excellence" as "somewhat obscure"); WERL, *On Tournaments for Appointing Great Justices to the U.S. Supreme Court*, 78 S. CAL. L. REV. 157, 171 (2004) (describing a "considerable gap . . . between what the numbers purport to measure and what they actually measure"); *see also infra* notes 85-110 and accompanying text.

69 Solum, *supra* note 68, at 1389.

70 *See id.* at 1392-93.

generating more citations.[71] Finally, "an opinion notorious for being 'wrong' might also lead to many cites."[72]

Because citation counts might be associated with judicial vices as well as judicial virtues, theory alone cannot validate the use of citation counts as a surrogate for quality. Determining which judicial characteristics constitute virtue and vice is a matter of normative theory. But for any conception of judicial quality, determining whether citations are more strongly associated with virtue or vice is an empirical question. Of course, this is impossible to test without first specifying a normative benchmark.[73]

Many studies simply assume the validity of citation counts as a surrogate for quality, acknowledging that citations could be correlated with judicial vice, but dismissing this possibility as unlikely.[74] A mere positive correlation, however, is not sufficient to validate citations as a surrogate for quality.[75] Scholars who seek to use citation measures to inform policy decisions must be able to convey uncertainty about their assessments of judicial merit. This cannot be done without a nuanced understanding of the relationship between citations and the conception of merit that is employed.

Choi and Gulati also defend the proposed use of citation counts in the selection of Supreme Court justices on the ground that "objective factors will do better than what we have now: a biased and nontransparent process overwhelmed by politics."[76] The use of objective measures, however, cannot displace normative debates about judicial merit. Citation counts cannot be validated as a surrogate without first articulating a conception of merit.

In addition, there are many objective measures that could potentially be used to evaluate judges. How could one choose among them without some

---

71 *See* Frank B. Cross & James F. Spriggs II, *The Most Important (and Best) Supreme Court Opinions and Justices*, 60 EMORY L.J. 407, 421 (2010) ("The first and most common criticism of citation usage is that it fails to capture dispositive rulings that conclusively resolve legal issues."); Montgomery N. Kosma, *Measuring the Influence of Supreme Court Justices*, 27 J. LEGAL STUD. 333, 339 (1998) (noting that unclear precedents may generate more litigation than clear precedents, resulting in more citations).

72 Robert Henry, *Do Judges Think? Comments on Several Papers Presented at the* Duke Law Journal*'s Conference on Measuring Judges and Justice*, 58 DUKE L.J. 1703, 1717 (2009); *see also* Frank B. Cross & Stefanie Lindquist, *Judging the Judges*, 58 DUKE L.J. 1383, 1391 n.25 (2009) ("[S]ome citations may be attributable to 'outrageously' bad decisions.").

73 *See* Solum, *supra* note 68, at 1368 (arguing that it is necessary to answer the normative question of what makes for excellence in judging before formally measuring judicial performance).

74 *See, e.g.*, Choi & Gulati, *Empirical Ranking*, *supra* note 21, at 70 (acknowledging but dismissing the concern that rewarding judges for citations might induce them to write "longer and more complex opinions that provide more citations"); Cross & Spriggs, *supra* note 71, at 421 ("While the 'settled case' phenomenon is theoretically problematic for any citation measure, its existence is questionable.").

75 *See supra* note 38 and accompanying text.

76 Choi & Gulati, *Tournament*, *supra* note 21, at 304.

normative baseline for comparison? To illustrate, compare the citation measures proposed by Choi and Gulati with alternative measures proposed by Robert Anderson.[77] Whereas Choi and Gulati do not distinguish among positive, negative, and neutral citations in their measures, Anderson interprets negative citations as evidence of low quality and ignores neutral citations. Choi and Gulati count only citations from outside a judge's circuit, whereas Anderson counts citations from both inside and outside a judge's circuit. Finally, Choi and Gulati count citations to all opinions authored by a judge, while Anderson counts citations to all decisions in which the judge was on the panel.

Not surprisingly, these two methodologies yield very different rankings.[78] Even if both of these measures could potentially be useful for measuring judicial performance, how could one know which measure to use? Is the difference between the two measures primarily methodological, in the sense that one method might be a better surrogate for a common conception of judicial quality? Or is the difference primarily normative, in the sense that the measures serve as surrogates for competing conceptions of judicial merit?

Anderson characterizes the differences between the measures as both methodological and normative. He justifies the exclusion of negative citations on normative grounds, arguing that negative citations may be appropriate for measuring *influence*, but that only positive citations are appropriate for measuring *quality*.[79] Similarly, he justifies examining panel membership rather than opinion authorship on the grounds that it "capture[s] collegial factors that *should* enter into a measure of *good judging*."[80] But he also claims that part of the difference is methodological, arguing that using panel membership is appropriate because it "mitigate[s] the effects of selection bias in opinion assignment."[81]

To the extent that the difference between the two measures is methodological, one cannot assess which is a better surrogate without specifying a conception of judicial merit. And to the extent the difference is normative,

---

77 *See generally* Robert Anderson IV, *Distinguishing Judges: An Empirical Ranking of Judicial Quality in the United States Courts of Appeals*, 76 MO. L. REV. 315 (2011) (proposing a method of ranking judges that distinguishes between positive and negative citations).

78 *See id.* at 349 ("The results of this analysis differ dramatically from those of prior judge ranking studies.").

79 *Id.* at 325-26; *see also* William M. Landes & Richard A. Posner, *The Influence of Economics on Law: A Quantitative Study*, 36 J.L. & ECON. 385, 389-90 (1993) ("When speaking of influence rather than quality, one has no call to denigrate critical citations.").

80 Anderson, *supra* note 77, at 329 (emphasis added).

81 *Id.*

the use of objective and quantifiable measures cannot displace normative debates about judicial merit. Either way, one cannot choose between these two measures without taking a position in the normative debate that the citation studies are purporting to circumvent.

Nevertheless, citation counts could conceivably be validated *subjectively*. Even though conceptions of judicial quality are inherently subjective, objective data could still be used to inform those subjective judgments. Scholars, for example, could survey informed observers about their perceptions of judges' relative competence or the quality of particular opinions. On certain dimensions of judicial quality, there is likely to be strong agreement. To take some extreme examples, everyone would agree that Chief Justice John Marshall was a greater judge than his contemporary Gabriel Duvall,[82] or that Learned Hand[83] was superior to his colleague Martin Manton, who went to prison for accepting bribes.[84] On other dimensions, however, assessments of judicial quality are likely to be disputed. For example, a comparison of Justices Sotomayor and Alito will likely depend on one's ideological leanings.

Such surveys could reveal the degree to which conceptions of judicial merit are shared and the degree to which they are disputed. The grounds for disagreement could potentially be approximated by a small number of salient dimensions, such as liberalism versus conservatism or pragmatism versus formalism. Empirical studies of citations can never tell us what kind of judge we ought to prefer, but they might conceivably shed light on how judges measure along these dimensions of judicial merit. To the extent that there are commonly shared conceptions of quality, these objective measures might at least be able to distinguish good judges on each side of the ideological spectrum from mediocre ones. Of course, survey responses do not indicate merit in an objective sense, but at least they would correspond to the conceptions of merit that are prevalent in scholarly dialogue or democratic deliberation.

---

82 *See* David P. Currie, *The Most Insignificant Justice: A Preliminary Inquiry*, 50 U. CHI. L. REV. 466, 466 (1983) ("Duvall's performance reveals . . . that he achieved an enviable standard of insignificance.").

83 Hand is widely recognized as one of the greatest judges in American history. *See* GERALD GUNTHER, LEARNED HAND: THE MAN AND THE JUDGE xv (2d ed. 2011) ("Learned Hand is numbered among a small group of truly great American Judges of the twentieth Century."); Posner, *supra* note 62, at 511 (describing Hand as the "third-greatest judge in the history of the United States, after Holmes and John Marshall").

84 *See generally* JOSEPH BORKIN, THE CORRUPT JUDGE 25-93 (1962) (describing Judge Manton's corruption); Allan D. Vestal, *A Study in Perfidy*, 35 IND. L.J. 17 (1959) (same).

It is essential, however, that citation measures be validated as surrogates for some discoverable measure of quality. In theory, it may seem plausible that good judges would be more productive and write better opinions, and that better opinions would generate more citations. But judicial craft is only one factor—and possibly a minor one—in determining how often a case is cited. Even a cursory examination can show that citation counts do not correspond very well to commonly held perceptions of judicial merit.

Consider two canonical torts cases—*Palsgraf v. Long Island Railroad Co.*[85] and *United States v. Carroll Towing Co.*[86]—which are taught in virtually every first-year law school torts class. Judge Cardozo's opinion in *Palsgraf*, which has been described as "[p]erhaps the most celebrated of all tort cases,"[87] has been cited 218 times in published opinions in federal and state courts.[88] Judge Hand's opinion in *Carroll Towing*, which formulated the "Learned Hand rule" for negligence liability and has been described as one of the "two most influential opinions that Hand ever wrote,"[89] has been cited a total of 177 times.[90] By comparison, the opinion in *Bonner v. City of Prichard*,[91] which holds that all Fifth Circuit decisions handed down prior to October 1, 1981 are binding precedent in the Eleventh Circuit, has been cited 4311 times.[92]

Similarly, *Marbury v. Madison*[93] has been cited 252 times in Supreme Court opinions, barely more than once per term.[94] *McCulloch v. Maryland*[95] has been cited 326 times in Supreme Court opinions,[96] less than twice per term. But *United States v. Detroit Timber & Lumber Co.*,[97] which held that the syllabus is not part of the opinion of the Court, has been cited 4362 times in the U.S. Reports.[98]

---

85 162 N.E. 99 (N.Y. 1928).

86 159 F.2d 169 (2d Cir. 1947).

87 William L. Prosser, *Palsgraf Revisited*, 52 MICH. L. REV. 1, 1 (1953).

88 Westlaw search for "162 N.E. 99" in SCT, CTAR, DCTR, and ALLSTATES databases through December 31, 2011.

89 Posner, *supra* note 62, at 513.

90 Westlaw search for "159 F.2d 169" in SCT, CTAR, DCTR, and ALLSTATES databases through December 31, 2011.

91 661 F.2d 1206, 1209 (11th Cir. 1981).

92 Westlaw search for "661 F.2d 1206" in SCT, CTAR, DCTR, and ALLSTATES databases through December 31, 2011.

93 5 U.S. 137 (1803).

94 Westlaw search for "5 U.S. 137" in SCT database through December 31, 2011.

95 17 U.S. 316 (1819).

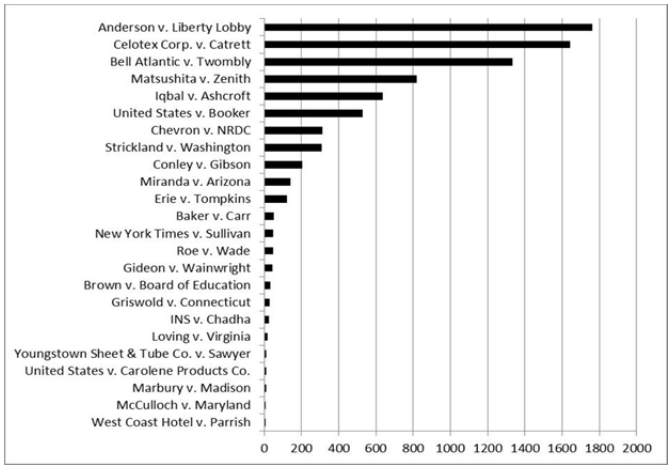
96 Westlaw search for "4 Wheat 316" in SCT database through December 31, 2011.

97 200 U.S. 321, 337 (1906).

98 Westlaw search for "200 U.S. 321" in SCT database through December 31, 2011.

These examples illustrate that frequency of citation does not necessarily correspond to commonly held perceptions of the importance of a holding or the quality of the written opinion. A more detailed comparison of Supreme Court decisions confirms this same pattern. Figure 1 compares a selection of Supreme Court decisions, displaying how often each case was cited per year in reported federal cases.[99] Canonical constitutional cases are dwarfed by holdings on frequently litigated issues such as standards for summary judgment and pleading requirements. *Brown v. Board of Education*[100] is cited 29 times per year, whereas *Anderson v. Liberty Lobby*[101] and *Celotex Corp. v. Catrett*[102]—two decisions providing standards for summary judgment—are each cited more than 1600 times per year. Since it was decided in 1986, *Anderson v. Liberty Lobby* has been cited almost 45,000 times, roughly as many times as every case decided by the Marshall Court combined.[103]

Figure 1: Federal Citations per Year for Selected Supreme Court Decisions



99 Westlaw search of SCT, CTAR, and DCTR databases through December 31, 2011.

100 347 U.S. 483 (1954).

101 477 U.S. 242 (1986).

102 477 U.S. 317 (1986).

103 A sample of pages in Shepard's spanning the Marshall Court yielded an estimate of roughly 45,000 federal citations.

Although judicial merit may well influence how often a judge's opinions are cited, these examples show that citation counts are strongly influenced by factors unrelated to merit, such as how often an issue is presented in litigation. Conceivably, such factors might be less relevant when comparing citation counts at the level of individual judges. If each judge decides a mix of high- and low-profile cases over time, then citation counts aggregated by a judge might conceivably better correlate with commonly held perceptions of merit. Such a claim is difficult to test, largely because judicial merit is contested, and even subjective perceptions are difficult to quantify. But citation statistics for Judge Learned Hand and his contemporaries on the Second Circuit, as reported in an article by Judge Richard Posner,[104] raise serious questions about the validity of these measures, even when aggregated by judge. Using Posner's results, I compiled statistics on opinions authored and citations per year for judges who were active from 1925 until 1939, when Learned Hand and Martin Manton served together.[105] The statistics are based on published majority opinions, and the citation counts only include citations by federal courts of appeals.

It may provide some reassurance that Learned Hand dominates his contemporaries, including Manton, in citations per year. But Manton has more citations per year than highly respected judges such as Thomas Swan and Augustus Hand.[106] Moreover, in opinions per year—the measure of "productivity" used by Choi and Gulati—Manton easily outpaces all of the other Second Circuit judges, including Learned Hand.

---

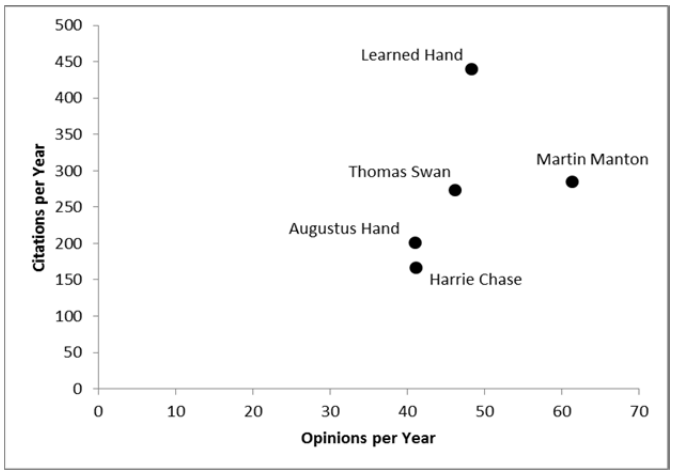
104 *See* Posner, *supra* note 62.

105 All analysis is based on Posner's published results, and not an independent analysis of the original data, which are no longer available.

106 *See, e.g.*, GUNTHER, *supra* note 83, at 242 (stating that once the "Learned Hand-Augustus Hand-Thomas Swan triumvirate was in place," the Second Circuit "symbolized the highest judicial quality in the nation"); MARVIN SCHICK, LEARNED HAND'S COURT 20-27 (1970) (describing widespread respect for Swan and Augustus Hand). Swan had been the dean of Yale Law School before joining the bench.

Figure 2: Opinion Counts and Citations to Opinions Authored by Second Circuit Judges, 1925-39

Note: Only judges who were active for at least 5 years during 1925–39 are included.



Manton's deficiencies were not merely ethical; he was held in low esteem even before evidence of his corruption had surfaced. Learned Hand had a poor opinion of Manton, perceiving him as "incapable of turning out memoranda and opinions that could earn him the respect from the bar or bench."[107] Chief Justice Taft believed that Manton "never should have been appointed to the bench in the first place."[108] Other prominent contemporaries described him as "unfit for the bench"[109] and "one of Wilson's worst appointments."[110] Yet in terms of two quantitative measures commonly used to evaluate judges—"opinion quality" and "productivity"—Manton compares quite favorably to most of his Second Circuit contemporaries. If sufficient weight were given to judicial "productivity," Manton might even rank above Learned Hand.

The fact that such quantitative measures cannot distinguish highly respected circuit judges from a judge widely regarded as one of the worst in history raises serious questions about whether these measures are valid

107 GUNTHER, *supra* note 83, at 237.

108 DAVID J. DANELSKI, A SUPREME COURT JUSTICE IS APPOINTED 45 (1964).

109 *Id.* at 51 (describing opinion of Attorney General Harry Daugherty).

110 *Id.* at 63 (quoting Elihu Root).

surrogates for quality. Perhaps a more careful analysis of judicial citations might yield useful information about some conceptions of judicial merit. The analysis here, for example, did not distinguish between positive and negative citations, or between in-circuit and out-of-circuit citations. It may also be possible to control for outlier opinions that are highly cited, such as those involving summary judgment. Various statistical adjustments could potentially lead to more refined citation measures that more accurately reflect some conception of judicial merit. The multiplicity of possible adjustments, however, presents a choice of which to apply, which requires some external conception of merit against which the various adjustments can be compared.

III. REVERSAL RATES

Reversal rates are a commonly used outcome measure in empirical studies of judicial decisionmaking and are widely used to justify normative claims about judges and courts. In the last decade, more than 1000 law review articles included some mention of reversal or affirmance rates, although many uses were purely descriptive.[111] Like citation counts, reversal statistics are easy to calculate but can be difficult to interpret.

Many scholars have advocated using reversal rates as indicators of judicial quality,[112] and some state courts use reversal rates in judicial performance evaluations.[113] Reversal rates are also commonly used to evaluate circuits, with one study even assigning letter grades to the various circuits based on how often they are reversed by the Supreme Court.[114] In debates about splitting the Ninth Circuit, scholars and judges have often discussed the Ninth Circuit's high reversal rate and debated its normative significance.[115]

---

111 Westlaw search for "reversal rate" or "affirmance rate" in JLR database from Jan. 1, 2002 until December 31, 2011.

112 *See* Cross & Lindquist, *supra* note 72, at 1402-05 (defending reversal rates as performance indicators for circuit judges); Floyd Feeney, *Evaluating Trial Court Performance*, 12 JUST. SYS. J. 148, 151-53 (1987) (describing reversal rates as an accepted indicator of trial court performance); Rebecca D. Gill et al., *Are Judicial Performance Evaluations Fair to Women and Minorities? A Cautionary Tale from Clark County, Nevada*, 45 L. & SOC'Y REV. 731, 751 (2011) (characterizing reversal rates as being "among the most relevant and quantifiable objective measures we have" of judicial performance).

113 *See* David C. Brody, *Judicial Performance Evaluations by State Governments: Informing the Public While Avoiding the Pitfalls*, 21 JUST. SYS. J. 333, 340 (2000) (describing use of reversal rates by the Alaska Judicial Council).

114 Roy E. Hofer, *Supreme Court Reversal Rates: Evaluating the Federal Courts of Appeals*, LANDSLIDE, Jan.-Feb. 2010, at 10.

115 *See* Jerome Farris, *The Ninth Circuit—Most Maligned Circuit in the Country—Fact or Fiction?*, 58 OHIO ST. L.J. 1465, 1465-70 (1997) (minimizing the significance of the Ninth Circuit's

Judges themselves have considered reversal rates in trial courts and administrative proceedings in determining whether procedures were adequate under the due process clause.[116] A growing literature in patent law has examined how often the Federal Circuit reverses claim construction decisions by district judges and debated the implications of the reversal rate.[117] One recent study evaluated economic training programs for district

reversal rate because most cases are not reviewed by the Supreme Court and many reversals occurred in cases in which "reasonable minds can differ"); Arthur D. Hellman, *Getting It Right: Panel Error and the En Banc Process in the Ninth Circuit Court of Appeals*, 34 U.C. DAVIS L. REV. 425, 432 (2000) (describing a circuit's "disproportionately high reversal rate" in the Supreme Court as a "matter of concern"); Marybeth Herald, *Reversed, Vacated, and Split: The Supreme Court, the Ninth Circuit, and the Congress*, 77 OR. L. REV. 405, 489 (1998) (denying that the Ninth Circuit's "high reversal rate is a problem that needs to be solved by a circuit split" because it is attributable to "ideological disagreement"); Richard A. Posner, *Is the Ninth Circuit Too Large? A Statistical Study of Judicial Quality*, 29 J. LEGAL STUD. 711, 712-13 (2000) (arguing that the Ninth Circuit's reversal rate is a "meaningless" statistic because reversals "often involve disagreement rather than the correction of error," but claiming that rates of summary reversal have normative significance); Kevin M. Scott, *Supreme Court Reversals of the Ninth Circuit*, 48 ARIZ. L. REV. 341, 342 (2006) (describing the Ninth Circuit's reversal rate as "cause for concern").

116 *See, e.g.*, Jones v. Barnes, 463 U.S. 745, 756 n.1 (1983) (Brennan, J., dissenting) ("[T]he reversal rate of criminal convictions on mandatory appeals in the state courts, while not overwhelming, is certainly high enough to suggest that depriving defendants of their right to appeal would expose them to an unacceptable risk of erroneous conviction."); Mathews v. Eldridge, 424 U.S. 319, 346-47 (1976) (viewing reversal rates as "relevant" but "not controlling" in assessing the adequacy of administrative procedures, especially when new evidence can be presented on appeal); Veterans for Common Sense v. Shinseki, 644 F.3d 845, 885 (9th Cir. 2011), *rev'd*, 678 F.3d 1013 (9th Cir. 2012) (holding that Department of Veterans Affairs policy for adjudicating claims for mental health benefits "has not worked, given the high reversal rates of those determinations").

117 *See, e.g.*, Christian A. Chu, *Empirical Analysis of the Federal Circuit's Claim Construction Trends*, 16 BERKELEY TECH. L.J. 1075, 1143 (2001) (finding that the Federal Circuit has reversed fewer cases overall, but has increased its claim construction modification and claim interpretation-based reversal rates); Richard S. Gruner, *How High is Too High?: Reflections on the Sources and Meaning of Claim Construction Reversal Rates at the Federal Circuit*, 43 LOY. L.A. L. REV. 981, 1071 (2010) (arguing that concerns about claim construction reversal rates are misplaced because most appeals involve hard cases); Kimberly A. Moore, *Are District Court Judges Equipped to Resolve Patent Cases?*, 15 HARV. J.L. & TECH. 1, 2-3 (2001) (arguing that the claim construction reversal rate is "problematic" because "it raises concerns about the efficiency of [the] adjudication system" and "creates doubt about the abilities of district court judges to adjudicate complex technical patent cases"); Kimberly A. Moore, Markman *Eight Years Later: Is Claim Construction More Predictable?*, 9 LEWIS & CLARK L. REV. 231, 245-47 (2005) (finding an increase in claim construction reversal rates and concluding that "[t]he fault . . . undoubtedly lies with the Federal Circuit" because it "is not providing sufficient guidance on claim construction"); David L. Schwartz, *Practice Makes Perfect? An Empirical Study of Claim Construction Reversal Rates in Patent Cases*, 107 MICH. L. REV. 223, 245-57 (2008) (examining whether claim construction reversal rates improve as a function of judicial experience); David L. Schwartz, *Pre-*Markman *Reversal Rates*, 43 LOY. L.A. L. REV. 1073, 1091-107 (2010) (examining the impact of *Markman v. Westview Instruments* on reversal rates, but urging caution in interpreting the results); Ted Sichelman, *Myths of (Un)certainty at the Federal Circuit*, 43 LOY. L.A. L. REV. 1161, 1171-84 (2010) (arguing that reversal rates in claim construction cases are problematic).

court judges by measuring how often their decisions in antitrust cases were appealed and reversed.[118]

Reversal rates have been prominently featured in debates about reforming asylum adjudication. In 2002, then-Attorney General John Ashcroft adopted "streamlining" rules for the Board of Immigration Appeals (BIA), which permitted decisions by immigration judges to be affirmed by a single BIA member in an unsigned opinion.[119] The rate at which the BIA reversed immigration judge decisions plummeted, leading some commentators to criticize the streamlining reforms for allowing errors to go uncorrected.[120] Ashcroft contended that these reversal rates had no significance[121] but then went on to claim that "the BIA streamlining reforms were a profound success" because fewer than ten percent of BIA decisions were reversed by circuit courts.[122] Yet in a widely noted opinion, Judge Richard Posner cited the BIA's reversal rate in the Seventh Circuit as evidence that immigration adjudication had "fallen below the minimum standards of legal justice."[123]

Legal scholars seem to think that reversal rates are worth discussing, but they rarely articulate why these rates are purportedly meaningful. Often, reversal rates are conflated with error rates or imbued with unwarranted normativity. One study, for example, found that more than two-thirds of death sentences in state courts are ultimately overturned on appeal.[124] After performing a highly sophisticated statistical analysis to examine what

---

118 *See* Michael R. Baye & Joshua D. Wright, *Is Antitrust Too Complicated for Generalist Judges? The Impact of Economic Complexity and Judicial Training on Appeals*, 54 J.L. & ECON. 1, 1-24 (2011) (finding that antitrust decisions by district judges and administrative law judges are more likely to be appealed and reversed in complex cases, but less likely to be appealed and reversed when the judge had participated in a program that provided basic economic training).

119 8 C.F.R. § 1003.1 (2010). The streamlining regulations also required the Board to review factual findings under the "clear error" standard, rather than de novo.

120 *See* DORSEY & WHITNEY LLP, STUDY CONDUCTED FOR THE AMERICAN BAR ASSOCIATION COMMISSION ON IMMIGRATION POLICY, PRACTICE AND PRO BONO RE: BOARD OF IMMIGRATION APPEALS: PROCEDURAL REFORMS TO IMPROVE CASE MANAGEMENT 40-41 (2003) ("The federal courts are describing obvious errors committed by the BIA: errors that would be comic, if they were not so tragic."); ANNA O. LAW, THE IMMIGRATION BATTLE IN THE AMERICAN COURTS 150 (2010) ("Exacerbating the worry that individual Board member adjudications would lead to more errors than a three-member panel was the increase in summary affirmances of immigration judges' decisions that went against the alien.").

121 *See* John D. Ashcroft & Kris W. Kobach, *A More Perfect System: The 2002 Reforms of the Board of Immigration Appeals*, 58 DUKE L.J. 1991, 2008 (2009) (agreeing with the Fourth Circuit's observation that reversal statistics are meaningless unless there is "an objectively correct percentage of reversals" to which an adjudicative body should aspire).

122 *Id.* at 2009.

123 Benslimane v. Gonzales, 430 F.3d 829, 830 (7th Cir. 2005).

124 *See* Andrew Gelman et al., *A Broken System: The Persistent Patterns of Reversals of Death Sentences in the United States*, 1 J. EMPIRICAL LEGAL STUD. 209, 214 (2004).

factors predicted reversal, the authors considered policy options to reduce reversal rates. First, they proposed rules requiring disclosure of exculpatory evidence and more funding for defense lawyers at the trial stage.[125] But then they noted that reversal rates could also be reduced by limiting the grounds for reversal and withdrawing funding for attorneys who represent death-row inmates at the post-conviction stage.[126] The authors did not actually advocate the latter proposals, acknowledging that "[t]he positive impact of such policies is questionable."[127] The fact that they simultaneously considered increasing funding for trial lawyers and defunding appellate lawyers, however, suggests that they were asking the wrong question. By conflating reversals with errors,[128] the authors lost sight of their normative goals. Defunding appellate lawyers may well reduce reversals, but this should serve as a reminder that the reduction of reversal rates is not a worthy end in itself.

As with citation counts, reversal rates do not have any intrinsic normative significance; they are only useful insofar as they can shed light on other normatively significant quantities, such as error rates. A reversal is a good outcome when the lower court was wrong, but it is a bad outcome when the lower court was correct. If the applicable law is indeterminate, a reversal reflects the fact that the higher and the lower courts are exercising discretion differently. Reversal rates, however, aggregate "good reversals" and "bad reversals," as well as "ambiguous reversals" when the law is indeterminate.

Although reversal rates are commonly used to measure error rates of lower courts, they accurately reflect error only when four conditions are satisfied: the law is always determinate, both courts are addressing the same legal question and relying on the same sources of law, all cases are appealed, and the higher court is always correct. Scholars can debate whether and when the first two conditions hold,[129] but the third is rarely satisfied and the fourth is almost always implausible. Thus, additional assumptions are necessary to draw normative conclusions from reversal rates.

---

125 *Id.* at 255.

126 *Id.* at 256.

127 *Id.* at 257.

128 *See id.* at 216-17 (defining the "total error rate" in terms of probabilities of reversal); *id.* at 218 ("We counted only *error* that actually resulted in *reversal* by the highest court with authority to review the verdict at the relevant stage of review." (emphasis added)).

129 Ronald Dworkin, most prominently, has adhered to the view that the law is determinate even in hard cases. *See* RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 81-130 (1977). But this viewpoint, while influential, is not widely held. Whether courts are addressing the same legal question will necessarily depend upon context.

The proportion of cases that are appealed is especially relevant when the reviewing court is the U.S. Supreme Court, which hears only a tiny fraction of petitioned cases. For example, Judge Jerome Farris observed that the Supreme Court reversed the Ninth Circuit in 28 out of 29 cases it reviewed in 1997.[130] Yet he defended the Ninth Circuit by arguing that the Court let stand more than 99% of all Ninth Circuit decisions from the previous year.[131]

A further complication is that reviewing courts do not necessarily consider the same legal issues as lower courts. When lower court decisions are reviewed under a deferential standard, a reversal might be stronger evidence of error, or at least strong disagreement. Failure to reverse, however, does not show that the higher court believed that the lower court judgment was correct.

In addition, higher courts and lower courts are often bound by different sources of law, even when resolving the same dispute. A circuit court panel may reach a result that is compelled by circuit precedent, but the Supreme Court would not be bound by the same circuit precedent. The Supreme Court also has the authority to overrule its own precedent, whereas a circuit court is obligated to follow such precedent until it is overruled by the Supreme Court.[132] Thus, reversal by the Supreme Court may well represent the application of different legal principles rather than disagreement about the same legal principles. In other words, the Supreme Court can overrule a circuit court, and both can still be correct.

Consider Judge Richard Posner's opinion in *Khan v. State Oil Co.*,[133] an antitrust case involving maximum resale price maintenance. Judge Posner believed the outcome was controlled by the Supreme Court's holding in *Albrecht v. Herald Co.*[134] Posner criticized *Albrecht* at length, describing it as "unsound when decided, and . . . inconsistent with later decisions by the Supreme Court."[135] He continued: "It should be overruled. Someday, we expect, it will be."[136] In a not-so-subtle signal to the Supreme Court, Posner

---

130 *See* Farris, *supra* note 115, at 1465.

131 *See id.*

132 *See* Rodrigues de Quijas v. Shearson/Am. Express, Inc., 490 U.S. 477, 484 (1989) ("[T]he Court of Appeals should follow the case which directly controls, leaving to this Court the prerogative of overruling its own decisions.").

133 93 F.3d 1358 (7th Cir. 1996), *vacated*, 522 U.S. 3 (1997).

134 390 U.S. 145 (1968), *overruled by* State Oil Co. v. Khan, 522 U.S. 3 (1997).

135 *Khan*, 93 F.3d at 1363.

136 *Id.*

wrote, "Yet despite all its infirmities, its increasingly wobbly, moth-eaten foundations, *Albrecht* has not been *expressly* overruled."[137]

Presumably, Judge Posner was not disappointed when the Supreme Court reversed him unanimously and overruled *Albrecht*,[138] relying extensively on his reasoning in the Seventh Circuit decision.[139] In this example, it would certainly be reasonable to assert that the Supreme Court was correct to overrule *Albrecht*, but that Posner was also correct to follow *Albrecht* despite his disagreement with its holding. From this point of view, the reversal does not reflect poorly on Posner; it resulted from the fact that the Seventh Circuit and the Supreme Court were bound by different sources of law. To the contrary, this reversal demonstrates Posner's influence, since he was able to convince the Court to hear the case and overrule a longstanding precedent that he disfavored.

To support any kind of conclusion about error rates, reversal rates must be interpreted in conjunction with some kind of assumptions about the relative competence of higher and lower courts and the determinacy of the law in the cases being analyzed.[140] In debates about the performance of the Ninth Circuit, for example, Judge Diarmuid O'Scannlain has cited the Ninth Circuit's high reversal rate in the Supreme Court as evidence that the Ninth Circuit "got it wrong" in a large majority of the cases that were reviewed.[141] Arthur Hellman has argued that, irrespective of whether the ultimate outcome is correct, "it is not healthy when an intermediate court is reversed repeatedly by the highest court in the structure."[142] But others have argued that the reversal rate reflects positively on the Ninth Circuit. According to Michelle Landis Dauber, the problem was "not that the 9th Circuit [was] getting the law wrong" but rather that "the Rehnquist Court [was] changing the law."[143] Judge Stephen Reinhardt, the most frequently reversed circuit judge in the federal courts,[144] is said to view his reversal rate as a "mark of distinction."[145] Judge Richard Posner, on the other hand,

---

137 *Id.* (internal citation omitted).

138 *See* State Oil Co. v. Khan, 522 U.S. 3 (1997).

139 *See id.* at 15-16, 20.

140 A high degree of disagreement between higher and lower courts may still provide evidence of substantial confusion about the meaning of the law and inefficiency in the resolution of disputes. *See* sources cited *supra* note 117 (discussing high reversal rates in patent cases).

141 Diarmuid F. O'Scannlain, *A Decade of Reversal: The Ninth Circuit's Record in the Supreme Court Since October Term 2000*, 14 LEWIS & CLARK L. REV. 1557, 1558 (2010).

142 Hellman, *supra* note 115.

143 Michele Landis Dauber, *The 9th Circuit Follows*, LEGAL TIMES, Aug. 19, 2002, at 37.

144 *See* Cross & Lindquist, *supra* note 72, at 1407 tbl.1 (ranking circuit court judges by frequency of reversal).

145 Heather K. Gerken, *Judge Stories*, 120 YALE L.J. 529, 530 (2010).

argues that reversal rates are meaningless statistics because "reversals by the Supreme Court often involve disagreement rather than the correction of error, and . . . the Supreme Court has neither the capacity nor the incentive to review more than a tiny percentage of federal courts of appeals decisions."[146]

The above commentators agree about what the Ninth Circuit's reversal rate is, but they have sharply differing views about its normative implications. Reversal rates may be objective, but they must be interpreted in conjunction with contestable assumptions about the relative competence of higher and lower courts, the institutional obligations of the lower courts, and the determinacy of the law in the cases under examination. Judge O'Scannlain's conclusions appear to be posited on a belief that the Supreme Court is usually correct when it disagrees with the Ninth Circuit; Judge Reinhardt's and Dauber's viewpoints are premised upon a more negative view of the Supreme Court. Hellman's position is predicated on a view that inferior courts should try to predict how the Supreme Court will rule, but Dauber disagrees, arguing that "the job of an intermediate court does not entail . . . trying to divine what the current members of the Supreme Court might do if and when they get the case."[147] Posner's view, on the other hand, reflects his view of the Court as a "political body"[148] rather than as a tribunal resolving legally determinate disputes.

As these conflicting interpretations demonstrate, scholars must be explicit about the premises that underlie their normative conclusions. These premises, moreover, must be plausible. Simple but implausible assumptions such as "the higher court is always correct" may support straightforward interpretations of reversal rates, but such conclusions have little value. What is needed are methods for combining objective data on reversals with plausible assumptions to generate useful conclusions that can inform policymaking.

A study of jury verdicts by Bruce Spencer provides an instructive example. Spencer examined disagreement between juries and judges in trial

---

146 Posner, *supra* note 115, at 712. Posner, however, argues that rates of summary reversal provide a useful indicator of circuit quality. *See id.* at 713.

147 Dauber, *supra* note 143, at 36. For sophisticated discussions about whether inferior courts ought to predict how a higher court would rule, see Evan Caminker, *Precedent and Prediction: The Forward-Looking Aspects of Inferior Court Decisionmaking*, 73 TEX. L. REV. 1 (1994), which argues that inferior courts may properly anticipate higher court rulings, and Michael C. Dorf, *Prediction and the Rule of Law*, 42 UCLA L. REV. 651, 673 (1995), which argues that the prediction approach is inconsistent with the rule of law.

148 *See* Richard A. Posner, *Foreword: A Political Court*, 119 HARV. L. REV. 32, 34 (2005) (observing that, on most constitutional issues, the Supreme Court behaves like a political body).

courts, but the same approach applies to disagreement between higher and lower courts within the judicial hierarchy. Using data in which trial judges had been surveyed about the correct outcome, Spencer estimated the accuracy of jury verdicts under the assumption that the judge is *at least as* likely to be correct as the jury.[149] Of course, he could have considered alternative assumptions as well. Stronger assumptions—such as that the judge is twice as likely as the jury to be correct—would have yielded sharper inferences. Similarly, weaker assumptions—such as that the judge is correct at least 10% of the time—would have yielded weaker inferences. By interpreting the data according to a variety of assumptions, empirical scholars can make their findings interpretable to an audience with a diverse range of viewpoints.

IV. MEASURING THE RULE OF LAW: STUDIES OF INTERJUDGE DISPARITY

A central feature of the rule of law is that the application of legal force is governed by publicized rules rather than "the predilections of the individual decisionmaker."[150] A large body of empirical research has sought to measure the degree to which systems of adjudication deviate from this ideal. Such studies have typically documented statistical disparities among judges—differences in their rates of reaching various types of dispositions—and concluded that the rule of law is violated. Such claims are typically followed by calls for legal or institutional reform.

The earliest example of such a study may be an annual report published by the criminal magistrates of New York City in 1914,[151] which provided detailed figures depicting the magistrates' conviction rates for various offenses. The magistrates reported large interjudge disparities in cases involving public intoxication, vagrancy, disorderly conduct, and peddling without a license, but more modest disparities in cases involving cruelty to animals and violations of the motor vehicle laws.[152]

---

149 *See* Bruce D. Spencer, *Estimating the Accuracy of Jury Verdicts*, 4 J. EMPIRICAL LEGAL STUD. 305, 310-14 (2007) (estimating the accuracy of jury verdicts from data on judge–jury agreement).

150 RONALD A. CASS, THE RULE OF LAW IN AMERICA 17 (2001).

151 *See* NEW YORK BOARD OF CITY MAGISTRATES, ANNUAL REPORT OF THE CITY MAGISTRATES' COURTS OF THE CITY OF NEW YORK (FIRST DIVISION) FOR YEAR ENDING DECEMBER 31, 1914 (1914) (compiling judicial outcomes with the intent of dissemination to the general public).

152 *See id.* at 50-61 (providing numerical and graphical representations of magistrates' decisions in various categories of cases).

The magistrates were not trying to advance any grand theories about law or adjudication; they merely hoped that publication of the statistics would help the magistrates "recognize [their] own personal peculiarities" and "correct any that cannot be justified in light of the records of [their] associates."[153] But their reports provided inspiration to legal realists such as Jerome Frank[154] and to political scientists such as Charles Grove Haines,[155] who viewed the results as confirmation that adjudication was inevitably idiosyncratic.

Since the publication of the magistrates' report in 1914, numerous studies have documented significant interjudge disparities in cases involving criminal law,[156] social security disability claims,[157] and asylum adjudication.[158] The original magistrates' report had modest normative goals, but many of these later studies advocated bold reforms. Disparity studies provided much of the impetus for the enactment of the U.S. Sentencing Guidelines[159] and the disability grid for social security disability claims.[160] More recently, scholars have been debating proposed reforms to address disparities in asylum adjudication.[161] Yet despite the large number of disparity studies that have been conducted and the prominence of the policy claims that have been advanced, there has been surprisingly little discussion about how observable disparities relate to normatively significant concepts.

---

153 George Everson, *The Human Element in Justice*, 10 J. CRIM. L. & CRIMINOLOGY 90, 98 (1919).

154 *See* JEROME FRANK, LAW AND THE MODERN MIND 124 (1930).

155 *See* Charles Grove Haines, *General Observations on the Effects of Personal, Political, and Economic Influences in the Decisions of Judges*, 177 ILL. L. REV. 96 (1922).

156 *See, e.g.*, WAYNE L. MORSE & RONALD H. BEATTIE, THE ADMINISTRATION OF CRIMINAL JUSTICE IN OREGON 151-69 (1932); James M. Anderson et al., *Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines*, 42 J.L. & ECON. 271 (1999); Emil Frankel, *The Offender and the Court: A Statistical Analysis of the Sentencing of Delinquents*, 31 AM. INST. CRIM. L. & CRIMINOLOGY 448 (1940); Frederick J. Gaudet et al., *Individual Differences in the Sentencing Tendencies of Judges*, 23 J. CRIM. L. & CRIMINOLOGY 811 (1933); Paul J. Hofer et al., *The Effect of the Federal Sentencing Guidelines on Inter-Judge Sentencing Disparity*, 99 J. CRIM. L. & CRIMINOLOGY 239 (1999); A. Abigail Payne, *Does Inter-Judge Disparity Really Matter? An Analysis of the Effects of Sentencing Reforms in Three Federal District Courts*, 17 INT'L REV. L. & ECON. 337 (1997); Ryan W. Scott, *Inter-Judge Sentencing Disparity After* Booker*: A First Look*, 63 STAN. L. REV. 1 (2010); Whitney North Seymour, Jr., *1972 Sentencing Study for the Southern District of New York*, 45 N.Y. ST. B.J. 163 (1973); Joel Waldfogel, *Does Inter-Judge Disparity Justify Empirically Based Sentencing Guidelines?*, 18 INT'L REV. L. & ECON. 293 (1998).

157 *See* JERRY L. MASHAW ET AL., SOCIAL SECURITY HEARINGS AND APPEALS: A STUDY OF THE SOCIAL SECURITY ADMINISTRATION HEARING SYSTEM 21 (1978).

158 *See* Ramji-Nogales et al., *supra* note 27.

159 *See* STITH & CABRANES, *supra* note 25, at 104-142.

160 *See generally* MASHAW, *supra* note 26.

161 *See* Stephen H. Legomsky, *Learning to Live with Unequal Justice: Asylum and the Limits to Consistency*, 60 STAN. L. REV. 413 (2007); Ramji-Nogales et al., *supra* note 27, at 378-89.

A. *The Normative Implications of Disparity*

In the century since the magistrates released their annual report, the methodology of disparity studies has barely changed. The studies count judges' decisions and report rates at which they reach various types of outcomes. Whenever disparities are found, the authors conclude that some reform is needed. Yet only a few of these studies have acknowledged that these statistical disparities by themselves do not have intrinsic normative significance. As one study of social security disability adjudication observed:

> Two judges with different [rates of reversing disability determinations] are probably behaving differently. But the reverse is not necessarily true: there is no reason to exclude the possibility that two judges with 50 percent [rates] are also behaving differently. Indeed, the likelihood is great that the existing statistics mask an indeterminate additional amount of nonuniformity in the judge-to-judge handling of [social security] claims.[162]

Thus, although large disparities among judges are problematic, small disparities do not necessarily indicate that a system of adjudication is functioning well. If two social security judges were deciding cases using coin flips, there would be no disparity, since both would reverse agency determinations 50% of the time. This means that any existing disparities in grant rates could be eliminated by ordering all judges to flip coins. The absurdity of such a proposal demonstrates that eliminating statistical disparity is not itself a worthy goal. Statistical disparity is only of interest insofar as it can shed light on other values.

To understand the normative implications of these studies, it is necessary to articulate the values at stake and to explain how they relate to the measureable statistics. Some of the prior scholarship has made efforts to identify the relevant values, such as consistency, correctness, determinacy, fairness, predictability, non-arbitrariness, and the rule of law.[163] But there has been little effort to explain how these values can be measured using available data. In fact, the relationships between these values and measureable statistics can be quite complex.

---

162 MASHAW ET AL., *supra* note 157, at 22.

163 *See id.* at 13-27 (consistency and correctness); Ramji-Nogales et al., *supra* note 27, at 299-300 (predictability, fairness, and rule of law); Jeremy Waldron, *Lucky in Your Judge*, 9 THEORETICAL INQUIRIES L. 185, 190-92 (2007) (predictability, non-arbitrariness, and fairness).

B. *Consistency, Predictability, and Comparative Justice*

Statistical disparities are of interest in part because they provide evidence of interjudge *inconsistency*—meaning that some cases would have been decided differently if they had been assigned to different judges. Indeed, many discussions of interjudge disparity focus on inconsistency as a normative concept.[164] Inconsistency has normative significance for two distinct reasons. The first is that it diminishes the predictability of adjudication. The rule of law requires that people have notice regarding how the law will be applied so that they can conform to its requirements and plan their affairs accordingly.[165] Notice will necessarily be inadequate to the extent that the application of the law depends upon which judge is deciding each case.[166]

Inconsistency among judges also implicates comparative justice.[167] Some legal rights may be comparative, in the sense that "a person's due is determinable only by reference to his relations to other persons."[168] In the sentencing context, for example, moral or legal principles may determine that two offenders are equally culpable and should therefore receive the same sentence, even if those principles do not uniquely determine what that sentence should be. If two such offenders receive different sentences only

---

164 *See, e.g.*, Rules for Adjudicating Disability Claims in Which Vocational Factors Must be Considered, 43 Fed. Reg. 55,349, 55,351 (Nov. 28, 1978) (justifying the grid rule for social security disability determinations on the ground that it would "better assure consistency of determinations"); MARVIN E. FRANKEL, CRIMINAL SENTENCES: LAW WITHOUT ORDER 7 (1973) (criticizing the federal sentencing process for failing to provide "any semblance of the consistency demanded by the ideal of equal justice"); Ramji-Nogales et al., *supra* note 27, at 299 ("Americans don't love consistent decisionmaking merely because we think that fairness to the parties requires that similar cases should have similar outcomes. We also like the predictability that stare decisis offers.").

165 *See* LON L. FULLER, THE MORALITY OF LAW 39-40 (1964) ("Government says to the citizen in effect, 'These are the rules we expect you to follow. If you follow them, you have our assurance that they are the rules that will be applied to your conduct.'"); Jules L. Coleman & Brian Leiter, *Determinacy, Objectivity, and Authority*, 142 U. PA. L. REV. 549, 582 (1993) (noting that predictability provides agents with "the opportunity to conform their behavior to law's demands"); Waldron, *supra* note 163, at 191 ("An important element of most theories of the Rule of Law is that those who make and administer state policy should do what they can to diminish its unpredictability and provide a solid and reliable basis for calculation by ordinary citizens.").

166 Consistency is a necessary but not a sufficient condition for predictability; even if all judges would decide a case the same way, that outcome might not be predictable. However, one might expect that a knowledgeable observer would be able to predict such an outcome with a reasonable degree of accuracy. *See* Coleman & Leiter, *supra* note 165, at 584-85 (discussing how "lawyers can and do predict, with a fairly high degree of accuracy, what outcomes judges will reach" by relying on a "'folk' social scientific theory of adjudication").

167 *See* Waldron, *supra* note 163, at 191-92 (discussing how inconsistent treatment of litigants implicates comparative justice).

168 Joel Feinberg, *Noncomparative Justice*, 83 PHIL. REV. 297, 298 (1974) (emphasis omitted).

because they were sentenced by different judges, such a result would constitute a violation of comparative justice.

Interjudge inconsistency, however, only captures one aspect of comparative justice. If two judges fail to treat like cases alike in precisely the same way—perhaps by exhibiting the same degree of racial bias—then they could be perfectly consistent with each other yet still violate comparative justice. Nevertheless, inconsistency provides some evidence of comparative injustice when the cases under examination present common legal or factual patterns.

Interjudge inconsistency appears to have an intuitive relationship with observable data. If two social security judges are granting benefits to claimants at very different rates, then they probably are treating the claimants inconsistently. Yet the relationship between measurable statistical disparity and inconsistency is far more complex than the disparity studies have acknowledged.[169] The difference between the judges' grant rates only determines lower and upper bounds for inconsistency, but cannot identify the precise level. Suppose, for example, that Judge *A* grants benefits to 30% of claimants and Judge *B* grants benefits to 40% of claimants. If these two judges saw a comparable mix of cases, then it follows that they would have reached different results in at least 10% of the cases. There is no reason, however, to presume that they would have disagreed exactly 10% of the time. In fact, they could have disagreed as much as 70% of the time if they would have granted benefits to entirely different sets of claimants.

Without more information, it is impossible to know whether the rate of inconsistency between Judges *A* and *B* is closer to 10% or 70%. Measuring inconsistency requires not only the judges' grant rates, but also the degree to which their decisions are correlated. There are no data that can provide estimates of correlation, however, because the two judges are never observed deciding the same case. Conceivably, one could administer surveys to Judge *A* and Judge *B* and compare their reactions to identical fact patterns, which could then be used to compute the correlation between their decisions. A few studies did administer such surveys in the 1970s and early 1980s,[170] but to my knowledge, no recent disparity study has sought to measure the correlation of judges' decisions using surveys.

---

169 *See* Joshua B. Fischman, *Measuring Inconsistency, Indeterminacy, and Error in Adjudication*, AM. L. & ECON. REV. (forthcoming Spring 2014) (manuscript at 6-20), *available at* http://aler.oxfordjournals.org/cgi/content/abstract/aht011?ijkey=f68R7vaaKTMP3yo&keytype=ref (constructing bounds on interjudge inconsistency from judges' rates of reaching different outcomes).

170 *See* ANTHONY PARTRIDGE & WILLIAM B. ELDRIDGE, THE SECOND CIRCUIT SENTENCING STUDY: A REPORT TO THE JUDGES OF THE SECOND CIRCUIT (1974), *available at*

This simple example involved only two judges and assumed that the judges' grant rates were known exactly. When there are more than two judges, the relationship between grant rates and inconsistency becomes far more complex.[171] Complex statistical problems arise when judges' grant rates are not known precisely, but must be inferred from judges' decisions in actual cases.[172]

C. *Determinacy and Correctness*

The concept of interjudge consistency was defined without reference to the content of law or any substantive conception of justice. This makes inconsistency easier to conceptualize but also limits its utility as a normative metric. Inconsistency may be most important in settings where predictability is paramount and correctness is a secondary concern. As Justice Brandeis wrote, it is sometimes "more important that the applicable rule of law be settled than that it be settled right."[173] But in many settings, assessing whether a system of adjudication satisfies the requirements of law and justice may be more important than whether it provides consistent results.[174]

Any attempt to measure whether decisions are correct or just will typically require addressing concepts that are not objectively measureable, at least whenever the meaning of law or the requirements of justice are disputed. Nevertheless, it is possible to make limited objective claims about correctness on the basis of empirical data. Consider, for example, one finding by Thomas Miles and Cass Sunstein in a study of circuit court cases reviewing administrative agency decisions for arbitrariness: judges who are

---

http://www.fjc.gov/public/pdf.nsf/lookup/2dcrstdy.pdf/$file/2dcrstdy.pdf (analyzing the results from a survey of district judges on sentencing severity); Kevin Clancy et al., *Sentence Decisionmaking: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity*, 72 J. CRIM. L. & CRIMINOLOGY 524 (1981) (same); Shari Seidman Diamond & Hans Zeisel, *Sentencing Councils: A Study of Sentence Disparity and Its Reduction*, 43 U. CHI. L. REV. 109 (1975) (analyzing sentences issued in districts that used sentencing councils, in which judges shared preliminary sentence recommendations for the same offenders). Some more recent studies have examined whether lay respondents have similar rank ordering for criminal offenses. *See, e.g.*, Paul H. Robinson & Robert Kurzban, *Concordance and Conflict in Intuitions of Justice*, 91 MINN. L. REV. 1829 (2007) (finding that lay respondents largely agree about the relative seriousness of various criminal offenses).

171 *See* Fischman, *supra* note 169, at 14-15 (defining and stating bounds for inconsistency with more than two judges).

172 *See id.* at 30-32 (deriving a method for making statistical inferences on inconsistency with observational data).

173 Burnet v. Coronado Oil & Gas Co., 285 U.S. 393, 406 (1932) (Brandeis, J., dissenting).

174 *See* RONALD DWORKIN, LAW'S EMPIRE 181 (1986) ("Suppose we can rescue only some prisoners of tyranny; justice hardly requires rescuing none even when only luck, not any principle, will decide whom we save and whom we leave to torture.").

Republican appointees are 14% more likely than Democratic appointees to vote to invalidate liberal agency decisions.[175] Although this empirical result cannot tell us how many of these agency decisions ought to have been invalidated, Miles and Sunstein argue that the disparity itself provides evidence of legal *error*: "We cannot rule out the possibility that one group has it essentially right. But it is not possible that both groups have it essentially right, and we suspect that errors can be found from both sides."[176]

This interpretation relies on the controversial premise that every case has a unique correct outcome. Under this assumption, if Democrat- and Republican-appointed judges would disagree in at least 14% of cases reviewing liberal agency decisions, then one side or the other must be wrong. This conclusion is justified even if we cannot know when Democrats and Republicans would disagree or which judges would be wrong. Since half of these cases, on average, would be decided by judges who are correct and half by judges who are wrong, we can expect that at least 7% of these cases will be wrongly decided.

Although Miles and Sunstein have situated studies such as theirs within a "New Legal Realism,"[177] there are important differences between their normative premises and those of the legal realists, who emphatically rejected the notion that the law was always determinate.[178] Miles and Sunstein offer an important interpretation, but the notion that judges should be evaluated on the basis of their adherence to "paper rules"[179] is not something that the legal realists would have endorsed.

Miles and Sunstein suggest an alternative interpretation of their results, which is more in line with the realist perspective: studies such as theirs "represent an effort to test certain intuitive ideas about the *indeterminacy* of law."[180] If judicial disagreement is taken as evidence that the law fails to provide a correct answer, then a 14% rate of disagreement between Democratic

---

175 *See* Thomas J. Miles & Cass R. Sunstein, *The Real World of Arbitrariness Review*, 75 U. CHI. L. REV. 761, 777 tbl.1 (2008) (comparing Democrat- and Republican-appointed circuit court judges' validation rates with respect to "liberal" agency decisions).

176 *Id.* at 807.

177 Thomas J. Miles & Cass R. Sunstein, *The New Legal Realism*, 75 U. CHI. L. REV. 831 (2008).

178 *See* Brian Leiter, *American Legal Realism* ("The Realists famously argued that the law was 'indeterminate.'"), *in* THE BLACKWELL GUIDE TO THE PHILOSOPHY OF LAW AND LEGAL THEORY 50, 51 (Martin P. Golding & William Edmundson eds., 2004).

179 *See* Llewellyn, *supra* note 4, at 448 (defining "[p]aper rules" as "the accepted *doctrine* of the time and place—what the books there say 'the law' is," in contrast to "real rules," which are "what the courts will do in a given case").

180 Miles & Sunstein, *supra* note 177, at 834 (emphasis added).

and Republican judges would mean that the outcome must be indeterminate in at least 14% of these cases.[181]

Claims about indeterminacy and error can be viewed as alternative interpretations of statistical disparity. If judges would disagree about how cases ought to be decided, then there must be some combination of legal indeterminacy and judicial error. Decomposing disparities into combinations of indeterminacy and error turns out to be a rather complicated statistical problem and still can only yield feasible combinations of indeterminacy and error rates.[182] As with inconsistency, the problem becomes more complicated with multiple judges and when grant rates must be inferred from judges' decisions.[183] Any effort to reach more precise interpretations of statistical disparity will necessarily require much stronger assumptions about the degree of legal determinacy and about the correct answers to various kinds of cases.[184]

D. *Conclusion*

Empirical studies on interjudge disparity have enormous potential for improving the quality of systems of adjudication. In order to justify reform, however, it is important for these studies to be precise about the values at stake and how they are measured. If one wants to promote consistency, then bureaucratic controls or even quotas would be an effective solution.[185] Selecting more capable judges or providing better training, as some have advocated,[186] might reduce error rates but could not reduce legal indeterminacy.

All too often, these studies have simply reported the disparities and let the audience reach judgments about the normative implications. The relationships between the data and the relevant normative concepts are far too complex, however, for intuition to be a reliable guide. Instead of

---

181 These disparities could be interpreted in terms of *epistemic indeterminacy*, meaning that the law is not knowable to competent judges. *See* Ken Kress, *A Preface to Epistemological Indeterminacy*, 85 NW. U. L. REV. 134, 138-39 (1990). They could also be interpreted as evidence of *causal indeterminacy*, meaning that the law fails to cause judges to reach the correct outcome. *See* Brian Leiter, *Legal Indeterminacy*, 1 LEGAL THEORY 481, 481-82 (1995).

182 *See* Fischman, *supra* note 169, at 21-27 (discussing how to determine feasible combinations of indeterminacy and error rates from observational data).

183 *See id.* at 21-27

184 *See id.* at 27-28 (discussing how assumptions can sharpen inferences about indeterminacy and error).

185 The fact that many scholars oppose such reforms in asylum adjudication suggests that consistency is not actually their primary concern. *See* Ramji-Nogales et al., *supra* note 27, at 379 (opposing bureaucratic controls and quotas).

186 *See id.* at 380-81.

"letting the data speak," we must start by focusing on the normative values at stake, and develop methods that can give useful answers about whether and how to reform legal institutions.

V. BRIDGING THE GAP BETWEEN 'IS' AND 'OUGHT'

Parts II–IV of this Article criticized a variety of empirical studies for failing to credibly reunite 'is' and 'ought.' Some studies sought to connect their findings to prescriptive claims but never explained how the phenomena measured related to any normative concepts. Other studies reported a variety of descriptive statistics without interpretation, placing the burden on unsophisticated audiences to decipher the implications of the findings.

It would be easy to fault the authors of these studies for claiming too much or for using flawed research designs. But the reality is that these studies are not aberrations. Many of them employed methods that are widely accepted in contemporary empirical legal studies, and several were even published in prestigious peer-reviewed journals.

The fundamental problem is that empirical legal methodology lacks frameworks for connecting empirical findings with normative conclusions. In this Part, I consider steps that scholars can take to make empirical research more relevant to the study of law. First, they should prioritize normative questions, and be explicit about the values that motivate their research. Second, they should allow substantive questions to drive their choice of methods, and not the other way around. Third, they need to be more explicit about how they are combining objective findings with contestable assumptions in order to reach normative conclusions. Finally, they should think more carefully about how empirical findings generalize from a research setting to a policy-relevant context.

A. *Prioritizing Normative Goals*

There is deep disagreement among legal scholars about whether and how empirical legal research should be guided by normative goals. These debates, of course, are not new. Llewellyn and Pound famously debated whether the 'is' could be divorced from the 'ought,' but Llewellyn's conception of a "temporary divorce" was controversial even among the legal realists. Felix Cohen agreed that empirical research should be guided by normative questions[187] but shared Pound's skepticism that 'is' and 'ought'

187 Felix S. Cohen, *Transcendental Nonsense and the Functional Approach*, 35 COLUM. L. REV. 809, 849 (1935) ("Legal description is blind without the guiding light of a theory of values.").

could ever be separated.[188] Others, such as Herman Oliphant and Underhill Moore, believed that empirical scholars should simply engage in a value-free description of the facts as they see them, without concern for any normative objectives.[189]

The value-free approach to legal research was not successful for Oliphant and Moore, whose empirical research was widely derided as pointless.[190] But their viewpoint is widely shared among contemporary empiricists.[191] There is much concern about keeping 'is' and 'ought' separate,[192] but far less emphasis on reuniting them. Many non-empiricists, however, have sharply criticized empirical legal scholarship for lacking relevance to normative questions in legal scholarship.[193]

This is, of course, a normative debate about the objectives of empirical legal scholarship. These commentators disagree about what constitutes good scholarship and how best to produce it. Resolving these debates, therefore, requires clarifying the goals of legal scholarship and considering how they can be advanced by empirical research.

---

188 *Id.* ("The relation between positive legal science and legal criticism is not a relation of temporal priority, but of mutual dependence." (citing Pound, *supra* note 6)).

189 *See* Underhill Moore, *Essay* ("[U]ntil [precise knowledge of the specific effects of law on behavior] is available, any discussion of the relative desirability of alternative social ends which may be achieved by law is largely day-dreaming."), *in* MY PHILOSOPHY OF LAW: CREDOS OF SIXTEEN AMERICAN SCHOLARS 203, 206-07 (Julius Rosenthal Found. for Gen. L., Nw. U. ed. 1987) (1941); Herman Oliphant, *Facts, Opinions, and Value-Judgments*, 10 TEX. L. REV. 127, 137 (1932) ("[I]t is no evidence that a student of law is deficient in moral sense if he merely observes and records the uniformities of social behavior with which the law is concerned . . . . It may, on the contrary, be substantial evidence of his desire to get on with what is his proper job at least, *viz.*, to identify rather than to evaluate the social consequences of particular legal measures and devices.").

190 *See supra* notes 11-16 and accompanying text.

191 *See* Theodore Eisenberg, *The Origins, Nature, and Promise of Empirical Legal Studies and a Response to Concerns*, 2011 U. ILL. L. REV. 1713, 1733 ("[S]tudies may be done largely because data are available, but no apology is needed for doing so."); Mark C. Suchman & Elizabeth Mertz, *Toward a New Legal Empiricism: Empirical Legal Studies and New Legal Realism*, 6 ANN. REV. LAW & SOC. SCI. 555, 574 (2010) ("[S]ome variants of the new legal empiricism often seem to be motivated less by systematic arguments about fundamental social processes than by casual curiosity, commonsense predictions, and readily available data.").

192 *See, e.g.*, Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 9 (2003) ("Too much legal scholarship ignores the rules of inference and applies instead the 'rules' of persuasion and advocacy.").

193 *See* Barry Friedman, *supra* note 20, at 262-63 ("Oftentimes positive scholarship seems to be struggling with the normative implications of its work only after the project is complete, if at all. One sees indications of a 'research now, justify later' approach. . . . Normative bite ought to define the problem, not be an afterthought."); *see also* C. Edwin Baker, *Viewpoint Diversity and Media Ownership*, 61 FED. COMM. L.J. 651, 663 (2009) ("Policy is misled if it relies on what is easy to measure rather than what is important.").

A comparison between law and other disciplines is instructive. The natural sciences, for example, are primarily descriptive. As Edward Rubin explains, "[T]he discourse of natural science . . . begins from the premise that there is a real world 'out there,' separate from conscious human control. Prescriptive statements about this world would be meaningless . . . ."[194] The role of normativity in the social sciences, however, remains controversial. As Rubin notes, some social scientists adopt the descriptive orientation of the natural sciences, proceeding as though their subject matter were "a fixed phenomenon, 'out there' beyond the control of rational decision-makers."[195] Many social scientists reject this approach, however, arguing that because the social sciences study institutions that serve social values, research questions in these fields must have some relevance to these values.[196]

Legal scholarship, like the social sciences, studies a social institution. An essential feature of law, however, is its normativity;[197] the law is developed

---

194 Edward L. Rubin, *Law And and the Methodology of Law*, 1997 WISC. L. REV. 521, 524.

195 *Id.* at 537. This detached perspective would accurately describe most quantitative research in political science. *See* IAN SHAPIRO, THE STATE OF DEMOCRATIC THEORY 2 (2003) ("Normative and explanatory theories of democracy grow out of literatures that proceed, for the most part, on separate tracks, largely uninformed by one another."); Robert A. Dahl, *The Behavioral Approach in Political Science: Epitaph for a Monument to a Successful Protest*, 55 AM. POL. SCI. REV. 763, 770-771 (1961) ("The empirical political scientist is concerned with what *is*, as he says, not with what *ought* to be."); John Gerring & Joshua Yesnowitz, *A Normative Turn in Political Science?*, 38 POLITY 101, 102 (2006) ("Traditionally, the scientific study of politics has been associated with a value-neutral approach to politics. One seeks to uncover what is, not what ought to be, in the political realm.").

196 *See, e.g.*, WILLIAM R. SHADISH ET AL., EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR GENERALIZED CAUSAL INFERENCE 476 (2002) ("Although scientists have frequently avoided value questions in the mistaken belief that they cannot be studied scientifically or that science is value free, we cannot avoid values even if we try. The conduct of experiments involves values at every step, from question selection through the interpretation and reporting of results."); MAX WEBER, THE METHODOLOGY OF THE SOCIAL SCIENCES 21 (Edward A. Shils & Henry A. Finch eds. & trans., 1949) ("The problems of the empirical disciplines are, of course, to be solved 'non-evaluatively.' . . . But the problems of the social sciences are selected by the value-relevance of the phenomena treated."); Robert A. Dahl, *Normative Theory, Empirical Research, and Democracy* ("Identifying a question that is important is a moral and normative issue, not a scientific issue."), *in* PASSION, CRAFT, AND METHOD IN COMPARATIVE POLITICS 113, 134 (Gerardo L. Munck & Richard Snyder eds., 2007); Robert Merton, *Technical and Moral Dimensions of Policy Research* ("The investigator may naively suppose that he is engaged in the value-free activity of research, whereas in fact he may simply have so defined his research problems that the results will be of use to one group in the society, and not to others. His very choice and definition of a problem reflects his tacit values."), *in* THE SOCIOLOGY OF SCIENCE: THEORETICAL AND EMPIRICAL INVESTIGATIONS 70 (1973); Gerring & Yesnowitz, *supra* note 195, at 112 ("*Art for art's sake* has some plausibility, and *science for science's sake* might also be argued in a serious vein. But no serious person would adopt as her thesis *social science for social science's sake*. Social science is science for *society's* sake.").

197 *See* Rubin, *supra* note 30 and accompanying text.

consciously by human decisionmakers for the purpose of guiding human conduct.[198] Law has no fixed reality that can be described without an understanding of its purposes.[199] Rather, the law is continually evolving, and one function of legal scholarship is to address how it ought to evolve.[200]

Even though empirical research is inherently descriptive, choices about what legal phenomena to examine and what relationships to analyze require evaluative judgments of importance.[201] Scholars have advanced various conceptions of importance, but all of these require some reference to values. First, empirical research could be considered important if it can guide legal reform, an assessment that necessarily requires a value judgment.[202] Second, empirical research may be important if it describes legal phenomena that participants in the legal system find important—which would require understanding their normative viewpoints.[203]

Finally, empirical research may be important if it contributes to the development of theories that can in turn guide legal reform or illuminate the

---

198 *See* Rubin, *supra* note 194, at 525 ("Modern legal scholars regard law as the product of conscious decision by public decision-makers, and possibly others.").

199 Although some classical formalists may have viewed law as being part of a fixed reality "beyond the reach of conscious decision-makers," virtually all contemporary legal scholars reject this perspective. *See id*. Even these formalists, of course, recognized the normativity of law. *See id.*

200 *See id*. ("There is thus no fixed reality, but rather an ongoing process by which people in certain positions make decisions; in bald terms, law is created, not discovered. This sentiment has led to prescriptive efforts to improve the quality of those decisions according to the scholar's own views about law or public policy.").

201 *See* JOHN FINNIS, NATURAL LAW AND NATURAL RIGHTS 17 (2d ed. 2011) ("[A] judgment of significance and importance must be made if [a description of law] is to be more than a vast rubbish heap of miscellaneous facts."); Cohen, *supra* note 187, at 848 ("The prospect of determining the consequences of a given rule of law appears to be an infinite task . . . unless we approach it with some discriminating criterion of what consequences are *important*."); Pound, *supra* note 6, at 697 ("[A] science of law must be something more than a descriptive inventory. There must be selection and ordering of the materials so as to make them intelligible and useful.").

202 *See* Cohen, *supra* note 187, at 848 ("[A] criterion of *importance* presupposes a criterion of values").

203 *See* ANDREI MARMOR, POSITIVE LAW AND OBJECTIVE VALUES 157 (2001) ("[A]n understanding of a normative social practice, like law . . . must comprise an understanding of its points, that is, of the values which would render the participants' beliefs in their reasons for action intelligible."); JOSEPH RAZ, THE AUTHORITY OF LAW 295 ("The explanation of human behavior related to law has to take account of the way people's beliefs about the law, normatively understood, affect their behavior"); Pound, *supra* note 6, at 700 ("Faithful portrayal of what courts and law makers and jurists do is not the whole task of a science of law. One of the conspicuous actualities of the legal order is the impossibility of divorcing what they do from the question what they ought to do or what they feel they ought to do."); Leslie Green, *Law and the Causes of Judicial Decisions* 33 (Oxford Legal Research Paper Series 2009), *available at* http://papers.ssrn.com/id=1374608 (endorsing Hans Kelsen's view that "if [the sociology of law] was to touch its intended subject, [it] would have to study beliefs and actions oriented towards legal norms as identified by jurisprudence").

legal system to its participants.[204] Certainly, such research need not have a direct normative payoff; a study may examine phenomena that seem narrow and unimportant, but the findings may generalize to a wide variety of contexts. Assessing what theories are useful, however, still requires a value-laden judgment; a theory that can only explain insignificant phenomena cannot itself be significant.

Empirical research may thus seek to advance an immediate policy prescription, to describe the legal system in a meaningful way, or to contribute to theory development. In any of these pursuits, however, the importance of the research must be assessed by reference to values. This is not to say that empiricists must personally take controversial positions in normative debates; one can acknowledge the viewpoints held by others without endorsing them. It is not too much to ask, however, that empirical research proceed in a conscious recognition of the values it intends to serve, and that scholars make efforts to clarify how their findings relate to the values that motivated their research.

B. *Rethinking Empirical Legal Methodology*

The studies criticized in Parts II–IV had worthy motivations, but they struggled to credibly connect their results to normative claims. As these discussions showed, the relationship between measureable objects and normative concepts is often complex. Because legal scholarship lacks its own empirical methodology, empiricists typically adhere to the methods of other disciplines, irrespective of whether they are suited to address the questions of legal scholarship.

Empirical legal methodology needs to be more closely tethered to the motivating questions in legal scholarship. Because the method used determines the question that is answered, the evaluation of methods and questions cannot be disentangled. Yet scholars have too often allowed the choice of method to determine the question that is asked, rather than having the substantive question determine the choice of method.

A 2002 critique of empirical legal scholarship by political scientists Lee Epstein and Gary King is often taken as representing the dominant view on empirical legal methodology. Declaring that the "state of empirical legal

---

204 David Collier et al., *Critiques, Responses, and Trade-offs: Drawing Together the Debate* ("[S]tudies that help advance theory in a way that gives scholars new leverage in conceptualizing and explaining significant outcomes would also be considered important."), *in* RETHINKING SOCIAL INQUIRY: DIVERSE TOOLS, SHARED STANDARDS 125, 127 (Henry E. Brady & David Collier eds., 2010).

scholarship [was] deeply flawed,"[205] they criticized legal scholars for their inattention to methodological concerns[206] and emphasized the need to develop empirical methods that were tailored to questions that arise in legal scholarship.[207]

Although Epstein and King were right about the need for an empirical legal methodology, they never explained why existing methodologies were inadequate. In fact, the bulk of their article was devoted to criticizing empirical legal scholars for failing to comply with standards that had been established *in other disciplines*. Of course, empirical methodologies need not be trapped within disciplinary boundaries, and many methods from other disciplines have proven useful in empirical legal studies.

Yet all too often, discussions of empirical legal methodology have been divorced from discussions of substantive questions. Even Epstein and King had little to say about the objectives of empirical legal scholarship.[208] In a brief section entitled "The Research Question," Epstein and King advocated a permissive approach toward research questions that sharply contrasted with their demanding rules for every other aspect of empirical scholarship.[209] They provided two criteria for research questions: "*they contribute to existing knowledge* and *they have some importance for the real world.*"[210] But the first criterion is nearly vacuous—how much research does not contribute in some way to existing knowledge?—and the meaning of "importance" in the second criterion was never articulated.[211] Epstein and King even declared these criteria to be entirely optional.[212] In their view, it is appropriate for "[i]nvestigators [to] conduct rigorous empirical research about any question, no matter how narrow it may be, no matter whether they are the only ones

---

205 Epstein & King, *supra* note 192, at 6 (emphasis omitted).

206 *See id.* at 11 ("[T]he complete list of all law review articles devoted to improving, understanding, explicating, or adapting the rules of inference is as follows: none.").

207 *See id.* at 11 ("The law is important enough to have a subfield devoted to methodological concerns, as does almost every other discipline that conducts empirical research. Scholars toiling in the social, natural, and physical sciences can help, but a whole field cannot count on others with differing goals and perspectives to solve all of the problems that law professors may face.").

208 *See* Jack Goldsmith & Adrian Vermeule, *Empirical Methodology and Legal Scholarship*, 69 U. CHI. L. REV. 153, 154 (2002) ("The reader of Epstein and King's 133-page article will find almost nothing that speaks to the simple question, 'What is legal scholarship *for*?'").

209 *See* Epstein & King, *supra* note 192, at 55-61.

210 *Id.* at 55.

211 With regard to the second criterion, they write, "This is a rule about which we need not say too much." *Id.* at 60.

212 *See id.* at 55 (writing that it "is not particularly problematic" that "many questions asked by academics and others about legal phenomena do not meet these standards").

interested in it, no matter if it has *virtually no implications for the real world*."[213]

Of course, Epstein and King were not actually advocating the pursuit of trivial questions. Nevertheless, by leaving the criteria of importance unexamined, they undermined their call for more rigorous methods in empirical legal research. Different methodological approaches will yield different estimates of causal or correlational relationships among observable variables. Any of these estimates can be rationalized as the answer to *some* research question, if perhaps an unimportant one. Evaluating empirical methods therefore requires some assessment of fit between statistical estimates that can be generated and *important* substantive research questions. This, in turn, requires criteria for determining what is an important question.

When standards for research questions are left unarticulated, it is all too tempting to allow the availability of data to define the research question.[214] Without some criterion of importance, one can start with a data set, apply a preferred statistical technique, and then rationalize a research question that is answered by the resulting estimate.[215] This may seem perverse, but it is in fact an inevitable consequence of a mindset that prioritizes adherence to methodological "rules" over normative relevance. If we combine exacting standards for deriving inferences from data with lax standards for relating those inferences to normative questions, there will be an inevitable pressure to reorient empirical research projects around phenomena that are conveniently measured and analyzed, rather than those that can genuinely inform policymaking. This leads to misplaced efforts at empirical sophistication, such as projects that explain the impact of judicial characteristics on citation counts using fixed-effects negative binomial regression with clustered

---

213 *Id.* at 55 (emphasis added). They do provide the caveat that "posing research questions in ways that attract the interest of others . . . is good career advice." *Id.*

214 *See* Friedman, *supra* note 20, at 262 ("[T]he temptation is great to rest on what data is readily available, allowing that to define the questions that are asked and the way in which they are answered."); Brian Leiter, *On So-Called "Empirical Legal Studies" and Its Problems*, BRIAN LEITER'S L. SCH. REPS. (July 6, 2010), http://leiterlawschool.typepad.com/leiter/2010/07/on-socalled-empirical-legal-studies.html ("[T]oo much of the work is driven by the existence of a data set, rather than an intellectual or analytical point.").

215 *Cf.* CHARLES F. MANSKI, PUBLIC POLICY IN AN UNCERTAIN WORLD: ANALYSIS AND DECISIONS 71 (2013) (criticizing reporting of the effects of an offer to treat patients where the effects of actual treatment should be the parameter of interest); Angus Deaton, *Instruments, Randomization, and Learning About Development*, 48 J. ECON. LIT. 424, 429 (2010) (criticizing the use of instrumental variables where the availability of an instrument—rather than the research question—determines the parameter of interest).

standard errors[216] or studies that predict reversals using multilevel hierarchical models and overdispersed logistic regression.[217] Defining citation counts or reversal rates as objects of interest may create an illusion of credibility, but it does not bring empirical research any closer to providing useful information that can improve the legal system.

There are many ways in which methodology can evolve to address questions unique to legal scholarship. In the following Sections, I briefly discuss two. First, because many meaningful legal phenomena cannot be objectively verified, empirical legal methodology will need to accommodate subjective phenomena. Second, because of the limitations of experimental approaches, legal empiricists will need to develop theories and methods that allow findings to generalize to diverse contexts.

C. *Accommodating Subjective Phenomena*

The studies of citation counts, reversal rates, and interjudge disparities discussed in Parts II–IV were all motivated by values internal to law, such as good judging and correctness. Such values have two important features in common: they are abstract and subjective. One can verify the contents of legal texts and judicial opinions, but the correctness of legal decisions and the quality of judicial reasoning will inevitably be disputed.

Many methodological approaches in empirical social science seek to avoid consideration of abstract concepts.[218] Although there are many advantages to concreteness in empirical research, such an approach is not always appropriate for addressing normative questions about law. Because many of the important normative goals are inherently subjective, any methodological approach that limits its focus to objectively measurable phenomena will have nothing to say about these goals. As Pound argued to Llewellyn,

> Those who long for an exact science analogous to mathematics or physics or astronomy have been inclined to seek exactness by excluding [the question of how justice ought to be administered] from jurisprudence altogether. But

---

216 *See* Choi et al., *supra* note 24, at 515-16 (examining the relationship between a judge's gender and citation counts).

217 *See, e.g.*, Gelman et al., *supra* note 124 (studying death sentence reversals).

218 *See, e.g.*, GARY KING ET AL., DESIGNING SOCIAL INQUIRY: SCIENTIFIC INFERENCE IN QUALITATIVE RESEARCH 109 (1994) (urging empirical social scientists to "maximize concreteness" and to "choose observable, rather than unobservable, concepts whenever possible"); *id.* at 111 ("If we have no alternative to using unobservable constructs . . . then we should at least *choose ideas with observable consequences*.").

> such a jurisprudence has only an illusion of reality. For the significant question is the one excluded.[219]

There can be no "exact science" of law. The legal empiricist's goal should not be to generate objective, assumption-free conclusions; there are few such conclusions that matter. Rather, legal empiricists need to find ways to combine objective findings with unverifiable assumptions to generate conclusions that are meaningful—at least according to some viewpoints—and to be explicit about how the assumptions are driving the results.

D. *Emphasizing Generalizable Results*

Empirical research examines what occurred in the past, but policymaking addresses what ought to be done in the future. To be relevant to normative questions, empirical research cannot merely explain what happened in the past, but must also interpret findings in ways that can inform decisions going forward. The normative goals of a research project will determine the extent to which the findings must be generalized to other contexts, which in turn will drive the study design.

Extrapolation may be unnecessary in a few instances, such as in research that examines purely historical questions. A study that sought to determine the authorship of the disputed Federalist Papers,[220] for example, had clear relevance to legal scholarship without any need to generalize the findings to other contexts. In other settings, what ought to be done in the future may depend directly on what happened in the past. In litigation, for example, whether the defendant caused the plaintiff's injury may directly determine whether the court ought to hold the defendant liable. To the extent that empirical research can shed light on such causal questions, it can provide direct guidance to legal decisionmakers without any need for generalizability.

Sometimes, extrapolating from the past to the future may be straightforward. If a drug proved effective in a well-controlled clinical trial, for instance, one might reasonably expect that it will have a similar effect in a comparable population in the future. In this setting, a simple comparison of the average effect on the treatment and control groups might be sufficient to determine whether the drug ought to be prescribed in the future.

---

219 Pound, *supra* note 6, at 703.

220 *See* Frederick Mosteller & David L. Wallace, *Inference in an Authorship Problem*, 302 J. AM. STAT. ASS'N 275 (1963).

Even this simple example, however, relies on several critical assumptions. First, it assumes that the impact of the medical treatment does not vary over time, so that the effect observed in the past accurately predicts the effect that will occur in the future. This may seem self-evident in the context of many medical trials, but it is less obvious in field research in law and the social sciences. Second, the example assumes that it was feasible to conduct a well-controlled trial with a study population that was representative of the target population. Third, it assumes that the measureable outcome has direct normative significance, so that a comparison of outcomes would provide sufficient information to guide policy. Fourth, it assumes that for normative purposes, we are only interested in the *average* effect of the treatment, and not any distributional effects.[221]

When these assumptions are satisfied, there is little concern about whether the results are generalizable, and researchers should design studies to have high internal validity.[222] When the assumptions do not hold, however, it is necessary to use research designs that provide both internal and external validity.[223] This necessarily requires making assumptions about how the results can be extrapolated.[224]

Many methodological differences among the disciplines stem from the plausibility of the above assumptions as applied to research questions within the respective disciplines. Some statisticians and social scientists have emphasized internal validity over external validity,[225] while many econometricians

---

221 *See* James J. Heckman, *The Scientific Model of Causality*, 35 SOC. METHODOLOGY 1, 20-21 (2005) (noting that standard statistical approaches that rely on randomization only estimate average treatment effects, and not distributional effects). Experimental studies can examine whether average effects vary among identifiable groups, but this would only capture part of the overall distributional impact.

222 A study design has "internal validity" if it "successfully uncovers causal effects for the population being studied." JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE, MOSTLY HARMLESS ECONOMETRICS: AN EMPIRICIST'S COMPANION 151 (2009).

223 A study design has "external validity" if the findings have predictive value in other contexts. *See id.*

224 *See* MANSKI, *supra* note 215, at 30-31 (describing the need for assumptions in extrapolating empirical findings and criticizing researchers who use untenable assumptions); Christopher A. Sims, *But Economics Is Not an Experimental Science*, 24 J. ECON. PERSP. 59, 60 (2010) ("We are always combining the objective information in the data with judgment, opinion and/or prejudice to reach conclusions.").

225 *See, e.g.*, PAUL R. ROSENBAUM, DESIGN OF OBSERVATIONAL STUDIES 56-57 (2010) ("The common view, which I share, is that internal validity comes first." (endnote omitted)); Donald T. Campbell, *Factors Relevant to the Validity of Experiments in Social Settings*, 54 PSYCH. BULL. 297, 310 (1957) ("If one is in a situation where either internal validity or representativeness must be sacrificed, which should it be? The answer is clear. Internal validity is the prior and indispensable consideration.").

have argued that both are essential.[226] Many economists insist that empirical findings must be interpreted in the context of a theoretical framework,[227] while some statisticians advocate reporting facts with minimal interpretation.[228] Much of the difference in perspectives stems from the fact that economics is primarily an observational science, while statistics is more oriented toward an experimental paradigm.[229] In medical research, treatments are typically tested in controlled trials and results may be easily extrapolated to other contexts. In such settings, simple comparisons between control and treatment groups may suffice. By contrast, economists are more often interested in forecasting the effects of policy interventions that cannot be implemented in advance,[230] which requires models than can predict what will happen under untestable counterfactuals.

The nature of theory required for extrapolating findings also varies by context. Toxicologists, for example, sometimes need to estimate the impact of exposure to minuscule amounts of environmental pollutants. When the impacts of such exposure are too small to be reliably measured in a controlled trial or an observational study, scientists and policymakers must

---

226 *See, e.g.*, MANSKI, *supra* note 215, at 37 (arguing that both internal and external validity are important goals of research design); Deaton, *supra* note 215, at 447-52 (same); Heckman, *supra* note 221, at 8 (same).

227 *See* Heckman, *supra* note 221, at 5 ("Blind empiricism unguided by a theoretical framework for interpreting facts leads nowhere."); Tjalling C. Koopmans, *Measurement Without Theory*, 29 REV. ECON. & STAT. 161, 162 (1947) (advocating "[f]uller utilization of the concepts and hypotheses of economic theory . . . *as a part of the processes of observation and measurement*").

228 *See* A.P. Dawid, *Causal Inference Without Counterfactuals*, 95 J. AM. STAT. ASS'N 407, 407 (2000) ("Nature is surely utterly indifferent to our attempts to ensnare her in our theories.").

229 *See* Guido W. Imbens, *An Economist's Perspective on Shadish (2010) and West and Thoemmes (2010)*, 15 PSYCHOL. METHODS 47, 48 (2010) ("Unlike biostatisticians, who often start from the perspective of a randomized clinical trial, economists start with the notion that individuals receive the treatments they received because they choose to."); Guido W. Imbens & Jeffrey Wooldridge, *Recent Developments in the Econometrics of Program Evaluation*, 47 J. ECON. LIT. 5, 19-20 (2009) (noting that in biostatistics, randomized experiments "are often viewed as the only credible approach to establishing causality," but that randomization "has played a much less prominent role" in economics and has "rarely been viewed as the sole method for establishing causality"); Sims, *supra* note 224, at 59 ("[E]conomics is not an experimental science and cannot be."). Nonetheless, many economists have been moving toward the experimental paradigm, and the use of laboratory and field experiments has grown dramatically in recent years. *See* ANGRIST & PISCHKE, *supra* note 222, at 12 (noting a trend toward randomized experimentation beginning in the 1980s).

230 *See* Heckman, *supra* note 221, at 17 ("Forecasting the effects of new policies is a central task of science and public policy analysis that the treatment effect literature ignores."); Imbens, *supra* note 229, at 48 (noting that economists often ask causal questions regarding the effects of novel treatments); Aviv Nevo & Michael D. Whinston, *Taking the Dogma Out of Econometrics: Structural Modeling and Credible Inference*, 24 J. ECON. PERSP. 69, 71 (2010) (advocating the use of structural modeling to "provide a way to extrapolate observed responses to environmental changes to predict responses to *other not-yet-observed changes*").

extrapolate from studies involving higher degrees of exposure. Often, the assumption is simply that there is a linear relationship between exposure to the pollutant and the likelihood of adverse health outcomes.[231] Here, the marginal effect of the pollutant is assumed to be constant within some relevant range, so that a measured marginal effect from one study is asserted to generalize to lower levels of exposure.

By contrast, economists often need more complex models to credibly extrapolate empirical findings to new contexts. Economists evaluating a proposed merger, for example, cannot feasibly test the effects of the merger using a controlled experiment. They could estimate the effects of past mergers, but such estimates may not reliably predict the effect of a future merger because firms have unique characteristics and industries are continually in flux.[232] Econometric studies of past data may reveal the structure of supply and demand in various markets, but economists must also rely on economic theory and game theoretic models of firm competition to predict how the proposed merger would affect the future behavior of firms and consumers.[233]

Recently, many scholars have advocated greater use of randomized trials in empirical legal research,[234] and some have conducted innovative field experiments that randomize legal representation[235] and law enforcement.[236]

---

231 *See* CASS R. SUNSTEIN, RISK AND REASON: SAFETY, LAW, AND THE ENVIRONMENT 164 (2002) (describing a decision by the Environmental Protection Agency to assume a linear relationship between arsenic exposure and cancer rates and arguing that the "assumption of linearity is not based on science . . ., but on a policy judgment, designed to err on the side of protecting health"); Adam M. Finkel, *A Second Opinion on an Environmental Misdiagnosis: The Risky Prescriptions of Breaking the Vicious Circle*, 3 N.Y.U. ENVTL. L.J. 295, 341-45 (1994) (arguing that the assumption of a linear relationship between exposure to carcinogens and cancer rates has a strong scientific basis).

232 *See* Nevo & Whinston, *supra* note 230, at 73-75 (describing the difficulties of using the causal effects of past mergers to predict the impact of future mergers).

233 *See id.* at 75 (describing the use of economic models to predict the impact of mergers).

234 *See, e.g.*, Michael Abramowicz et al., *Randomizing Law*, 159 U. PA. L. REV. 929 (2011) (arguing that policymakers and governments should test laws and regulations with randomized trials); D. James Greiner & Cassandra Wolos Pattanayak, *Randomized Evaluation in Legal Assistance: What Difference Does Representation (Offer and Actual Use) Make?*, 121 YALE L.J. 2118, 2127 (2012) ("[R]andomized trials . . . can provide credible answers on a far wider range of questions than is currently appreciated."); D. James Greiner et al., *The Limits of Unbundled Legal Assistance: A Randomized Study in a Massachusetts District Court and Prospects for the Future*, 126 HARV. L. REV. 901, 956 (2013) ("We believe such direct, randomized comparisons should be pursued, as they represent a powerful way to assess whether judicial best practices change case outcomes, litigant perceptions, and other outcomes of import.").

235 *See, e.g.*, W. VAUGHAN STAPLETON & LEE E. TEITELBAUM, IN DEFENSE OF YOUTH: A STUDY OF THE ROLE OF COUNSEL IN AMERICAN JUVENILE COURTS 49-51 (1972) (describing the study design for a randomized trial measuring the impact of counsel in juvenile hearings); Greiner & Pattanayak, *supra* note 234 (using randomization to evaluate the effects of an offer of

Although such experiments have the potential to offer credible estimates of the effects of various interventions, there are serious practical and ethical constraints on intentional randomization in the legal process.[237] These constraints are most acute for studies of adjudication, where randomization would be in deep tension with the need for reasoned decisionmaking.[238] Some studies exploit naturally occurring sources of randomness within the legal system,[239] but because natural experiments are not designed with research objectives in mind, scholars must exercise judgment in extrapolating the results to policy questions of interest. Even the most careful randomized

---

legal representation from a law school clinic on outcomes in hearings for unemployment benefits); Greiner et al., *supra* note 234 (using randomization to evaluate the effects of an offer of legal representation in housing cases); Carroll Seron et al., *The Impact of Legal Counsel on Outcomes for Poor Tenants in New York City's Housing Court: Results of a Randomized Experiment*, 35 L. & SOC'Y REV. 419 (2001) (using randomization to evaluate a legal assistance program for low-income tenants in New York City).

236 *See* Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 AM. SOC. REV. 261 (1984) (randomizing arrests of domestic violence suspects to estimate the impact of arrest on recidivism).

237 *See* FEDERAL JUDICIAL CENTER, EXPERIMENTATION IN THE LAW: REPORT OF THE FEDERAL JUDICIAL CENTER ADVISORY COMMITTEE ON EXPERIMENTATION IN THE LAW 25-30 (1981) (cataloging ethical issues inherent in randomizing legal experiments); Phyllis Jo Baunach, *Random Assignment in Criminal Justice Research: Some Ethical and Legal Issues*, 17 CRIMINOLOGY 435 (1980) (discussing ethical and legal concerns, such as arbitrary assignment and potential denial of benefits); Edna Erez, *Randomized Experiments in Correctional Context: Legal, Ethical, and Practical Concerns*, 14 J. CRIM. JUST. 389, 391-92 (1986) (surveying ethical issues associated with random assignment in criminal justice studies); Pascoe Pleasence, *Trials and Tribulations: Conducting Randomized Experiments in a Socio-legal Setting*, 35 J.L. & SOC'Y 8, 24-26 (2008) (discussing several ethical concerns of random control trials, such as the denial of benefits and compulsory participation); *see also* Greiner & Pattanayak, *supra* note 234, at 2127-32 (arguing that legal ethics permits randomization of an offer of legal representation but not randomization of the actual use of representation).

238 Judith Resnik, *Tiers*, 57 S. CAL. L. REV. 837, 840-41 (1984) (describing condemnation of a judge who flipped a coin to determine a defendant's sentence); Adam M. Samaha, *Randomization in Adjudication*, 51 WM. & MARY L. REV. 1, 5 (2009) ("[J]udges strongly condemn randomization for their own merits decisions.").

239 *See* David S. Abrams & Albert H. Yoon, *The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability*, 74 U. CHI. L. REV. 1145 (2007) (exploiting the random assignment of public defenders in felony cases to estimate the impact of attorney performance); James M. Anderson & Paul Heaton, *How Much Difference Does the Lawyer Make? The Effect of Defense Counsel on Murder Case Outcomes*, 122 YALE L.J. 154 (2012) (exploiting the random assignment of defense attorneys to assess the performance of public defenders relative to appointed counsel in murder cases); Donald P. Green & Daniel Winik, *Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism Among Drug Offenders*, 48 CRIMINOLOGY 357 (2010); Jeffrey R. Kling, *Incarceration Length, Employment, and Earnings*, 96 AM. ECON. REV. 863, 865-66 (2006) (exploiting the random assignment of judges to estimate the impact of incarceration on subsequent labor market outcomes); *see also supra* notes 156-158 and accompanying text (describing studies that examine the impact of randomly assigned judges on case outcomes).

trials can be compromised by subject attrition,[240] crossover between the treatment and control groups,[241] spillover effects,[242] or even conscious efforts by nonparticipants to undermine the research.[243] As Angus Deaton writes,

> [C]onducting good [randomized controlled trials] is exacting and often expensive, so that problems often arise that need to be dealt with by various econometric or statistical fixes. There is nothing wrong with such fixes in principle . . . but their application takes us out of the world of ideal [randomized controlled trials] and back into the world of everyday econometrics and statistics.[244]

The use of randomized trials is a welcome development in empirical legal scholarship, but many important research questions cannot be resolved with standard experimental methods. To address such questions, there is no alternative but to rely on theory, contestable assumptions, and empirical estimates that can be extrapolated from other contexts.

For example, it would be difficult to estimate the effects of proposed sentencing guidelines through a randomized trial. One could examine past changes in sentencing guidelines, but this would not necessarily predict the impact of future guidelines; the guidelines at issue might not be identical, the composition of the judiciary would have changed, and mandatory minimum penalties might be different. Any effort to predict the effect of future policies would require a model of judicial sentencing behavior that enables results from prior studies to be generalized to new contexts.

---

240 *See* Abramowicz et al., *supra* note 234, at 957-59 (noting that attrition "may bias experimental results when the attrition rate depends on selection for treatment").

241 *See id.* at 959-60 (discussing the problems related to crossover, such as those that occur "if well-connected people . . . thwart random assignment.").

242 *See id.* at 960 (discussing how a controlled trial can be contaminated by spillover effects if the control group is also influenced by the treatment); Edward Miguel & Michael Kremer, *Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities*, 72 ECONOMETRICA 159, 160 (2004) ("[I]f externalities benefit the comparison group, outcome differences between the treatment and comparison groups will understate the benefits of treatment on the treated.").

243 *See* Richard D. Schwartz, *Foreword* to STAPLETON & TEITELBAUM, *supra* note 235, at xii (noting judicial opposition to Stapleton and Teitelbaum's randomized study of representation in juvenile delinquency hearings); Dave Hoffman, *Experiments in Lawyering: Does the Harvard Legal Aid Bureau Deserve a Merit Badge?*, CONCURRING OPINIONS (Dec. 21, 2010), http://www.concurringopinions.com/archives/2010/12/experiments-in-lawyering-does-the-harvard-legal-aid-clinic-deserves-a-merit-badge.html (describing how a legal services organization stopped referring clients to Harvard Legal Aid Bureau due to the Bureau's participation in a randomized trial).

244 Deaton, *supra* note 215, at 447.

In addition, as I have emphasized throughout this Article, the measureable outcomes in empirical legal research often do not have direct normative significance. Although a randomized trial can provide a credible estimate of the average effect of an intervention, this estimate may not be meaningful if the underlying outcome variable lacks normative significance. Finally, legal scholars are not merely interested in the average effects of treatments, but also distributional effects; the degree to which likes are treated alike[245], for example, cannot be assessed by estimating the average effect of any legal policy.[246]

The diverse research questions in empirical legal scholarship will almost certainly require diverse methods. Simple methods may be most appropriate in some settings and more technical approaches in others. But the methods used in empirical legal scholarship should be determined by substantive research questions, not the other way around.

CONCLUSION

Pound and Llewellyn shared a worthy goal: using empirical social science to improve the law. Both understood that social scientists must engage with values in order to advance legal reform. They disagreed sharply about values, but at least they were debating the right questions.

The legal realists and sociological jurists never succeeded in establishing empirical research within the mainstream of legal scholarship. Now, contemporary scholars are bringing new energy to the empirical study of the legal system. They have been steadily improving in methodological sophistication, but in the process, they have lost much of their connection with law's normativity. All too often, research into what *is* fails to inform debates about what *ought* to be.

This Article has argued that empirical legal scholars must clarify the normative issues at stake in their research and be more explicit about the asserted connections between measurable data and normative claims. Doing this well will require careful thinking about the questions that animate empirical legal research and the development of new frameworks and methods that can provide meaningful answers. This is no doubt a challenging task. But if empirical studies lose sight of the goals of legal scholarship, what will the counting be good for?

---

245 *See supra* Section IV.B.

246 *See supra* note 221 and accompanying text.