The Rules of Inference

The Rules of Inference

Lee Epstein & Gary King[*]

Although the term "empirical research" has become commonplace in legal scholarship over the past two decades, law professors have, in fact, been conducting research that is empirical—*i.e.*, learning about the world using quantitative data *or* qualitative information—for almost as long as they have been conducting research. For just as long, however, they appear to have been proceeding with little awareness of, much less compliance with, many of the rules of inference, and without paying heed to the key lessons of the revolution in empirical analysis that has been taking place over the last century in other disciplines. The tradition of including some articles devoted exclusively to the methodology of empirical analysis, so well represented in journals in traditional academic fields, is virtually nonexistent in the nation's law reviews. As a result, readers learn considerably less accurate information about the empirical world than the studies' stridently stated, but often overly confident, conclusions suggest.

To remedy this situation—both for the producers and consumers of empirical work—we adapt the rules of inference used in the natural and social sciences to the special needs, theories, and data in legal scholarship, and explicate them with extensive illustrations from existing research. We also offer suggestions for how the infrastructure of teaching and research at law schools might be reorganized so that it can better support the creation of first-rate empirical research without compromising other important objectives.

Introduction

Just as research in the social and natural sciences addresses a wide array of theoretical, methodological, and substantive concerns, so too does scholarship produced by legal academics. The law reviews are replete with articles ranging from the normative to the descriptive, from narrow doctrinal analyses to large-*n* statistical investigations. Some studies advocate legal reform; others intend solely to add to the store of academic knowledge. And, yet, in all this variation in approach, in all this diversity in purpose, effect, and even intended audience, many, if not most, of these studies evince a common characteristic: a concern, however implicit, with *empiricism*—basing conclusions on observation or experimentation—and *inference*[1]—using facts we know to learn about facts we do not know.

This may seem a puzzling, even odd, statement to legal academics. After all, in this community the word "empirical" has come to take on a particularly narrow meaning—one associated purely with "statistical techniques and analyses," or quantitative data.[2] But empirical research, as natural and

---

[*] Lee Epstein (epstein@artsci.wustl.edu, http://www.artsci.wustl.edu/~PoliSci/Epstein/) is the Edward Mallinckrodt Distinguished University Professor of Political Science and Professor of Law, Washington University in St. Louis. Gary King (king@harvard.edu, http://GKing.Harvard.edu) is Professor of Government, Harvard University and Senior Science Advisor, Evidence and Information for Policy Cluster, World Health Organization. We thank the National Science Foundation (SBR-9729884, SBR-9753126, and IIS-9874747) and the National Institutes of Aging (P01 AG17625-01) for research support; Micah Altman, Anne Joseph, Susan Appleton, Stuart Banner, Shari Diamond, Bill Eskridge, Barry Friedman, Brian Glenn, Jack Goldsmith, Andrew Holbrook, Dan Keating, Bob Keohane, Jack Knight, Bert Kritzer, Ron Levin, Joan Nix, Eric Posner, Dan Schneider, Nancy Staudt, Sid Verba, and Adrian Vermeule for helpful comments and discussions; Dana Ellison, Andrew Holbrook, and Jeff Staton for research assistance; and Scott Norberg for his data.

[1] See Gary King, Robert O. Keohane, & Sidney Verba, Designing Social Inquiry: Scientific Inference in Qualitative Research (1994), at 46.

[2] See, for example, Michael Heise, "The Importance of Being Empirical," 26 Pepperdine L Rev 807 (1999): "[W]hen I speak of empirical legal scholarship I refer only to the subset of empirical legal scholarship that uses statistical techniques and analyses. By statistical techniques and analyses I mean studies that employ data (including systematically coded judicial opinions) that facilitate descriptions of or inferences to a larger sample or population as well as replication by other scholars;" Craig A. Nard, "Empirical Legal Scholarship: Reestablishing a Dialogue between the Academy and

social scientists recognize, is far broader than these associations suggest. The word "empirical" denotes evidence about the world based on observation or experience. That evidence can be numerical (quantitative) or non-numerical (qualitative);[3] neither is any more "empirical" than the other. What makes research empirical is that it is based on observations of the world—*i.e., data*, which is just a term for facts about the world. These facts may be historical or contemporary, based on legislation or case law, the results of interviews or surveys, or the outcomes of secondary archival research or primary data collection. Data can be precise or vague, relatively certain or very uncertain, directly observed or indirect proxies, and they can be anthropological, interpretive, sociological, economic, legal, political, biological, physical, or natural. As long as the facts have something to do with the world, they are data, and as long as research involves data that is observed or desired, it is empirical.

Under this definition of "empirical," assertions that "the amount of theoretical and doctrinal scholarship…overwhelms the amount of empirical scholarship,"[4] ring hollow. For, as even the most casual reader of the nation's law reviews must acknowledge, a large fraction of legal scholarship makes at least some claims about the world based on observation or experience.[5]

In fact, in terms of legal scholarship, it is only the *purely* normative or theoretical that is *not* empirical. But even many articles whose main purpose is normative often invoke empirical arguments to shore up their normative points—such as offering the positive empirical implications of adopting their preferred policy. Staudt's essay on the Federal Income Tax Code is typical.[6] A largely normative piece exploring conventional wisdom among tax scholars that housework should not or cannot be taxed, it ends with several unambiguously empirical claims:

> This reform, together with the market-oriented reform, would go far toward changing society's views of the value of productive activities carried out both in the home and in the market, and more importantly, it would represent a critical step in achieving greater economic security for women.[7]

The same holds for most doctrinal work, including the many studies that take issue with a particular line of court decisions or the logic used in them. Such is Sklansky's investigation of the U.S. Supreme Court's use of "new originalism"[8] to resolve Fourth Amendment search and seizure

---

the Profession," 30 Wake Forest L Rev 347 (1995): Empirical "scholarship [is] based on a detailed statistical study and analysis from which one could draw conclusions and formulate or reformulate policy." Even Peter H. Schuck, "Why Don't Law Professors Do More Empirical Research?," 39 J Legal Educ 323 (1989), at 323, who recognizes that many forms of legal scholarship are, in fact, "empirical," defines empirical scholarship as "primarily" "statistical studies."

[3] Many scholars divide empirical research into two types or styles: quantitative, which uses numbers and statistical methods, and qualitative, which does not rely on numbers but on historical materials, intensive interviews, and the like. See, for example, Earl Babbie, The Basics of Social Research (1999); William J. Dixon, "Research on Research Revisited: Another Half Decade of Quantitative Research on International Organizations," 31 Intl Org 65 (1977). This distinction may be relevant for some purposes but not for ours. That is because the rules governing empirical research, to which we devote most of this Article, apply with equal force to qualitative and quantitative work. See King, Keohane, & Verba, Designing Social Inquiry , supra note 1.

[4] Heise, "The Importance of Being Empirical," supra note 2.

[5] We conducted a systematic, but still preliminary, analysis of articles published over the last five years in six top law reviews (*Chicago, Columbia Harvard, NYU, Stanford, and Yale*). Virtually all the articles we have examined thus far include at least some empirical analyses or empirical hypotheses. Just under half used empirical evidence to evaluate their empirical hypotheses, and about half of these are solely qualitative.

[6] Nancy C. Staudt, "Taxing Housework," 84 Georgetown L J 1571 (1996).

[7] Id., at 1647.

[8] According to David A. Sklansky, "The Fourth Amendment and Common Law," 100 Colum L Rev 1739 (2000), under "new originalism" the Court, assesses whether searches and seizures are "unreasonable" in violation of the Fourth Amendment by determining whether 18th century common law allowed them.

cases. The author's primary objective is to provide support for the claim that this "new" approach "should be unattractive even to those who are generally sympathetic to originalism."[9] As a prelude to making this normative argument, however, Sklansky must demonstrate the empirical claim that the Court has, in fact, adopted a new originalist approach. And that demonstration calls for an inference or claim about the real world, as does Staudt's move from "this policy should be changed" to "if this policy were changed the following problems may be ameliorated."

We applaud these studies and, indeed, all those that express a concern with empirics whether implicitly or explicitly. And the producers of these studies, legal academics, apparently agree, hoping to see and intending to produce more of this sort of work.[10] To them, along with other members of the legal community, including judges and lawyers,[11] research that offers claims or make inferences

---

[9] Id., at 1745.

[10] See Nard, "Empirical Legal Scholarship: Reestablishing a Dialogue between the Academy and the Profession," supra note 2 for the results of his telephone survey of forty "randomly selected" law professors. Given the procedures Nard used, we would not want to make a strong claim that his sample reflects the relevant population. (For more on sampling, see infra Part VIII). Nonetheless, we are impressed with the number of articles, produced by legal academics, proclaiming the need for more real-world research—and research on a wide-array of legal topics at that. A few examples include: Heise, "The Importance of Being Empirical," supra note 2; Derek Bok, "A Flawed System of Law Practice and Training," 22 J Legal Educ 570 (1983); Michael Dorf, "The Supreme Court, 1997 Term—Foreword: The Limits of Socratic Deliberation," 112 Harv L Rev 4 (1998); Schuck, "Why Don't Law Professors Do More Empirical Research?," supra note 2; Michael A. Livingston, "Reinventing Tax Scholarship: Lawyers, Economists, and the Role of the Legal Academy," 83 Cornell L Rev 365 (1998); Joseph A. Guzinksi, "Government's Emerging Role as a Source of Empirical Information in Bankruptcy Cases," 1 Am Bankr Instit J 1 (1998); Kevin R. Reitz, "Sentencing Guideline Systems and Sentence Appeals: A Comparison of Federal and State Experiences," 91 Nw U L Rev 1444 (1997); Nancy S. Marder, "The Myth of the Nullifying Jury," 93 Nw U L Rev 877 (1999).

[11] Judge Richard A. Posner is perhaps most closely associated with this position. See, for example, Richard A. Posner, "The Summary Jury Trial and Other Methods of Alternative Dispute Resolution: Some Cautionary Observations," 53 U Chi L Rev 366 (1986); Richard A. Posner, "Madison Lecture: Against Constitutional Theory," 73 NYU L Rev 1 (1998); Richard A. Posner, The Problematics of Moral and Legal Theory (1999). For others in the long list of distinguished members of the legal community to call for more work grounded in real-world observations, see Heise, "The Importance of Being Empirical," supra note 2. Indeed, such calls trace at least as far back as Oliver W. Jr. Holmes, "The Path of Law," 10 Harv L Rev 457 (1897) (predicting that "For the rational study of law the black letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics") and the Legal Realists. See Michael Rustad & Thomas Koenig, "The Supreme Court and Junk Science: Selective Distortion in Amicus Briefs," 72 NC L Rev 91 (1993); John H. Schlegel, "American Legal Realism and Empirical Social Science: From the Yale Experience," 28 Buff L Rev 459 (1979).

We also note that even members of legal community who have not explicitly exhorted law professors to produce more empirical research, have acknowledged its increasing role in the study and practice of law. Justice Stephen Breyer, in his introduction to the Federal Judicial Center's Reference Manual on Scientific Evidence (available online at: http://www.air.fjc.gov/public/fjcweb.nsf/pages/16), opened with the following lines: "In this day and age of science, science should expect to find a warm welcome, perhaps a permanent home in our courtrooms. The reason is a simple one. The legal disputes before us increasingly involve the principles and tools of science." He went on to provide several examples of U.S. Supreme Court cases that required the justices to understand statistics and, more generally, the rules of inference. By the same token, when he served on the on the U.S. Court of Appeals for the Fifth Circuit, John Minor Wisdom wrote: "What seemed at first to be antagonism between social science and law has now developed into a love match. What began in the field of education spread to many other fields. In case after case the Fifth Circuit, among other courts, has relied on studies…to show pollution, unlawful exclusion of blacks from the jury system, employment discrimination, arbitrary or discriminatory use of the death penalty, discrimination against women, the need for reapportionment, and the cure for malapportionment of various public bodies." John Minor Wisdom, "Random Remarks on the Role of Social Sciences n the Judicial Decision-Making Process in School Desegregation Cases," 39 L & Contemp Probs 143 (1975).

At least at the level of the U.S. Supreme Court, data seem to support Justice Breyer's and Judge Wisdom's observations. After examining citations to evidence drawn from the real world in the Court's abortion and sex discrimination cases, not only do Rosemary J. Erickson & Rita J. Simon, The Use of Social Science Data in Supreme

based on observations about the real world—on topics ranging from the imposition of the death penalty[12] to the effect of court decisions on administrative agencies[13] to the causes of fraud in the bankruptcy system[14] to the use of various alternative dispute resolution mechanisms[15]—"can play an important role in public discourse…and can affect our political system's handling" of many issues.[16] At the very least, legal academics are making extensive use of existing studies: Citations in the law reviews to real-world research, even under highly restrictive definitions of what constitutes "real-world research," have nearly doubled over the last two decades.[17]

At the same time, *an immense body of work in empirical legal scholarship is deeply flawed.* We base this claim primarily on a review of the legal literature we conducted[18]—a review that revealed many proceeding with research agendas, however diverse their goals might be, with little awareness of,

---

Court Decisions (1998), conclude that "the use of...data by the Supreme Court has increased over time." They also claim that "[a] scientifically-reliable study…is almost assured of being entered into the court record," at 153.

[12] See, for example, Ronald J. Tabak, "Empirical Studies of the Modern Capital Sentencing System: How Empirical Studies Can Affect Positively the Politics of the Death Penalty," 83 Cornell L Rev 1431 (1998) (claiming that "empirical studies concerning the death penalty can play an important role in public discourse on capital punishment…But constructive discourse will occur only if opponents of the death penalty educate themselves about what the empirical studies show.").

[13] See, for example, Peter H. Schuck & Donald Elliot, "To the Chevron Station: An Empirical Study of Federal Administrative Law," 1990 Duke L J 984 (1990) (asserting that "although there may be widespread agreement that judicial review of agency action matters, there is no consensus about precisely *how* and *under what circumstances* it matters.); Jerry Mashaw & David Harfst, "Regulation and Legal Culture: The Case of Motor Vehicle Safety," 4 Yale J Reg 257 (1987) ("The normative expectations of administrative lawyers have seldom been subjected to empirical verification of a more than anecdotal sort. And different observers rely upon different anecdotes.").

[14] See, for example, Guzinksi, "Government's Emerging Role as a Source of Empirical Information in Bankruptcy Cases," supra note 10, at 1 (noting that "Efforts to change the bankruptcy laws have been hampered by the lack of reliable answers to any number of empirical questions…).

[15]See, for example, Posner, "The Summary Jury Trial and Other Methods of Alternative Dispute Resolution," supra note 11, at 393 (writing that "Beginning with the promulgation of the Federal Rules of Civil Procedure in 1938 and acceleration with the caseload explosion that began around 1960, the federal courts have been an arena of massive experimentation in judicial administration. The milestones include liberalized class actions, one-way attorney's fee shifting, expansive pretrial discovery, managerial judging, the six-person jury, limited publication of appellate opinions, greater use of judicial adjuncts, and now 'alternative dispute resolution,' illustrated by the summary jury trial and court-annexed arbitration. Very few of these experiments have been conceived or evaluated in a scientific spirit and this may help explain why the federal courts remain in a state of crisis. Maybe a dose of social science is the thing, or one of the things, that the system needs.")

[16] Tabak, "Empirical Studies of the Modern Capital Sentencing System: How Empirical Studies Can Affect Positively the Politics of the Death Penalty," supra note 12, at 1431. A recent example of this phenomenon is Gratz v. Bollinger, __ F. Supp. __ (E.D. Mich. 2000), in which Judge Patrick J. Duggan relied heavily on real-world evidence to uphold the University of Michigan's current affirmative action program. According to Judge Duggan, the defendants had presented him with "solid evidence regarding the educational benefits that flow from a racially and ethnically diverse student body." In commenting on the case, Derek Bok, coauthor of a study that offered social-science evidence to defend affirmative action, Derek Bok & William G. Bowen, The Shape of the River: Long Term-Consequences of Considering Race in College and University Admissions, (1998), observed that in Regents of the University of California v. Bakke, 438 U.S. 265 (1978) "Justice Powell said there was no evidence of the educational benefits of diversity, but he was willing to accept the judgment of educators that there is. Courts are now able to look to data in order to see how much weight to put on this claim." Quoted in Jacques Steinberg, "Defending Afirmative Action with Social Science," New York Times (2000), at 41.

[17] See Robert C. Ellickson, "Trends in Legal Scholarship: A Statistical Study," 29 J Legal Stud 517 (2000). He writes that the citation data "hint that law professors and their students have become more inclined to produce…quantitative analyses."

The search terms he used to produce the citations to which we refer in the text are "statistic" and "significance." Unlike, say, "empirical" these are, according to Ellickson, more likely to appear in articles containing original research. Of course, these terms only apply to quantitative empirical research.

[18] See infra Part I.

much less compliance with, the rules of inference that guide empirical research in the social and natural sciences. The sustained, self-conscious attention to the methodology of empirical analysis, appearing in articles devoted solely to methodology in the vast majority of traditional academic fields, is virtually nonexistent in the nation's law reviews. As a result, readers learn considerably less accurate information about the empirical world than the studies' stridently stated, but overly confident, conclusions suggest.

Given that legal scholarship—perhaps more so and more quickly than most other research—has the potential to influence public policy as it is promulgated by judges, legislators, and bureaucrats,[19] this is highly problematic. It is especially so when that influence comes in studies assessing the likely consequences of particular changes in public policy, evaluating the impact of existing public programs, or any others that affect the real world in a timely manner.[20] But even if the content of the

---

[19] See generally Scott Martin, "The Law Review Citadel: Rodell Revisited," 71 Iowa L Rev 1093 (1986); Richardson, "Law Reviews and the Courts," 5 Whittier Law Review 385 (1983); Max et al. Stier, "Law Review Usage and Suggestions for Improvement: A Survey of Attorneys, Professors, and Judges," 44 Stan L Rev 1467 (1992); Alex Kozinski, "Who Gives a Hoot about Legal Scholarship?," 37 Houston L Rev 295 (2000).

[20] For particular examples, see Erickson and Simon, *The Use of Social Science Data in Supreme Court Decisions*, supra note 10 ; Wallace D. Loh, Social Research in the Judicial Process (1984); John Monahan & Laurens Walker, Social Science in Law (1998); Paul L. Rosen, The Supreme Court and Social Science (1972).

Unfortunately, a good many of the examples reported in these works reveal the problems that can ensue when policy makers base their decisions on poorly designed research. Consider, for example, the plight of three studies, all appearing in law reviews and all reaching the same conclusion: No significant differences exist between the decisions reached by six- and twelve-member juries. Note, "Six-Member and Twelve-Member Juries: An Empirical Study of Trial Results," 6 University of Michigan Journal of Law Reform 671 (1973); Note, "An Empirical Study of Six and Twelve-Member Jury Decision-Making Processes," 6 University of Michigan Journal of Law Reform 712 (1973); Bermant & Coppock, "Outcomes of Sex- and Twelve-Member Jury Trials: An Analysis of 128 Civil Cases in the State of Washington," 48 Wash L Rev 593 (1973). Within a matter of months of publication, the trio found their way into a U.S. Supreme Court decision, *Colgrove v. Battin* 413 U.S. 149 (1973), which, in line with the studies' results, held that six-person juries satisfy the Seventh Amendment's guarantee of a jury trial in civil cases. It mattered not that two of the three articles were unsigned student notes or that none was published in a top law review: All housed timely and strong claims about the world, and the justices responded accordingly.

Of course we do not know whether these three studies, rather than legal arguments or other factors, are what convinced the U.S. Supreme Court to rule the way it did. What is beyond speculation, however, is that the majority opinion cited the three with approval: "…very recent studies have provided convincing empirical evidence of the *Williams* [v. Florida, 413 U.S. 149 (1973)] conclusion that there is no discernible difference between results reached by the two different-sized juries." *Colgrove,* 413 U.S. at 160. What is also beyond speculation, these then "very recent studies" were seriously flawed. One actually relied on an experiment in which "jurors" (actually college students) viewed a videotaped trial that was so obviously and heavily biased that not one of the sixteen experimental juries reached a verdict favoring the plaintiff. Note, "An Empirical Study of Six and Twelve-Member Jury Decision-Making Processes." Hence, even if six- and twelve-person juries do, in fact, render distinct verdicts in the more commonly occurring cases (*i.e.,* those that are more evenly balanced between the defense and the plaintiff), the experiment would have missed the effect of size entirely. See Michael Saks, "Ignorance of Science is No Excuse," 10 Trial 18 (1974) and Hans Zeisel & Shari Seidman Diamond, "Convincing Empirical Evidence on the Six Member Jury," 41 U Chi L Rev 281 (1974).

This is but one illustration. Other problems with the jury research cited in *Colgrove* were so transparent that Michael Saks, a law professor with a Ph.D., was driven to write:

> The quality of…scholarship displayed would not win a passing grade in a high school psychology class…. The Court did look at empirical studies and did understand the stated findings. What the Court did not realize was that not all empirical studies are equal…. Studies using poor methods tell one nothing; but they can seriously mislead because their findings still may properly be called 'empirical.' The empirical studies cited in *Colgrove v. Battin,* because of their faulty methods, said much less than the Court thought they were saying.

Saks, "Ignorance of Science is No Excuse," at 18. For other critiques of the jury studies cited in *Colgrove*, see Zeisel and Diamond, "Convincing Empirical Evidence on the Six Member Jury"; Richard O. Lempert, "Uncovering "Nondiscernible" Differences: Empirical Research and the Jury Size Cases," 73 Mich L Rev 644 (1975); Robert H.

concluding sections in law review articles—often prescriptions or policy implications of the research—went largely ignored or were geared primarily to other academics, our concerns about the present state of legal scholarship would remain; after all, regardless of the purpose, effect, or intended audience of the research, academics have an obligation to produce work that is reliable.

Miller, "Six of One is Not a Dozen of the Other: A Reexamination of Williams v. Florida and the Size of State Criminal Juries," 146 U Pa L Rev 621 (1998).

The studies to which Saks refers are quantitative (numerical), but empirical research relying on other more qualitative (non-numerical) forms of evidence has been just as influential, even though it may be equally problematic—with research on the Second Amendment nicely illustrating the point. In response to court decisions, see, *e.g.,* United States v. Miller, 307 U.S. 174, 178 (1939) (holding that the purpose of the Second Amendment is "to assure the continuation and render possible the effectiveness" of state militias); Burton v. Sills, 53 N.J. 86, 99 248 A. 2d 521, 527 (1968); appeals dismissed, 394 U.S. 812 (1969) (finding that "Congress…may regulate interstate firearms so long as the regulation does not impair the maintenance of the active, organized militias of the states"); U.S. v. Hale, 978 F. 2d. 1016, 1019 (1992), cert denied, 507 U.S. 997 (1993) (refusing to "conclude that the Second Amendment protects the individual possession of military weapons"), and law review articles suggesting that the Amendment guarantees only a *collective* right of the states to arm their militias (only thirty-seven of the forty-one law review articles published between 1980 and 1994 apparently took an individual rights approach to the Second Amendment. See Amicus Curiae Brief of Academics for the Second Amendment at 7, n4, filed in United States v. Lopez [No. 93-1260 in the U.S. Supreme Court]; cited in Glenn Harlan Reynolds, "A Critical Guide to the Second Amendment," 62 Tenn L Rev 461 [1995]), Levinson, Lund and other legal academics argue that judges should begin interpreting the Second Amendment as establishing an *individual* right to keep and bear arms—a right that, like those of religion speech, and others contained in the U.S. Constitution, Congress cannot abridge. SeeSanford Levinson, "The Embarrassing Second Amendment," 99 Yale L J 637 (1989); Nelson Lund, "The Second Amendment, Political Liberty, and the Right of Self Preservation," 39 Ala L Rev 103 (1987); Nelson Lund, "The Past and Future of the Individual's Right to Arms," 31 Ga L Rev 1 (1996); David E. Vandercoy, "The History of the Second Amendment," 28 Valparaiso L Rev 1007 (1994). Supporting their position, they claim, is a vast array of documentary evidence indicating, among other things, that the framers intended the Amendment to safeguard an individual right. Attorneys representing various anti-gun control groups have, naturally enough, worked these arguments into their legal briefs. See, for example, Brief Supporting Appellee of Amicus Curiae National Rifle Association of America, filed in United States v. Emerson (No. 99-10331 in the U.S. Court of Appeals for the Fifth Circuit; available at: www.nraila.org/research/20000120-SecondAmendment-001.shtml), and, in turn, at least one federal judge made extensive use of them to support his holding that the Second Amendment prohibits the government from denying gun ownership to people under restraining orders, United States v. Emerson, 46 F. Supp. 2d 598 (1999) (dismissing an indictment against Timothy Emerson, who was charged with possession of a firearm while being under a restraining order, in violation of 18 U.S.C. § 922(g)(8)).

Historians and other academics have now jumped into the debate. After examining the evidence offered in the Second Amendment articles, Garry Wills—who was surprised to learn that most legal periodicals are not peer reviewed—wondered whether "our law journals were being composed by Lewis Carroll using various other pseudonyms." Garry Wills, "To Keep and Bear Arms," *New York Review of Books*, September 21 1995. (For other critiques, see, e.g., Saul Cornell, "Commonplace or Anachronism: The Standard Model, the Second Amendment, and the Problem of History in Contemporary Constitutional Theory," 16 Const Commen 221 (1999); Michael A. Bellesiles, Arming America: The Origins of the National Gun Culture (2000)). Others, e.g., Joyce Lee Malcolm, To Keep an Bear Arms (1994), have lent support to particular claims made by Levinson and other legal academics.

Surely this debate will continue. We only wish to note here that the law review articles which seem to have precipitated it (at least those making claims about the desires of the framers of the U.S. Constitution) violate many of the rules we discuss in the balance of this Article—e.g., the authors do not typically explain why they marshal particular pieces of evidence and neglect others— and that these violations can have dramatic effects on the authors' conclusions. See also Morgan Cloud, "Searching through History; Searching for History," 63 U Chi L Rev 1707 (1996) making a similar point about "lawyers' histories of the Fourth Amendment" when he writes that "they have been incomplete, reviewing only a small fraction of the relevant historical data, and they have been partisan, selectively deploying fragments of the historical record to support their arguments about the Amendment's meaning." For other recent treatments of the problems in legal scholarship that seeks to uncover original intent or understandings see Martin S. Flaherty, "History Right? Historical Scholarship, Original Understanding, and Treaties as 'Supreme Law of the Land'," 99 Colum L Rev 2095 (1999) and Emil A. Kleinhaus, "History as Precedent: The Post-Originalist Problem in Constitutional Law," 110 Yale L J 121 (2000).

Empirical scholarship, again regardless of its purpose, effect, or intended audience, that does not follow the time-honored rules of inference is unlikely to fulfill this obligation. Unfortunately, too much legal scholarship falls into this category. Too much legal scholarship ignores the rules of inference and applies instead the "rules" of persuasion and advocacy; these have an important place in legal studies, but not when the goal is to learn about the empirical world.

One source of the problem almost certainly lies in the training law professors receive,[21] and their general resulting approach to scholarship. While Ph.D.s are taught to subject their favored hypothesis to every conceivable test and data source, seeking out all possible evidence *against* their theory, attorneys are taught to amass all the evidence *for* their hypotheses and distract attention from anything that might be seen as contradictory information. An attorney who treats a client like a hypothesis would be disbarred; a Ph.D. who advocates a hypothesis like a client would be ignored.[22] But when attorneys—as law professors—move from the courtroom to the faculty commons, to a world where the truth, and not just a particular version of it, matters, defending theories and hypotheses as if they were clients in need of the best possible representation and dismissing competitors out of hand or ignoring them entirely is highly problematic.[23] That is because in empirical research, challenging a theory with the best possible opposing arguments is what makes the strongest case for a theory.[24]

---

[21] See, for example, Heise, "The Importance of Being Empirical," supra note 2; Nard, "Empirical Legal Scholarship: Reestablishing a Dialogue between the Academy and the Profession," supra note 2; Gerald N. Rosenberg, "Across the Great Divide (Between Law & Political Science)," 3 The Greenbag 267 (2000). Lack of training may be the primary reason but it is not the only one. Others that scholars suggest—such as their perception that real-world research is more time consuming (see, e.g., Julius G. Getman, "Contributions of Empirical Data to Legal Research," 35 J Legal Educ 489 (1985)), and harder to conduct (see, e.g., Lawrence M. Friedman, "The Law and Society Movement," 38 Stan L Rev 763 (1986); Heise, "The Importance of Being Empirical," supra note 2), than other forms of legal scholarship —are, to us, seriously flawed. Nonetheless, they do implicate the infrastructure law schools could, but do not presently, provide for such work. We offer suggestions for building the necessary infrastructure in Part IX.

[22] These training differences have led to numerous misunderstandings and accusations being lobbed back and forth across the disciplinary divide. What critics miss, however, is that these differences in perspective are consistent with the markedly differing goals of the two sides. Among other things, scientists aim to conduct good empirical research, to learn about the world, and to make inferences. Lawyers and judges, and hence law professors, in contrast specialize in *persuasion*. Lawyers need to persuade judges and juries to favor their clients, and the rules of persuasion in the adversary system are different from than the rules of empirical inquiry. As actors who lack the power of enforcement, judges attempt to enhance the legitimacy of their actions by persuading the parties to lawsuits, the executive branch, the public, and so on that judicial decisions have a firm basis in the established prior authority of law rather than in the personal discretion of judges—even when that authority is inconsistent, illogical, historically inaccurate, or nonexistent. Law reviews, in turn, are filled in part with shadow court opinions (with many articles written by former law clerks), rearguing, supporting, or practicing this art of political persuasion. For the purposes of political persuasion, "judges can make claims about history, philosophy, economics, and political science that would be regarded as shallow or discreditable by practitioners of those disciplines but that do not offend the minimal standards of acceptability for performance of their own distinctive craft." Richard H. Fallon, Jr. Non-Legal Theory in Judicial Decisionmaking, 17 Harv J L & Publ Pol 93. We can see the lack of contradiction only by recognizing that the Ph.D.s' goals of learning about the empirical world differ from the J.D.s' goal of political persuasion. Of course, lawyers also understand that ideas about the world can be wrong, and Ph.D.s use persuasion to convince others of the importance of their ideas, but the institutional differences in the fields nevertheless remain stark.

[23] See Arthur S. Miller, "The Myth of Objectivity in Legal Research and Writing," 18 Cath U L Rev 290 (1969) arguing that "objectivity" in their legal writings is impossible for lawyers to achieve because of their commitment to advocacy.

[24] Our point is not that lawyers are worse at applying the rules of inference than scholars in the natural and social sciences. Such may often or even usually be true, but these other fields have plenty of their own problems, none of which, however, are the subject of the present Article.

But enough. Our purpose here is not to lambaste law professors or the scholarship they generate;[25] it is rather to make a productive contribution intended to mitigate existing problems in the literature. We attempt to accomplish this by adapting the rules of inference used in the social and natural sciences to the special needs, theories, and data in legal scholarship, and explicating them with extensive illustrations from existing research.[26] In so doing, we hope to speak to various constituencies within the community. Beginning with law professors, given their propensity to draw inferences about the real world, it is clear that they have a strong interest in learning how to conduct empirical research properly. But it is equally clear that many have not turned that interest into productive research practices. We believe the rules and guidelines we set out can begin to help everyone do so, and, in the longer term, improve the quality of legal scholarship,[27] thereby enabling law professors to contribute to a credible, valid, common, and, ultimately, more valuable research enterprise.

At the same time, we want to encourage greater self-conscious attention to methodology in legal studies. The law is important enough to have a subfield devoted to methodological concerns, just 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 problems that law professors have. Unfortunately, the complete list of all law review articles devoted exclusively to improving, understanding, explicating, or adapting the rules of inference is as follows: none. Many apply some of the rules of inference in the context of their specific research projects, but there appear to be no methods articles in the field at all. Thus, in addition to perusing this Article, we hope others will take up the challenge of attempting to explain, adapt, and extend the rules of inference in legal scholarship, and to write other articles about the methodology of empirical analysis in this field.

---

[25] Many scholars already have evaluated the state of legal scholarship, and have done so from a variety of perspectives. For recent examples, see Edward L. Rubin, "The Practice and Discourse of Legal Scholarship," 86 Mich L Rev 1835 (1988); Posner, "Madison Lecture: Against Constitutional Theory," supra note 11; Harry T. Edwards, "The Growing Disjunction between Legal Education and the Legal Profession," 91 Mich L Rev 34 (1992); Lawrence M. Friedman, "Law Reviews and Legal Scholarship: Some Comments," 75 Denver U L Rev 661 (1998); Richard S. Harnsberger, "Reflections about Law Reviews and American Legal Scholarship," 76 Neb L Rev 681 (1997); Daniel A. Farber & Suzanna Sherry, Beyond All Reason: The Radical Assault on Truth in American Law (1997); Ronald J. Krotoszynski, "Legal Scholarship at the Crossroads: On Farce, Tragedy, and Redemption," 77 Tex L Rev 321 (1998).

[26] In so doing, we draw on King, Keohane, & Verba, Designing Social Inquiry (1994), supra note 1, and related works, e.g., Mark Lichbach, ed., "The Qualitative-Quantitative Disputation," 89 Am Pol Sci Rev 2 (1995); Henry E. Brady & David Collier, eds., Rethinking Social Inquiry: Diverse Tools, Shared Standards (2001); and Mark Lichbach & Ned Lebow, Theory and Evidence in International Relations (2001). Our purpose is to adapt, elaborate, and expand on these works (and the rules they contain) in ways that seem especially productive for research in law.

[27] At the very least law professors, judges, and lawyers will now have something on which to hang their hats other than the problematic sources cited in so many law review articles. See, for example, Orin S Kerr, "Shedding Light on Chevron: An Empirical Study of the Chevron Doctrine in the U.S. Courts of Appeals," 15 Yale J Reg 1 (1998) (relying on a 1955 statistics book, R. Clay Sprowls, Elementary Statistics); Leandra Lederman, "Which Cases Go to Trial? An Empirical Study of the Predictors of Failure to Settle," 49 Case Western Reserve 315 (1999) (justifying the use of (the highly unwise) stepwise regression—see Gary King, "How Not to Lie with Statistics," 30 Am J Pol Sci (1986)— on the basis of claims made in Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers (1990) about the importance of "parsimony"—a widely-misunderstood concept—see King, Keohane, & Verba, Designing Social Inquiry (1994), supra note 1); Gregory C. Sisk, Michael Heise, & Andrew P. Morriss, "Charting the Influences on the Judicial Mind: An Empirical Study of Judicial Reasoning," 73 NYU L Rev 1377 (1998) (citing George W. Bohrnstedt & David Knoke, Statistics for Social Data Analysis (1988), among others, to support their (incorrect) understanding of multicollinearity and their concomitant flawed decision to exclude an independent variable if it and other another variable(s) correlate at the .5 level or higher).

We also hope the rules we offer will prove useful to members of the legal community (*e.g.,* attorneys and judges) who do not necessarily conduct research but certainly consume it.[28] In so doing, it is critical that they are able to distinguish between work that is "poor" and that which is not.[29] This is especially so for trial judges, who, under *Daubert v. Merrell Dow Pharmaceuticals*,[30] can no longer exclusively rely on a consensus in the scientific community to evaluate the quality of research presented by experts in their courtrooms; they are now also required to judge the research *themselves*, to evaluate its credibility, to assess its methods, and to appraise its design.[31] And while there are signs

---

[28] Scores of studies document the use judges and lawyers make of empirical evidence—some of which lawyers (and amici curiae) amass independently, some of which judges and attorneys take from law review articles and other scholarly work. See, for example, Rustad & Koenig, "The Supreme Court and Junk Science: Selective Distortion in Amicus Briefs," supra note 11; David Kaye, "Statistics for Lawyers and Law Statistics," 89 Mich L Rev 1520 (1991); John Monahan & Laurens Walker, "Obtaining, Evaluating, and Establishing Social Science in Law," 134 U Pa L Rev 477 (1986); Rosen, The Supreme Court and Social Science supra note 11; Ronald Roesch, "Social Science and the Courts: The Role of Amicus Curiae Briefs," 15 L & Human Beh 1 (1991); Michael Saks and Charles H. Baron, eds., The Use, Nonuse, and Misuse of Applied Social Research in the Courts (1980); Erickson and Simon, The Use of Social Science Data in Supreme Court Decisions, supra note 11. In fact, as Rustad & Koenig, "The Supreme Court and Junk Science: Selective Distortion in Amicus Briefs," at 94, note: "Once heretical, the belief that empirical studies can influence the content of legal doctrine is now one of the few points of general agreement among jurists."

Even more interesting are claims that judges and lawyers would make greater use of research grounded in real-world observations if more studies existed or if they were of a higher quality. These come from legal academics, for example, David Faigman, "Normative Constitutional Fact-Finding: Exploring the Empirical Component of Constitutional Interpretation," 139 U Pa L Rev 541 (1991), as well as from jurists. See, for example, Chandler v. Florida, 449 U.S. 560, 578 (1981) (lamenting that "at present no one has been able to present empirical data sufficient to establish that the mere presence of the broadcast media inherently has an adverse effect" on the administration of justice).

We are heartened by these claims but want to emphasize that if courts make use of data resulting from improperly-conducted studies (*i.e.,* a healthy portion of the work published in the law reviews; see infra Part I), they can open themselves up to severe criticism. See Rustad and Koenig, "The Supreme Court and Junk Science: Selective Distortion in Amicus Briefs"; David M. O'Brien, "The Seduction of the Judiciary: Social Science and the Courts," 64 Judicature 8 (1980); Saks & Baron, eds., The Use, Nonuse, and Misuse of Applied Social Research in the Courts . For critiques of particular decisions, see, for example, Saks, "Ignorance of Science is No Excuse," supra note 20; Zeisel & Diamond, "Convincing Empirical Evidence on the Six Member Jury," supra note 20; Miller, "Six of One is Not a Dozen of the Other: A Reexamination of Williams v. Florida and the Size of State Criminal Juries," supra note 20.

[29] See Saks, "Ignorance of Science is No Excuse," supra note 19, at 18.

[30] 509 U.S. 579 (1993).

[31] Moreover, in *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999), the U.S. Supreme Court extended the trial judge's obligation under *Daubert* to cover virtually all proffers of testimony based on specialized knowledge.

Judge Alex Kozinski's opinion in *Daubert v. Merrell Dow Pharmaceuticals,* 43 F.3d 1311 (1995), at 1315-1316, highlights the "uncomfortable position" in which (at least some) judges now find themselves when called to evaluate research:

> Federal judges ruling on the admissibility of expert scientific testimony face a far more complex and daunting task in a post-*Daubert* world than before… Under *Daubert,* we must engage in a difficult, two-part analysis. First, we must determine nothing less than whether the experts' testimony reflects "scientific knowledge," whether their findings are "derived by the scientific method," and whether their work product amounts to "good science." Second, we must ensure that the proposed expert testimony is "relevant to the task at hand," i.e., that it logically advances a material aspect of the proposing party's case. The Supreme Court referred to this second prong of the analysis as the "fit" requirement.
>
> The first prong of *Daubert* puts federal judges in an uncomfortable position. The question of admissibility only arises if it is first established that the individuals whose testimony is being proffered are experts in a particular scientific field; here, for example, the Supreme Court waxed eloquent on the impressive qualifications of plaintiffs' experts. Yet something doesn't become "scientific knowledge" just because it's uttered by a scientist; nor can an expert's self-serving assertion that his conclusions were "derived by the scientific method" be deemed conclusive, else the Supreme Court's opinion could have ended with footnote two. As we read the Supreme Court's teaching in *Daubert,* therefore, though we are largely untrained in science and certainly no match for any of the witnesses whose testimony we are reviewing, it is our responsibility to determine whether those experts' proposed testimony amounts to "scientific knowledge," constitutes "good science," and was "derived by the scientific method."
>
> The task before us is more daunting still when the dispute concerns matters at the very cutting edge of scientific research, where fact meets theory and certainty dissolves into probability. As the record in this case illustrates, scientists often have vigorous and

that they are *attempting* to perform these crucial tasks—a 1991 Federal Judicial Center survey revealed that only 25% of judges claimed that they had "screened out" flawed expert evidence in their most recent trial; by 1998, that figure was 41%[32]—we know of no law review article that helps them to do so, that is, to differentiate the good from the bad and to evaluate the degree of certainty to accord different research practices.[33] This Article provides a first cut at enhancing their ability to make that distinction and perform the requisite evaluation.[34]

To accomplish these various ends, we proceed in three steps. First, because we acknowledge that our rules will only help empirical researchers who are unaware of or otherwise fail to follow them, we begin, in Part I, with a brief report on our survey of methodological problems in the nations' law reviews.

Next, in Parts II through VIII we devote considerable attention to clarifying the rules of inference and their application to legal inquiry. We do so out of the belief that legal scholars conducting empirical research need to understand, analyze, use, and, where necessary, learn to extend the guidelines we offer. These rules must not be merely framed and hung on a wall or trotted out in footnotes; they need to be read, taught, internalized, and put to use whenever and wherever legal scholarship is conducted. Part II considers the goals of empirical research: amassing and summarizing data and making descriptive and causal inferences. After outlining the research design process in Part III, we articulate rules that govern the components of that process: Research Questions (Part IV), Theories and their Observable Implications (Part V), Rival Hypotheses (Part VI), Measurement and Estimation (Part VII), and Selecting Observations (Part VIII). Although no certain path to valid inference exists, or even could exist in principle, adhering to these rules should greatly increase the validity and value of empirical legal scholarship.

We conclude with suggestions on how the legal community can reorganize the conduct of inquiry and teaching so that it can better support the creation of first-rate research without compromising its other objectives, such as training lawyers. The opportunities here are extremely promising and very exciting, perhaps even more so than in many of the more advanced areas of academic research. We expand on this point in some detail in Part IX, as well as offer a set of

---

sincere disagreements as to what research methodology is proper, what should be accepted as sufficient proof for the existence of a "fact," and whether information derived by a particular method can tell us anything useful about the subject under study.

Our responsibility, then, unless we badly misread the Supreme Court's opinion, is to resolve disputes among respected, well-credentialed scientists about matters squarely within their expertise, in areas where there is no scientific consensus as to what is and what is not "good science," and occasionally to reject such expert testimony because it was not "derived by the scientific method." Mindful of our position in the hierarchy of the federal judiciary, we take a deep breath and proceed with this heady task.

[32] Reported in "Judges Shunning Bad Science," *Science* 929 (3 Nov 2000). The original study can be found at /air.fjc.gov/public/fjcweb.nsf/pages/336.

[33] There are, of course, some textbooks and manuals that address related, but not identical, issues. See, for example, Steven M. Crafton and Margaret F. Brinig, Quantitative Methods for Lawyers (1994), Wayne C. Curtis, Statistical Concepts for Attorneys: A Reference Guide (1983), Finkelstein and Levin, Statistics for Lawyers, supra note 27; Hans Zeisel and David Kaye, Prove it with Figures: Empirical Methods in Law and Litigation (1997); Federal Judicial Center, Reference Manual on Scientific Evidence (2000), available at //air.fjc.gov/public/fjcweb.nsf/pages/16. These works explain various statistical procedures and discuss (some of the) rules of statistical inference but, unlike this Article, they are not typically geared toward legal scholarship or concerned with inference using qualitative evidence, research design, or research infrastructure.

[34] That judges (and we might add, lawyers and law professors) have difficulty making that distinction has been documented by numerous scholars. See supra note 20. See also Rustad & Koenig, "The Supreme Court and Junk Science: Selective Distortion in Amicus Briefs," supra note 11, at 152 (claiming that "Justices are not immune from the human tendency to 'overvalue vivid anecdotes when making important decisions'."); Faigman, "Normative Constitutional Fact-Finding," supra note 28, at 544 (asserting that "The Court often fails to distinguish between normative principles and empirical propositions, analyzing empirical research as it might arguments about text or precedent.").

recommendations regarding the creation of a serious infrastructure for legal research. These include fostering the development of skill sets necessary to conduct "real world" inquiries, moving to an alternative model of journal management, devising standards for data archiving, and several others that speak to the future of scholarship in the legal academy.

I. The Extent of Methodological Problems in Empirical Research

If all scholars who conduct empirical legal research followed the rules of inference, then self-conscious attention them would be unnecessary. Accordingly, in deciding whether to write this Article, we conducted our own survey of current practices in the nation's law reviews. Our purpose was to identify where the problems were. Our answer? So far as we can tell, everywhere. Everywhere we find empirical research in the law reviews, in articles written by members of the legal community, serious problems of inference and methodology abound.

We now report on how we conducted our survey, but readers should be aware that our goal is to improve legal research, not to excoriate selected individual authors for sins committed by everyone. A blood sport may be fun to watch, but it would serve no useful purpose here and hardly seems fair. In the balance of this Article, we criticize numerous individual studies but we do so only when it helps us make a point that would benefit others. Usually this means illustrating a rule of inference and how its violation could be corrected. Indeed, the optimal article for our purposes here is not one that violates each and every rule of inference; that would produce a mess, require long qualifications, and be almost useless from an expository perspective. The best example from our perspective is an article that is perfect in nearly every respect save one—the one that illustrates the rule we are explicating. That way we can demonstrate most cleanly and clearly the advantages of following a particular rule of inference. Thus, readers should understand that most of the works we discuss explicitly in the balance of this Article are well above average, even though they fail in the one area we delineate and regardless of whether they were penned by the holder of an endowed chair or an entry-level professor or appeared in the flagship journal of the Nation's #1-ranked law school or in a more specialized outlet produced by a fourth-tier institution. In this brief section, we summarize our survey but do so without explicit references.[35] Finding examples for a methods piece like this is difficult: there is much good work in the law that follows each of the specific rules we delineate, and a lot of work that violates a complex combination of different rules.

Our goal was not to assess the average law review article or to characterize all legal scholarship: this might be useful, but for present purposes we only needed to ascertain whether there existed sufficient problems to warrant an explicit discussion of methods in a law review article. We thus began by casting the net very widely, reading all 231 articles, published in *all* U.S. law reviews between 1990-2000, that had the word "empirical" in their title.[36] We inventoried these articles because, by virtue of their titles, they at least claimed to be conducting research based on real-world observations. "Empirical research" apparently has become a term of art in legal scholarship, and many of those using it in their titles appear to be intentionally identifying their work with this movement. We have since been supplementing this search strategy with a narrower one intended to uncover and evaluate some of the best in empirical legal research. This search, still in progress,

[35] We also do not tabulate the particular types of mistakes we identified. We tried to develop coding schemes to do so, but most articles make such a complicated and interrelated set of mistakes that we could not find a way to quantify errors without doing injustice to the original article. Of course, tabulations like these would serve no material purpose: Errors are everywhere, every rule we discuss below is violated numerous times in the law review literature, and almost every one can be corrected.

[36] We amassed these from a LEXIS search, conducted on July 17, 2000, of title [empirical] in the LAWREV library, ALLREV file; we excluded articles not intended to be empirical, such as reviews.

includes all empirical articles from six top law reviews (*Chicago, Columbia Harvard, NYU, Stanford, and Yale*) published between 1995 and 2000. It also includes the top fifty cited articles (according the Legal Resource Network) that were written by legal academics and appeared in the law reviews.[37] We added to these formal lists via a much more informal approach; namely, by reading widely through law reviews, following citations, and reading further. When legal academics learned we were working on this project, many were kind enough to send us their empirical work or to refer us to others, and we read these as well. Finally, we examined studies in four peer-reviewed law journals—*Journal of Law and Economics*; *Journal of Law, Economics, & Organization*; *Journal of Legal Studies*, and *Law & Society Review*—even though social scientists and business school faculty authored most of the articles in them, not members of the legal community who constitute the primary audience for this Article.[38]

We have obviously not evaluated anything close to all empirical research in the law, but we have searched extensively in something very roughly approximating a representative sample of all empirical research in the law reviews. We also focused deeply in several ways in areas where quality should be high, so much so that, for this sample, any conclusions we draw should be biased against a finding of methodological problems. Nonetheless, our results are discouraging. While it is certainly true that some articles in the law reviews are better than others, and some meet the rules of inference better than others, every one we have read thus far—every single one—violates at least one of the rules we discuss in the balance of this Article.[39] Since all—every single one—have the potential to find their way into a court case, an administrative proceeding, or a legislative hearing we can only imagine the serious consequences for public policy (not to mention for the development of knowledge) that may have already resulted—or still may result.[40]

---

[37] See http://www.ssrn.com.

[38] Specifically, of the 72 research articles published in these journals (in numbers 1 and 2 of volumes published in 2000; for the *Journal of Legal Students*, parts 1 of numbers 1 and 2) only 18% (n=13) were authored by a law professor or a team of law professors; the remaining 59 were authored (or coauthored) by at least one scholar outside the legal academy. (Our count of research articles excludes exchanges, symposium pieces, review essays, and book reviews.) The *Journal of Legal Studies,* housed at the University of Chicago, accounts for the lion's share of articles produced by legal academics, 10 of the 13, but 4 of the 10 were authored by members of Chicago's faculty.

[39] For those who have already read Parts III-VIII, we clarify that the target of our inference here is the extent of methodological problems in future legal literature if this Article were not read. In other fields, methodological sophistication changes gradually, and we see no reason for major differences in law. We therefore seem safe in our assumption that recent methodological practices will accurately forecast future practices, and so we focus on the task of assessing the recent past. If we take the sample of all articles with "empirical" in their titles as a rough approximation to a random sample of all empirical articles in law reviews, then we can be very confident that the population of empirical articles is similarly problematic. For example, the probability of observing a random sample of 231 articles in which 100% violate at least one of the rules of inference, given a target population with only 90% of articles violating at least one rule, is astronomically small (fewer than 1 in 37 billion). This result, and our supplementary samples biased in favor of better research, makes it seems highly unlikely that even as many as 10% of law review articles in the near future will not violate the rules of inference (at least without further attention to these rules).

[40] The staying power of flawed and discredited legal studies can be extraordinary. For example, as Tabak, "Empirical Studies of the Modern Capital Sentencing System: How Empirical Studies Can Affect Positively the Politics of the Death Penalty," supra note 12, at 1431 points out "Scholars conducting valid studies on the subject of deterrence have failed to find any deterrent effect from capital punishment…Yet in the political discourse, the proponents of the death penalty often claim the contrary—that a deterrent effect exists. They still cite studies done by Isaac Ehrlich and his student, Stephen Layson." This is so even though a panel appointed by the National Academy of Sciences, along with many other scholars, have condemned their work." See, for example, Lawrence R. et al. Klein, "The Deterrent Effect of Capital Punishment: An Assessment of the Estimates," in *Deterrence and Incapacitation*, ed. Alfred et al. Blumstein (1978).

We are aware that there is an econometrics literature that pursues the question of deterrence with indeterminant results. The only point here is that scholars continue to cite the Ehrlich study, even though higher-quality studies—reaching the same and different results—exist. Legal academics have made a similar point about recent scholarship on

In writing this, we do not mean to suggest that empirical research appearing in law reviews is always or even usually worse than articles in the journals of other scholarly disciplines. Such comparisons are irrelevant. Our only point is that our survey indicates that many legal scholars conducting empirical research—which, under a standard definition of "empirical," means virtually all legal scholars—would profit from a greater familiarity with the rules of inference. Other fields seem to benefit from devoting sustained attention to their methodological problems in article explicitly and exclusively devoted to methods, and we have uncovered no reason why the legal academy should not follow suit.

This conclusion and the results of our survey more generally will probably come as no surprise to many law professors whose own evaluation of legal scholarship has been, if anything, more negative and extreme than ours. Indeed, there seems to be a long tradition of legal academics denigrating articles published in their journals.[41] Over the years, they have referred to the content in them as "Junk" and "Junk Stream,"[42] "manure,"[43] not "readable by humans,"[44] "turgid legaldegooky garbage,"[45] "junk science,"[46] "spinach,"[47] "boring, too long,"[48] including "assertions unconnected to an empirical basis,"[49] relying only on "anecdotes,"[50] "opaque," [51] and an "open scandal."[52]

---

the impact of *Miranda v. Arizona,* 384 U.S. 436 (1966), that relies on older studies to reach conclusions about the decision's impact. See, for example, Stephen J. Schulhofer, "Miranda's Practical Effect: Substantial Benefits and Vanishingly Small Social Costs," 90 Northwestern Law Review 500 (1996), at 506-507 (quoting Richard Angelo Leo, Police Interrogation in America (1994), Ph.D. dissertation, University of California at Berkeley), "virtually all [the studies] were conducted by lawyers or law professors not trained in the research methods of social science [and they] are replete with methodological weaknesses…The methodological weaknesses of virtually all of the *Miranda* impact studies should necessarily temper, and in some instances should cause us to question, the conclusions that have been drawn from these studies."

[41] Among the most famous attacks is Fred Rodell, "Goodbye to Law Reviews," 23 Va L Rev 38 (1936), in which he claimed that were two problems with legal writing: "One is its style. The other is its content." Rodell's article may be the most well known—indeed, and ironically enough, at least according to Bernard J. Hibbitts, "Last Writes? Reassessing the Law Review in the Age of Cyberspace," 71 NYU L Rev 615 (1996), Rodell's work would "ultimately…become the most-cited law review article on law reviews"—but many others express similar sentiments. For a review of their concerns, see Hibbitts, "Last Writes? Reassessing the Law Review in the Age of Cyberspace."

[42] Kenneth Lasson, "Scholarship Amok: Excesses in the Pursuit of Truth and Tenure," 103 Harv L Rev 926, 928 (1990).

[43] Id., at 931.

[44] James Lindgren, "An Author's Manifesto," 61 U Chi L Rev 527, 531 (1994).

[45] Rodell, "Goodbye to Law Reviews," supra note 41.

[46] Peter Huber, Galileo's Revenge: Junk Science in the Courtroom (1991) uses the term "junk science" to refer to suspect expert testimony at trials. But see Kenneth J. Chesebro, "Galileo's Retort: Peter Huber's Junk Scholarship," 42 Am U L Rev 1637 (1993), who calls Huber's work "junk scholarship" and writes: "Galileo…would quickly become exasperated at the unsupported thesis of Huber's book, its numerous material misrepresentations and omissions, and its manipulative and evasive method of argument. Galileo would find Huber's criticism of purported errors of scholarship by others to be hypocritical, as Huber himself repeatedly violates the standards he holds out for the world at large. After full review, Galileo would not ratify the message of Galileo's Revenge." Rustad & Koenig, "The Supreme Court and Junk Science: Selective Distortion in Amicus Briefs," supra note 11, at 98-99, invokes "junk science" to describe the contents of some amicus curiae briefs. According to Rustad & Koenig amicus curiae briefs that present empirical evidence often do not meet Huber's standard for good science—"the science of publication, replication, and verification, the science of consensus and peer review." We believe the same is true of a healthy portion of legal scholarship published in the law reviews.

[47] Rodell, "Goodbye to Law Reviews," supra note 41.

[48] Elyce Zenoff, "I Have Seen to the Enemy and They Are Us," 36 J Legal Educ 21 (1986).

[49] Heise, "The Importance of Being Empirical," supra note 2, at 808.

[50] Michael Saks, "Do We Really Know Anything about the Behavior of the Tort Litigation System—and Why Not?," 140 Pennsylvania Law Review 1147 (1992), at 1159 puts it this way with regard to the use of anecdotal evidence in legal scholarship on the tort litigation system: "[A]necdotal evidence is heavily discounted in most fields, and for a perfectly good reason: such evidence permits only the loosest and weakest of inferences about matters a field is trying to

Fortunately, at least for future research, remedying the problem with empirical work in law reviews will not require as drastic or dramatic measures as this litany of complaints suggests—only greater attention and adherence to the rules of inference we begin to articulate in the next Part.

II. Common Features of Empirical Research

Virtually all good empirical research shares two features. First, the researcher typically has a specific goal(s) in mind—such as collecting data or making inferences. And second, no matter what the specific goal might be, she or he must follow some general rules to arrive there—or at least arrive there with some known degree of confidence.

In what follows, we consider these features in some detail. We begin with a discussion of the goals of empirical research and move to the guidelines to which all researchers seeking to achieve one or more of these goals ought adhere.

A. The Goals

Regardless of the type of data employed, all empirical research seeks to accomplish one of three ends, or (more typically) some combination thereof: *amassing data* for use by the researcher or others; *summarizing data* so they are easier to comprehend; and *making descriptive or causal inferences,* which entails using data we have observed to learn about data we would like to gather.

1. Amassing Data

The legal world produces enormous quantities of data every day. Hundreds of court cases are litigated or settled; scores of decisions are handed down; dozens of rulings are implemented (or not); handfuls are reviewed by legislatures and other policy-making bodies. And members of the legal community—perhaps to a greater extent than most others—have been instrumental in ensuring that records of these events exist over long periods of time. To provide just one illustration, we have the exact text of U.S. Supreme Court opinions back to the 1790s, but, for comparison, precinct-level presidential election results do not exist nationally before 1984.[53]

---

understand. Anecdotes do not permit one to determine either the frequency of occurrence of something or its causes and effects. They do no better in enlightening us about the behavior of the tort litigation system." See also Mashaw & Harfst, "Regulation and Legal Culture: The Case of Motor Vehicle Safety," supra note 13.

[51] Posner, "Madison Lecture: Against Constitutional Theory," supra note 11 at 3, which is devoted to making the point that "constitutional theory is not responsive to, and indeed tends to occlude, the greatest need of constitutional adjudicators, which is the need for empirical knowledge." "Constitutional theory, he continues, "today circulates in a medium that is largely opaque to the judge and practicing lawyer." Posner (at 11) "would like to see an entirely different kind of constitutional theorizing"—one based largely on answering questions relating to the effect of law and legal decisions. Edwards, "The Growing Disjunction between Legal Education and the Legal Profession," supra note 25 at 34, agrees that legal academics have overemphasized "abstract theory" but has different ideas of the direction into which they should head. He makes a case for "practical" legal scholarship—scholarship that is "prescriptive" and "doctrinal."

The difference between Posner's and Edward's positions may be wide on many dimensions but it is not particularly relevant to our goal of improving inquiry based on real-world observations. For, regardless of whether scholarship focuses on, say, the products of legal decisions (doctrine) or their effect, it should be conducted in accord with the rules we outline.

[52] John Henry Schlegel, "Searching for Archimedes—Legal Education, Legal Scholarship, and Liberal Ideology," 34 J Legal Educ 103 (1984).

[53] Gary King & Bradley Palmquist, "The Record of American Democracy, 1984 and 1990," 26 Sociological Methods and Research 424 (1998).

Yet, merely preserving records, albeit important, is not typically a separate goal associated with empirical scholarship; that goal is, rather, to translate or amass information in such a way that researchers can make use of them. Consider an investigation by Liebman and his colleagues that seeks to assess, via an examination of state and federal appellate court decisions, the extent of errors in capital sentencing.[54] Fortunately for the Liebman team, published decisions in these cases are available from any number of sources, *e.g.,* F. 3d, S.E. 2d. Less fortunately, these decisions are just that: judicial opinions handed down in capital cases without any attempt to systematize various features of interest. That is the important task Liebman et al. undertake: They use information in the federal and state reporters to create a data base.[55] So, for example, while they include the name of the individual whose capital judgment was under review (as does, say, F 3d.), they also characterize details about the victim and the basis for the court's decision (if it reversed),[56] among other systematically collected attributes.

In Liebman's case, amassing the data was a, but not the ultimate, goal of the research; he and his colleagues are also interested in summarizing the data in much the way we describe in the next section. To the extent that most researchers do not collect data for the sake of collecting data but have other goals in mind (summarizing or making inferences) Liebman is not atypical. But there are exceptions, most notably the so-called "multi-user" or "public-use" databases. The idea behind these is straightforward enough: Rather than collect data designed to answer particular research questions—*e.g.,* In how many capital cases are errors made?—amass large data bases so rich in content that multiple users, even those with distinct projects, can draw on them.

In addition to opening access to many researchers, large public-use databases have what is known a *combinatoric* advantage. To see this, consider one useful method of analyzing data—the *crosstabulation* or a table that summarizes information about various quantities of interest. For example, a crosstabulation of U.S. Supreme Court action over a sample of certiorari (cert.) petitions (cert. granted or not) by the U.S. government's involvement in those petitions (the petitioner or not), as Table 1 shows, produces four "cell" values: the number of cert. petitions the Court granted when the federal government was the petitioner (26) and the number it denied (0), and the number it granted when the government was not the petitioner (85) and the number it denied (1078). This is in addition to information about each variable taken separately (*e.g.,* the fraction of petitions in which the government was the petitioner [26/1189] and of those granted cert. [111/1189]).

Table 1. Crosstabulation of U.S. Supreme Court Action over Petitions for Certiorari by U.S. Government Involvement in the Petitions[57]

		Did the Court Grant Cert.?		
		Yes	***No***	**Total**
Was the Government the Petitioner?	***Yes***	26	0	26

[54] James S. Liebman et al., "Capital Attrition: Error Rates in Capital Cases," Tex L Rev (2000).

[55] Actually, they created two data bases, the Direct Appeal Database (DADB) and the Habeas Corpus Database (HCDB). DADB houses "the name of the individual whose capital judgment was under review; the sentencing state; the year, outcome, citation, and subsequent judicial history (rehearing, certiorari) of the decision finally resolving the appeal; and information about the basis for reversal if a reversal occurred." HCDB contains "the name of the individual whose capital judgment was under review; the sentencing state; the timing of the habeas petition and its adjudication at the various stages; the outcome at the various stages; information about the petitioner, lawyers, judges, courts, victim, and offense; the aggravating and mitigating circumstances found at trial; procedures used during the habeas review process; and the asserted and the judicially accepted bases for and defenses to habeas relief." Id.

[56] Id.

[57] Data Source: H.W. Perry, Deciding to Decide (1991), at 136. Perry drew these from a "random sample" of cert. petitions.

	No	85	1078	1163
Total		111	1078	1189

The relevant point is that two factors, U.S. Supreme Court Action and U.S. Government Involvement, each with two categories, produce information on (2×2=) 4 cells. Now consider a situation in which two teams of researchers want to study the factors that lead the Court to grant cert. One is mainly interested in assessing the effect of the federal government's role as a petitioner and the other, in examining the impact of conflict in the U.S. Courts of Appeals over the answer to the question at issue in the cert. petition ("intercircuit conflict").[58] Suppose, for the sake of simplicity, that each of these factors has two categories (whether the government was the petitioner or not; whether intercircuit conflict existed or not). Suppose further that the teams are not working together but rather draw two independent samples of cert. petitions, as well as collect data on the factor of specific interest (federal government involvement for Team 1 and intercircuit conflict for Team 2) and four other (non-overlapping) factors (also with two categories each) that they also suspect affect the Court's decision to grant cert.[59] The result is that crosstabulations for each study could produce 2^5, or 32, cells. Note, though, that since the researchers drew their samples of cert. petitions independently, crosstabulations of the factors in one sample with the factors in the other would not be possible. Now consider what would happen if the two teams combined forces and collected the same ten factors on the same set of petitions in one large data set. The combinatoric advantage accrues: If the ten factors were collected together on the same petitions, the ten factors of two categories each would generate 2^{10} (1,024) different cells—or ($2^{10}/(2\times2^5)$=) 16 times as much information as the two data bases produced separately.[60]

Aware of these advantages, social scientist Harold J. Spaeth, nearly two decades ago, asked the National Science Foundation (NSF) to fund a multi-user data base on the U.S. Supreme Court, one that would contain scores of attributes of Court decisions, handed down since 1953, ranging from the date of the oral argument to the identities of the parties to the litigation to how the justices

[58] Scholars suggest that these factors are related to the Court's decision over cert. Specifically, when the U.S. government is the petitioner, they argue that the Court is more likely to grant cert. than when, say, a private party is the petitioner. See, for example, Gregory A. Caldeira & John R. Wright, "Organized Interests and Agenda Setting in the U.S. Supreme Court," 82 Am Pol Sci Rev 1109 (1988); Joseph Tanenhaus, et al., The Supreme Court's Certiorari Jurisdiction: Cue Theory, in Judicial Decision Making, ed. Glendon Schubert (1963).; Virginia Armstrong & Charles A. Johnson, "Certiorari Decisions by the Warren and Burger Courts: Is Cue Theory Time Bound?," 15 Polity 141 (1982). Likewise, they suggest that the presence of intercircuit conflict increases the odds that the Court will grant cert. See, for example, Robert M. Lawless & Dylan Lager Murray, "An Empirical Analysis of Bankruptcy Certiorari," 62 Mo L Rev 101 (1997); S. Sidney Ulmer, "The Supreme Court's Certiorari Decisions: Conflict as a Predictive Variable," 78 Am Pol Sci Rev 901 (1984); Perry, Deciding to Decide, supra note 57; Caldeira & Wright, Organized Interests and Agenda Setting in the U.S. Supreme Court; Doris Marie Provine, Case Selection in the United States Supreme Court (1980).

[59] Examples of other factors include: whether the U.S. Solicitor files an amicus curiae brief in support of cert; whether a judge(s), in the court that handed down the decision at issue in the cert. petition, dissented; and whether the case involved a civil liberties issue. Scholars posit that some or all of these may all effect the Court's decision to grant cert. or not. See, for example, Provine, Case Selection in the United States Supreme Court, supra note 58; Caldeira & Wright, "Organized Interests and Agenda Setting in the U.S. Supreme Court," supra note 58; Gregory A. Caldeira, et al., Sophisticated Voting and Gate-Keeping in the Supreme Court, 15 Journal of Law, Economics, & Organization 549 (1999).

[60] This "combinatoric" advantage is only one of many that accrue when scholars cooperate and work to foster a vibrant scholarly community in their attempt to build knowledge—a topic we revisit and an effort we encourage throughout this Article. See especially infra p. 32 and Part IX.

voted.[61] With support from the NSF, along with guidance from a board of overseers, Spaeth went about the task of collecting and coding the data and, finally, assembling the data base. In the late 1980s, he made it (and the documentation necessary to use it) publicly available; since then, he has updated the data annually. He also has backdated the data to cover the Vinson Court era (1946-1952 terms).

As one might expect, Spaeth has made great use of his database to answer his specific research questions,[62] but a multitude of other scholars have used it to study questions of their own as well.[63] If the Liebman team makes its database publicly available, even though it was originally designed primarily for their research, scholars probably will take advantage of its labors just as they have done of Spaeth's. Seen in this way, the Spaeth, Liebman, and many other data-collection efforts are important contributions to the scholarly community in their own right. We should fully recognize them as such.

There is yet another similarity between the Liebman et al. and Spaeth data sets. In both instances, the creators obtained their data from public or other readily available sources: in Liebman et al.'s case, from state and federal reporters and in Spaeth's, from *U.S. Reports*. This is quite common, but it is not the only way that scholars—especially social scientists—approach data collection. Indeed, many studies on, say, voting behavior rely on data created by the investigator rather than on data the investigator obtains from other sources. These data may come from surveys, interviews, or experiments.

Regardless of the source, evaluating data-collection efforts depends largely on the purpose the researcher has in mind. For, as we suggested earlier, scholars often do not view amassing data as an end in and of itself. Nonetheless, whatever the goal, some fairly basic rules apply. First, the process by which the data came to be observed must be fully recorded. This is the scientific equivalent of insisting in court that the "chain of evidence" be fully documented and unbroken. Second, the more data, the better. In almost any conceivable empirical usage more data will not hurt the researcher's goals. We elaborate and explain the importance of both rules in Parts II and VII.

---

[61] For a full list of attributes (variables) in the Spaeth data base, and the data themselves, navigate to: http://www.ssc.msu.edu/~pls/pljp/sctdata.html. Information about the data base is in Harold J. Spaeth & Jeffrey A. Segal, "The U.S. Supreme Court Judicial Data Base: Providing New Insights in the Court," 83 Judicature 228 (2000).

[62] See, for example, Harold J. Spaeth & Jeffrey A. Segal, Majority Rule or Minority Will : Adherence to Precedent on the U.S. Supreme Court (1999); Jeffrey A Segal & Harold J. Spaeth, The Supreme Court and the Attitudinal Model (1993); Jeffrey A. Segal et al., "Ideological Values and the Votes of U.S. Supreme Court Justices Revisited," 57 J Politics 812 (1995); Jeffrey A. Segal & Harold J. Spaeth, "The Influence of Stare Decisis on the Vote of United States Supreme Court Justices," 40 Am J Pol Sci 971 (1996); Lee Epstein, et al. The Supreme Court Compendium: Data, Decisions, and Developments (2001).

[63] See, for example, Jeffrey A Segal, "Separation-of-Powers Games in the Positive Theory of Law and Courts," 91 Am Pol Sci Rev 28 (1997); Robert L. Boucher, Jr. & Jeffrey A. Segal, "Supreme Court Justices as Strategic Decision-Makers: Aggressive Grants and Defensive Denials on the Vinson Court," 57 J Politics 812 (1995); Lee Epstein & Jack Knight, The Choices Justices Make (1998); Lee Epstein & Jeffrey A. Segal, "Measuring Issue Salience," 44 Am J Pol Sci 66 (2000); Caldeira, et al., Sophisticated Voting and Gate-Keeping in the Supreme Court, supra note 59; Kevin T. McGuire, "Repeat Players in the Supreme Court: The Role of Experienced Lawyers in Litigation Success," 57 J Politics 187 (1995); James Meernik & Joseph Ignagni, "Judicial Review and Coordinate Construction of the Constitution," 41 Am J Pol Sci 447 (1997); Richard C. Kearney & Reginald S. Sheehan, "Supreme Court Decision Making: The Impact of Court Composition on State and Local Government Litigation," 54 J Politics 1008 (1992).

2. Summarizing Data

In his study of consumer bankruptcy, Norberg raises two empirical questions: How do creditors fare, and what factors account for their success or failure?[64] To address these, Norberg begins by summarizing data he collected on each of seventy-one chapter 13 cases filed in the U.S. Bankruptcy Court for the Southern District of Mississippi between 1992 and 1998. Table 2 below, which details creditor collections per case by the type of claim, is exemplary of the tack he took.

Table 2. Creditor Collections Per Case, as Reported in Norberg[65]

Nature of the Claim	Range	Mean	Median	Standard Deviation
Secured	$0 - 66,183	$9,313	$3,914	$13,766
Priority	0 - 4,965	193	0	713
General	0 - 7,645	865	146	1,544
Total	0 - 67,130	10,367	4,076	14,379

Given that one of his research questions ("How do creditors fare?") begged for a descriptive answer—a direct summary of the relevant data—Norberg's exercise was a reasonable place to start. But even when researchers pose questions they cannot answer with descriptive information alone, an important step in the analysis, whether of quantitative or qualitative evidence, usually is to provide such summaries. That is because alternatives—in the Norberg example, a list of 213 dollar values (71 cases times 3 categories)—are beyond the direct comprehension of most human beings; we cannot even hold this many numbers in our heads at one time, much less simultaneously interpret them. Accordingly, virtually all studies use data summaries rather than present data in their original raw form—including the most extensive numerical databases in quantitative research or the longest, most detailed verbal accounts of any real phenomenon in qualitative inquiries—in order to understand and communicate what the data are about.

Summaries take various forms. With numerical data, we often summarize many numbers with only a few, and indeed this activity accounts for much of the academic field of statistics. Some simple statistics typically invoked by legal scholars include the mean, median, mode, range, and standard deviation.[66] The first three are measures of *central tendency*, that is, they tell us the "center" of the distribution of dollar values of secured claims (or any factor of interest). The mean is the simple average, such that the $9,313 in Table 2 was the average amount of money the creditors with secured claims collected; the median is the case in the middle of the distribution of cases, such that half of the secured creditors collected more than $3,914 and half collected less; and the mode (which Norberg does not supply in his table) is simply the most frequently occurring value, such that if Norberg had only three cases in his priority claim category with collections of $0, $4,965, and $0, $0 would be the mode. The latter two, the range and the standard deviation, are measures of dispersion or variability, that is, they tell us the degree to which the data are spread around the typical values. The range is simply the minimum and maximum values for each claim type, such that the least

[64] Scott F. Norberg, "Consumer Bankruptcy's New Clothes: An Empirical Study of Discharge and Debt Collection in Chapter 13," 7 Am Bankr Instit J 415 (1999).

[65] Id., at 429. Norberg also reports the 25th and 75th percentiles. For purposes of presentation, we exclude these from Table 1 above.

[66] See Heise, The Importance of Being Empirical, supra note 2, at 826.

amount of money collected by creditors with secured claims was $0 and the most, $66,183; the standard deviation is a statistic that captures the distance between the values and the mean.[67]

Each numerical summary gives the reader a feel for the whole distribution of 71 cases for each claim type, but necessarily omits other features of the distribution. A simple example is to note that the median includes information about the central tendency, but omits information about variation. In other words, knowing the median amount of money collected by creditors with secured claims tells us nothing about the degree to which creditors deviated from that median: Were they tightly clustered around that $3,914 amount or not? Moreover, even though the table provides measures of central tendency and variation, other features, such as skewness (the degree to which the distribution is symmetric around the mean) or the dollar figures from individual cases, are necessarily lost. This is easy to see since the entire list of 213 numbers cannot be reproduced from Table 2, or equivalently, many different lists of 213 numbers are consistent with the numerical summaries it displays.

The fact that information is lost in summarization is not a problem in and of itself; the only difficulty comes if researchers discard useful information. Since information useless for one purpose can be useful for another, however, we cannot know solely from the numbers reported if the author did a good job at summarizing.

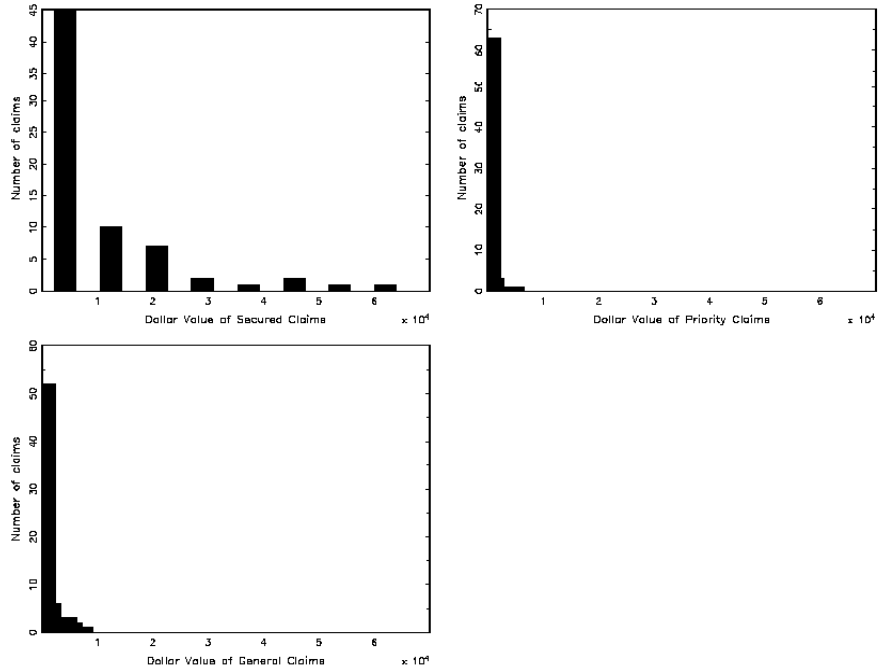
In some cases, graphic or other presentations of data can convey the right information without losing much of anything. This is not typical, but when possible we should take advantage of it, as we can for Norberg's data.[68] Rather than present the data in tabular form as he did, we might depict them in three separate *histograms* for collections in secured, priority, and general claims. Histograms are visual representations of the entire distribution, in this case, of the number of dollars creditors collected per case in each category.

Figure 1 presents examples with the horizontal axis in each histogram representing a dollar value and the vertical axis, the number of cases in each category. The only information lost in this presentation is the dollar variation that may exist within each bar, which groups together cases (the width of the bar is a choice made by the investigator). But note the gain: Features of the data that Norberg's table (*i.e.,* his choice of summary statistics) obscures turn out to be quite consequential. For example, we can now see (much more clearly than we could even calculate from Table 2) that the modal number of dollars received by creditors, for all three types of claims, is zero. This fact reveals, among other things, that means and medians (the statistics Norberg provides) are not particularly good summaries of this feature of the data since they obscure the spike at zero. In all likelihood, the processes that generated these data are disjointed: In creditor claims, it seems likely that one set of factors explains whether creditors will collect *any* money, and a partially separate set of factors explains *how much* creditors will collect (assuming some amount of money will be paid). Whether this result holds under more systematic analysis is an interesting question but one that arose only after we extracted valuable information about the distribution using a method that summarizes data while not discarding features of interest.

[67] The standard deviation is the square root of the variance. The variance is the average of the squared differences from the mean.

[68] We are grateful to Scott Norberg for providing us his data.

Figure 1. Histograms of Norberg's Data



However scholars choose to present their summaries, or whatever they decide to include in them, a cautionary note is in order: They should not reify their numbers. That is because single-number summaries need not exactly represent even one case.[69] So, for example, if we observe nine of twelve jurors voting to convict a defendant, we would not summarize this information by saying that the average juror was three-quarters in favor of a conviction.[70]

While this may seem obvious, it is surprising just how often legal scholars do reify numbers—especially when they attempt to present "profiles" of the "average" rather than summaries of the components. Consider how Eisenberg and his colleagues, in a study of whether jurors in capital cases assume responsibility for the sentences they impose, present their data:

> [R]esponses to several different interview questions suggest a relatively consistent picture of juror sentencing responsibility. The average juror understands and accepts the key role he plays in determining the defendant's sentencing; does not view the law as forcing him to reach a particular decision; does not view a death decision as

[69] Reification is one of the oldest statistical mistakes on record. See Stephen Stigler, The History of Statistics (1986).

[70] For another illustration, consider a recent election in Los Angeles, where 1.8 million voters cast yes-or-no ballots for 27 different referenda. For any one referendum, we might think that the "will of the voters" can be characterized by the fraction voting in favor, but of course not any individual voter. For example, if 66% of voters support a referendum, a claim that "the average voter is two-thirds in favor" would be absurd. The situation is even starker when considering the set of all the referenda, since the set of majority referenda winners does not necessarily correspond to the preferences of a majority of voters. In fact, in this election, not even a single person voted with the majority in every referenda. Steven J. Brams, D. Marc Kilgour, & William S. Zwicker, The Paradox of Multiple Elections, 15 Social Choice and Welfare 211 (1998).

something that the courts will likely reverse; and finds his service on a capital jury emotionally upsetting. On the other hand, he does not think it very likely that any death sentence he imposes will actually be carried out.[71]

In fact, the authors provide no evidence that their profile of the average juror accurately characterizes a majority of jurors, a few jurors, or any jurors at all. In this research, as in most, the "average juror" (as distinct from the average response of jurors to a survey question), is a creation of the researcher.

This caveat noted, when properly collected, presented, and understood, summary statistics are useful and often necessary ways of characterizing large data sets. Likewise, summaries usually play an important role in qualitative empirical research. Here description can take the form of a verbal summary, for example, when a researcher attempts to summarize whatever precedent (which may include two, three, or many more cases) is relevant to his or her concerns. Such is the Gelacak et al.[72] investigation, which focuses primarily on how U.S. district court judges use the departure provisions of the Sentencing Reform Act of 1984.[73] Recognizing that these judges may be influenced by appellate case law in their circuit, the authors also survey departure "jurisprudence" among the U.S. Courts of Appeals.[74] The following is how they describe the Ninth Circuit's approach:

> The Ninth Circuit was…notable in its departure review standards, as the court imposed rather stringent procedural requirements on district courts' decisions to depart. The Ninth Circuit required district courts to state the reasons for departure on the record [footnote omitted] and to explain the extent of the departure [footnote omitted]. The Ninth Circuit strictly applied these requirements and vacated numerous departures, especially upward departures, for the sentencing court's failure to provide an adequate explanation."[75]

In these three sentences, Gelacak et al. attempt to summarize the doctrine established in five cases, spanning a three-year period. Their brief description necessarily omits considerable detail in these cases, some of which the authors describe in accompanying footnotes.[76] What makes it nonetheless useful is that the features it retains are those the authors wish to communicate to readers, if indeed the quote accurately reflects the cases it was written to represent.

3. Making Descriptive Inferences.

Describing observations via summaries, as we discuss above, is a critical part of most research projects. But it is often not the primary goal; that goal, rather, is *inference—the process of using the facts we know to learn about facts we do not know*. There are two types of inference: descriptive and causal.

While researchers often use data summaries to make descriptive inferences, descriptive inferences are different than data summaries. We do not make them by summarizing facts; we make them by using facts we know to learn about facts we do not observe. To see the distinction, consider Milhaupt and West's investigation of organized crime, which seeks to understand why it "emerge[s],

---

[71] Theodore Eisenberg, Stephen P. Garvey, & Martin T. Wells, "Jury Responsibility in Capital Sentencing: An Empirical Study," 44 Buff L Rev 339 (1996).

[72] Michael S. Gelacak, Ilene H. Nagel, & Barry L. Johnson, "Departures Under the Federal Sentencing Guidelines: An Empirical and Jurisprudential Analysis," 81 Minn L Rev 299 (1996)

[73] 18 U.S.C. 3553(b) (1994) (providing that a sentencing court may depart from the sentencing guidelines, if the court finds "that there exists an aggravating or mitigating circumstance of a kind, or to a degree, not adequately taken into consideration by the Sentencing Commission in formulating the Guidelines that should result in a sentence different from that described…")

[74] Gelacak, Nagel, & Johnson, Departures Under the Federal Sentencing Guidelines: An Empirical and Jurisprudential Analysis, supra note 72, at 336.

[75] Id.

[76] See, for example, Id., at n. 156.

and what function…it plays in an economy."[77] The researchers focus on Japan to address these questions—that is, they amass and summarize data from that country for the years 1972-1997—but they also hope to use "the Japanese experience" to develop "insights into organize crime in environments as diverse as Russia and Sicily."[78] They are thus seeking to make two descriptive inferences: (1) they want to use summaries of the specific data they collect on Japan to learn about organized crime in Japan generally and (2) they want to use what they learn about Japan to learn about the rest of the world.

Lederman engages in an analogous enterprise. In attempting to answer a *general* research question—"Which cases go to trial?"—she draws a random sample of 400 cases docketed by a *specific court* (the Tax Court) during a *discrete* period (1990-1995).[79] Any summary of these 400 cases, or indeed all 400 cases, are of little interest in and of themselves. Their purpose, to Lederman, was to help her learn about data she did not have, all Tax Court cases that do and do not go to trial and, ultimately, all cases filed in all courts.

As substantively distinct as Milhaupt and West's and Lederman's studies are, they thus share a common feature: In both, the researchers are attempting to make a descriptive inference. By collecting observations on one country (Japan) or on one court (the Tax Court) at a particular point in time they are trying to learn something—to make a descriptive inference—about that particular country or court in general (not simply at the point in time they are studying). At the same time, they are seeking to generalize about the world based on testing just a small part of it. That is, they want to learn something about the other countries or other courts, or other time periods they are not observing. The facts that they do not observe or know are sometimes called features of a *population* (*e.g.,* all countries of interest, including Japan), the values of which they are seeking to learn by taking measurements on a *sample* (*e.g.,* Japan in the years 1972-1997).

That Milhaupt and West and Lederman understand this distinction is important since the critical first step in making a descriptive inference is to identify the target of the inference—the fact we would like to know, such as organized crime in all countries or settlement in all cases. When scholars do not take this step, they open up themselves to a virtual Pandora's Box of ills, many of which we detail in Part VIII. Of immediate concern here is that failure to follow this principle can lead to what should be unnecessary criticism of the research. Such was the fate of a Bradley and Rosenzweig's study of the Bankruptcy Code.[80] To support their primary argument that Congress ought repeal Chapter 11, the researchers examined the behavior of a sample of publicly traded companies that filed for bankruptcy. This led Warren, in a damning critique of the research, to raise the obvious question: Since publicly traded companies constitute less than 1% of all Chapter 11 cases and since previous research indicates that "the experience of large, publicly traded companies in bankruptcy differs sharply from that of small, private companies," is it legitimate for Bradley and Rosenzweig to advocate sweeping legislative change on the basis of their sample?[81] To Warren, the answer was a definitive no, the researchers cannot "claim that their data apply with equal force to all corporations choosing Chapter 11."[82]

If, in fact, the target of Bradley and Rosenzweig's inference was all companies filing for bankruptcy, then Warren makes a reasonable point; if, in fact, the researchers intended to make claims only about publicly traded companies then Warren is stretching—but by no fault of her

---

[77] Curtis J. Milhaupt & Mark D. West, "The Dark Side of Private Ordering: An Institutional and Empirical Analysis of Organized Crime," 67 U Chi L Rev 41 (2000).

[78] Id., at 44.

[79] Lederman, Which Cases Go to Trial?, supra note 27, at 327.

[80] Michael Bradley & Michael Rosenzweig, "The Untenable Case for Chapter 11," 101 Yale L J 1043 (1992).

[81] Elizabeth Warren, "The Untenable Case for Repeal of Chapter 11," 102 Yale L J 437 (1992), at 443.

[82] Id.

own.[83] It is up to the researchers, not the readers, to specify the target of their inference. Should that target be elusive or unclear to the investigators, what they might do is imagine how they would proceed with an unlimited budget and no limits on the amounts of time and effort they could expend. If, in that hypothetical situation, they find that they are unable to clarify the quantity to be estimated with great precision, then they should rethink the project at a much earlier stage, if not from scratch. Indeed, without an unambiguously identified target of inference, a research project cannot be reasonably evaluated and, hence, cannot be successful.

Most of the studies we have considered so far happen to rely on numerical evidence, but scholars using qualitative information seek to make descriptive inferences just as often. Consider doctrinal analyses of particular areas of the law. In many, researchers attempt to make an inference about the "state of law" by focusing on "a few…exemplary" or "key cases."[84] But, in nearly as many, the researchers leave their attempt at descriptive inference unstated, often relying instead on the ever-present "string citation"[85]—replete with "a few…exemplary" cases—to do the work. Such was Black's examination of doctrine governing state regulation of political parties, which asserts "These cases demonstrate that the Court's early intervention into the state regulation of political parties did not clearly benefit major parties over minor parties,"[86] "These cases," we learn in footnotes, are nine , a small fraction of all such decisions.[87] Others eschew string citations for deep analyses of the "key" or "exemplary" cases. To demonstrate that the U.S. Supreme Court has, in fact, adopted the new originalism approach to the Fourth Amendment that he wants to critique,[88] Sklansky examines eleven[89] of the fifty-four (a figure we calculated[90]) search and seizure cases the Court decided during the period under analysis. Whether these eleven cases or Black's five are "exemplary," "key," what exemplary or key would mean in this context, or whether they adequately or fairly represent any cases other than those selected by the author we do not know.

What we do know is that descriptive inference is not the near-trivial task that these authors implicitly (via string citations or deep analysis) make it out to be. It could be that Black tells us something important about state regulation from his focus on five key (or exemplary) cases or that Sklansky has captured a willingness on the part of the Court to invoke "new" originalism to resolve Fourth Amendment cases from his eleven. But it also could be that we are led astray by the research.

---

[83] Bradley & Rosenzweig, "The Untenable Case for Chapter 11," make claims that could support either target. In note 80 of their article, they write: "Our data, of course, are also limited to public corporations. We therefore make no empirical case against Chapter 11 insofar as it applies to nonpublic corporations." This, of course, works against the relevant portion of Warren's critique. Supporting her point is, among other parts of the Bradley & Rosenzweig article, the Conclusion, which calls for the unqualified repeal of Chapter 11.

[84] Schuck & Elliot, To the Chevron Station, supra note 13, at 1060.

[85] "String citation" is a term of art in the legal literature referring to a list of citations offered to support a point in the text.

[86] Benjamin D. Black, "Developments in the State Regulation of Major and Minor Political Parties," 82 Cornell L Rev 109 (1996), at 124.

[87] See, for example, cases reported in Lee Epstein & Charles D. Hadley, "On the Treatment of Political Parties in the U.S. Supreme Court, 1900-1986," 52 J Politics 413 (1990); Nathaniel Persily and Bruce E. Cain, "The Legal Status of Political Parties: A Reassessment of Competing Paradigms," 100 Colum L Rev 775 (2000).

[88] See supra note 8.

[89] He also explores a dissent Scalia wrote while serving as a judge on the U.S. Court of Appeals for the District of Columbia. That dissent came in the case of *Ollman v. Evans,* 750 F.2d 970 (D.C. Circuit 1984) (en banc), cert. denied, 471 U.S. 1127 (1985). Moreover, in addition to the eleven cases that he examines in detail, mentions six in passing. See Sklansky, The Fourth Amendment and Common Law, supra note 8, at 1749-1768.

[90] We derived this figure of fifty-four from the U.S. Supreme Court Judicial Data Base, supra note 61, for the 1986-1999 terms (the time period of interest to Sklansky). To parallel Sklansky's "exemplary" cases as closely as possible, we selected only orally argued search and seizure cases (issue=16, 17, or 18) resulting in signed opinions of the Court (analu=0; dec_type=1 or 7).

Hillman[91] suggests as much as regard to Farber and Matheson's work on whether reliance remains relevant to promissory estoppel cases.[92] While Farber and Matheson report the demise of reliance in "key" cases—thereby putting a serious dent into the prevailing conventional wisdom and, in the longer run, creating a "new consensus"[93]—Hillman, conducting a qualitative and quantitative investigation of all U.S. promissory estoppel cases decided between 1994 and 1996 (rather than of just "key" cases), reaches precisely the opposite conclusion:

> Analysts have…reported the unimportance of reliance as a substantive element of promissory estoppel and the invariable award of expectancy damages in successful cases. The reality, at least during the mid-1990s, was very different. A showing of reliance was crucial to recovery and the remedy was not exclusively expectancy damages.[94]

If Hillman's analysis is valid, Farber and Matheson's inference is clearly flawed.

The same eventually may be said of Sklansky's study.[95] While we did not conduct a detailed analysis in the style of Hillman's, we did consider the possibility that the eleven cases Sklansky selected for deep study may not be exemplary or, perhaps, not even key but rather those that were most visible to him. Specifically, we compared the eleven cases in Sklansky's article to the population of fifty-four on a simple indicator of "visibility": whether the case received a headlined story in the *New York Times* on the day following the Court's decision.[96] The results of our investigation are illuminating: Of the eleven cases Sklansky analyzed, nine (or 82%) were the subject of a *Times* article, but in the population, 52% received coverage.

Does this mean that Sklansky's conclusions about the Court's adoption of new originalism miss the mark? No. Even though he may have selected the eleven cases because they were especially visible to him, it is *possible* that he was able to reach a valid descriptive inference if visibility is unrelated to the variables in his study. But possible is different than conclusive especially since he provides no argument or evidence about this, and therein lies the problem: We have no way to tell whether his eleven did or did not support the inference he reaches.

How might Sklansky, Farber and Matheson and the others have made their descriptive inferences more accurate and less uncertain? The answer is simple: They needed to reveal far more about the process by which they generated and observed their data—the whole process from the time the world produced the phenomena of interest to the moment when the data were in their possession and considered final. If, for example, we can be assured that the five cases Black

---

[91] Robert A. Hillman, "Questioning the 'New Consensus' on Promissory Estoppel: An Empirical and Theoretical Study," 98 Colum L Rev 580 (1998).

[92] Daniel A. Farber & John H. Matheson, "Beyond Promissory Estoppel: Contract Law and the 'Invisible Handshake'," 52 U Chi L Rev 903 (1985). The researchers, at 907, note 14, apparently reviewed over 200 promissory estoppel cases. But, as Hillman, "Questioning the "New Consensus" on Promissory Estoppel: An Empirical and Theoretical Study," supra note 91, points out, Farber and Matheson "discuss only a few that are supposed to show that courts generally stretch their analysis to find reliance in successful promissory estoppel cases. In fact, Farber and Matheson state…in their conclusion: 'In key cases promises have been enforced with only the weakest showing of any detriment to the promisee.' But the reader is left wondering why the cases discussed are 'key'."

[93] *See* Id., at 582, claiming that Farber & Matheson, Beyond Promissory Estoppel, supra note 92, along with Edward Yorio & Steven Thel, "The Promissory Basis of Section 90," 101 Yale L J 111 (1991), "have been enormously influential," with later studies treating them as creating a "new consensus." In other words, sucessor studies "assume the accuracy of [Farber & Matheson and Yorio & Thel] in building their own theses."

[94] Hillman, Questioning the "New Consensus" on Promissory Estoppel, supra note91, at 619.

[95] Sklansky, The Fourth Amendment and Common Law, supra note 8.

[96] Social scientists have used this indicator to determine the "salience" of cases and laws. See, for example, Lee Epstein and Jeffrey A. Segal, "Measuring Issue Salience," 44 Am J Pol Sci 66 (2000); David Mayhew, Divided We Govern (1991). We could not use other possible indicators of visibility—such as the number of citations in the law reviews, treatises, case books or in other cases—because many Sklansky's cases are of a relatively recent vintage.

observed are just like those to which he was inferring, then inferences could be of high quality.[97] Unfortunately, in Black's (and in Sklansky's) study, the cases were selected only by some private, undisclosed decision by the investigator, and thus readers have no way to assess the quality of the inferences.

A logical exception that might be raised to this point is that the quality of inferences can be judged by knowing the reputation of the investigators: We ought believe Sklansky because he is a distinguished professor at a well-regarded law school (UCLA) but perhaps not Black, a "mere" student at Cornell. This kind of appeal-to-authority, however, is wholly irrelevant to valid inference. Even if it were true that famous authors are wrong less often than obscure authors (a hypothetical with little supporting evidence!), they are wrong sometimes and only the evidence—the process by which the data came to be observed—not the investigator's fame, job, status, or income is the stuff of serious scientific inference.

4. Making Causal Inferences

The examples above are just a few of the many we could have selected to illustrate descriptive inferences, for making them is often a critical part of research programs. The same may be said of causal inference, such as in studies in which the scholar, lawyer, or judge wants to know whether one factor or set of factors leads to (or causes) some outcome. Did the U.S. Supreme Court's decision in *Miranda v. Arizona*[98] bring about to a decline in the number of confessions?[99] Do laws (and other types of penalties) aimed at reducing alcohol-impaired driving cause a decline in the rate of alcohol-related traffic fatalities?[100] Did *Chevron v. NRDC*[101] lead federal appellate courts to give greater weight to agency construction of laws?[102] Do various rules encourage parties to settle their disputes?[103]

All these questions ask whether a particular "event"—the presence or absence of which we refer to as the *key causal variable* (in the examples above, Supreme Court decisions, driving laws, and settlement rules)—caused a particular "outcome," or *dependent variable* (in the examples above, confession rates, traffic fatalities, deference, settlement). The events and outcomes can be characterized as variables that take on different values, that is, they vary: Either *Miranda* exists or it does not; confessions can increase, decrease, or stay the same.

---

[97] Black, "Developments in the State Regulation of Major and Minor Political Parties," supra note 86.

[98] 348 U.S. 436 (1966).

[99] See, for example, Paul G. Casell, "Miranda's Social Costs: An Empirical Assessment," 90 Nw U L Rev 387 (1996) and studies cited therein; Paul G. Casell & Richard Fowles, "Handcuffing the Cops? A Thirty-Year Perspective on Miranda's Harmful Effects on Law Enforcement," 50 Stan L Rev 1055 (1998); John J. III Donohue, "Did Miranda Diminish Police Effectiveness?," 50 Stan L Rev 1181 (1998).

[100] See, for example, Laurence H. Ross, "Administrative License Revocation in New Mexico: An Evaluation," 9 Law & Policy 5 (1987); Frank A Sloan, Bridget A. Reilly, & Christoph Schenzler, "Effects of Tort Liability and Insurance on Heavy Drinking and Drinking and Riving," 38 J L & Econ 49 (1995).

[101] Chevron, U.S.A., Inc. v. NRDC, 467 U.S. 837, 842-843 (1984) (providing guidance for court review of an agency's construction of a statute in the form of a two-step test: "First, always, is the question whether Congress has directly spoken to the precise question at issue. If the intent of Congress is clear, that is the end of the matter; for the court, as well as the agency, must give effect to the unambiguously expressed intent of Congress. If, however, the court determines Congress has not directly addressed the precise question at issue, the court does not simply impose its own construction on the statute, as would be necessary in the absence of an administrative interpretation. Rather, if the statute is silent or ambiguous with respect to the specific issue, the question for the court is whether the agency's answer is based on a permissible construction of the statute.").

[102] See, for example, Schuck & Elliot, To the Chevron Station, supra note 13; Ellen P. April, "Muffled Chevron: Judicial Review of Tax Regulations," 3 Fla Tax Rev 51 (1996).

[103] See, for example, David A. Anderson & Thomas D. Jr. Rowe, "Empirical Evidence on Settlement Devices: Does Rule 68 Encourage Settlement?," 71 Chi Kent L Rev 519 (1995).

To address these common questions in legal research, what many scholars do is observe what occurs before and after a change in the causal variable. This is the tack often taken in studies of the effect of *Miranda,* as well as of *Chevron.* In a doctrinal analysis of trial and appellate court cases involving tax regulations, for example, April concludes that "*Chevron* has not worked the revolution in the balance of powers between agencies and the courts that some commentators feared. It has not displaced judges from a key role in the review of agency regulations. Instead…*Chevron* may have decreased deference to administrative action by encouraging courts themselves to decree the meaning of a statute."[104] In contrast, Schuck and Elliot, relying on quantitative evidence, report that the percent of administrative agency decisions that appellate courts affirmed in 1984 (prior to *Chevron)* was 70.9%; after *Chevron,* that figure rose to 81.3.[105]

Does this necessarily mean that *Chevron* caused an increase (or in April's case, a decrease) in judicial deference to agencies? Of course not. It is possible that the increase (decrease) would have occurred (or, perhaps would have been greater) in the absence of *Chevron.* Only by rerunning history and holding everything else in the world constant except *Chevron* (*i.e.,* in one version of our recreated history, the Court handed down *Chevron* and the other, it did not) would we be able to define the total causal effect. If, in the version of our history without *Chevron,* we observed no decline in deference but in the version with *Chevron* we observed deference, then we might conclude that *Chevron* had a causal effect.[106]

Since we cannot rerun history in this fashion, we must rely on causal inference. Causal inference too is about using facts we do know to learn about facts we do not know. In fact, *a causal inference is the difference between two descriptive inferences*—the average value the dependent variable (*e.g.,* the percentage of cases decided between 1984 and 2000 in which judges defer to the agency) takes on when a "treatment" is applied (*e.g., Chevron* is introduced) and the average value the dependent variable takes when a "control" is applied (*e.g.,* when *Chevron* is not introduced). The *causal effect*—the goal of the process of causal inference—is this difference, the amount that the percentage of judicial deference increases or decreases when we move from a world without *Chevron* to a world identical in all respects except for the presence of *Chevron.*

Learning the values of the dependent variable when the key causal variable indicates treatment and when it indicates control requires two inferences since neither quantity typically can be directly observed. But there is an additional complication: Researchers can only estimate directly the actual value of the degree of judicial deference when either the treatment or the control is applied but not both, since either *Chevron* was introduced in the real world or it was not. This is known as the *fundamental problem of causal inference*,[107] and it is indeed a fundamental problem since no matter how perfect the research design, no matter how much data we collect, no matter how much time, effort, and research resources we expend, we will never be able to make causal inferences with certainty. At most, one of the two descriptive inferences will be based on "factual" information, and at a minimum one will require "counterfactual" inference. (Of course, counterfactual inference is of

---

[104] April, Muffled Chevron: Judicial Review of Tax Regulations, supra note 102, at 55. Qualitative (doctrinal) examinations of the effect of *Chevron* are not uncommon. See, for example, Damien J. Marshall, "The Application of Chevron Deference in Regulatory Preemption Cases," 87 Georgetown L J 263 (1998); John F. Coverdale, "Court Review of Tax Regulations and Revenue Rulings in the Chevron Era," 64 Geo Wash L Rev 35 (1995); Linda Matarese, "Has the *Chevron* Deference Made a Difference When Courts Review Federal Banking Agency Interpretations of the Glass-Steagall Act," 33 Howard L J 195 (1990).

[105] Schuck & Elliot, To the Chevron Station, supra note 13, at Table 3.

[106] We adapt the language here and in the next two paragraphs from King, Keohane, & Verba, Designing Social Inquiry, supra note 1, at 79.

[107]Paul Holland, "Statistics and Causal Inference," 81 Journal of the American Statistical Association 945 (1986).

interest in and of itself, such as when we ask what the world would have been, or would be, like had certain policies been promulgated.)

The fundamental problem of causal inference, combined with all the difficulties of making even descriptive inferences, has led some scholars to suggest that we should never frame research in terms of causal questions, hypotheses, and statements, that we should talk only in the language of correlations or associations, as in "*Chevron seems associated* with an increase in judicial deference," and not "*Chevron caused* an increase in judicial deference." Sometimes this is all that can be done, but we disagree that the goal should be changed: Simply because uncertainty cannot be eliminated does not mean we cannot or should not draw causal inferences when the research necessitates it. Legal researchers, lawyers, the courts, and legislators need to make causal inferences, and so giving up and redefining the goal is not an option. Moreover, generating useful, policy-relevant research topics is among the things that legal scholars do best. We thus recommend that researchers not change the object of their inferences because causal inference is difficult. Instead, they should make their questions as precise as possible, follow the best advice science has to offer about reducing uncertainty and bias, and communicate the appropriate level of uncertainty readers should have in interpreting their results—much of which we summarize in this Article.[108]

B. General Guidelines

In Parts III-VIII we explicate specific rules governing different components of the research process. Immediately below, we provide guidelines to which all empirical research should adhere regardless of whether the goal is to amass or summarize data, make descriptive or causal inferences, or some combination thereof. Indeed, these guidelines apply to so many areas and stages of a research project that together they almost describe a general attitude or approach to thinking about empirical scholarship. In this regard, we note that in the social sciences, faculty frequently tell their Ph.D. students a tale that bears repeating in legal circles:

> When you publish your paper, out there somewhere will be a graduate student holding a yellow-lined pad of paper with your name scrawled at the top. Expect everything you do to be scrutinized, any rival explanation you have not explored to be tested, and every possible way you could be proven wrong to be dissected. Anticipate. Get to each of these areas before this graduate student. When you do empirical research, follow the rules of inference.

The guidelines that follow provide a crucial starting point for heeding this advice.

1. Research must be replicable

Good empirical work adheres to the *replication standard*: Another researcher should be able to understand, evaluate, build on, and reproduce the research without any additional information from the author. This rule does not actually require anyone to replicate the results of an article or book; it only requires that researchers provide information—in the article or book or in some other publicly available or accessible form—sufficient to replicate the results in principle.[109] Unfortunately, the present state of legal scholarship nearly always fails this most basic of tests.

Consider two examples. In the first, "Explaining the Pattern of Secured Credit," the author, Mann, explains the procedures he used to amass his data:

---

[108] This point was also made by King, Keohane, and Verba, Designing Social Inquiry (1994), supra note 1, at 76ff.

[109] Gary King, "Replication, Replication,"with comments and rejoinder, 28 P.S.: Political Science and Politics 443 (1995).

> I constructed a set of interviews designed to mirror as closely as possible the lending market as a whole. On the borrowing side, I viewed borrowers as differentiated by the size of the company, the company's access to publicly traded debt, and the company's line of business. Accordingly, I interviewed responsible individuals at large and small companies, ranging from two publicly traded Fortune 100 companies to several small, closely held companies (including one that has never turned a profit). I also interviewed several borrowers that borrow exclusively in the private debt market, a borrower in the process of issuing its first public debt issue, and several borrowers that are active in the public debt markets. Finally, I interviewed individuals operating in various lines of business, including real estate, pharmaceuticals, industrial tooling, manufacturing, and computer technology and service. On the lending side, I interviewed representatives of each of the major types of institutional lenders in our economy—insurance companies, banks, and asset-finance companies—as well as several noninstitutional lenders who extend significant amounts of trade credit to their customers.[110]

Likewise, in their study—"Jury Responsibility in Capital Sentence: An Empirical Study"—Eisenberg and his colleagues describe how they gathered the data to answer their primary research question (Do jurors in capital cases assume responsibility for the sentences they impose?):

> Jurors who sat in forty-three South Carolina murder cases were randomly sampled with a goal of four juror interviews per case. The sample includes twenty-three cases resulting in sentences and twenty cases resulting in life sentences. The cases in the study consist of all South Carolina capital cases brought from enactment of the South Carolina Omnibus Criminal Justice Improvements Act of 1986 to when interviews were terminated in the summer of 1993….A total of 153 live interviews were completed by interviewers trained to work with the [fifty-one page] interview instrument…. Jurors were interviewed after they had served, not before.[111]

Despite their designers' laudable efforts at explaining their research procedures, neither of these studies is replicable: Another researcher could not reproduce them without talking to the authors. This is obvious in the case of the Mann essay: Since he does not provide a definition of "responsible individuals" we have no way of determining if we would identify the same type of "responsible individuals" as did he. We might be able to infer a definition from his list of subjects but, alas, since 6 of his 23 requested anonymity, this would be hazardous. Moreover, while Mann provides us with the categories of companies from which he chose his "sample" of subjects, he never tells us how he selected particular companies within those categories. Fully consistent with his written description are strategies of selecting companies at random (and so probably unrelated except by chance to any other variable), on the ground of where Mann lives (and so convenient to him but possibly biased in unspecified ways), on the companies' borrowing and lending strategies (and so biased for inferences about these strategies), or on the basis of how friendly the companies were toward the researcher (and thus presumably biased towards firms with better public relations' departments). The choice among these strategies could produce almost any empirical result, and so without Mann explicitly delineating his selection criteria, readers cannot know how he conducted his study, and thus cannot see the relationship between his data and the target of his inferences. The result of this lack of replicability, then, is a set of conclusions that have little known empirical basis: They may apply to the companies he chose but readers cannot evaluate whether they apply to the population of companies that are of interest.

The lack of replicability is less apparent in the Eisenberg et al. study but no less problematic. To see why, think about what information an investigator would need to know, but that the authors do not provide, to replicate their work. A partial list includes:

---

[110] Ronald Mann, "Explaining the Pattern of Secured Credit," 110 Harv L Rev 625 (1997), at 631-632.

[111] Eisenberg, Garvey, & Wells, Jury Responsibility in Capital Sentencing, supra note 71, at 350.

(1) What do the authors mean by "randomly sampled"? How did they obtain the random numbers? Was it a simple random sample with equal probability of selection, a stratified sample, or something else?[112]

(2) How did they approach the jurors? Did they tell them that they were doing a study on juror responsibility? How many times was each juror contacted, and at what point did they give up? Were they contacted by telephone or in person? Did someone of the same sex and race contact each?

(3) What does "a goal of four juror interviews per case" mean? If they could not get four, how did they proceed? If more than four volunteered, where they all interviewed? Were the jurors who were ultimately contacted those who felt responsible for their decision and comfortable talking about it? What was the refusal rate?

More generally, how did the authors get from a population of 516 jurors (43 cases with 12 jurors each) to a sample of 153? How do the 153 differ from the remaining (516-153=) 363? If Eisenberg et al. had free reign in choosing their sample, they could have produced biases of almost any type and, as a result, drawn virtually any conclusion. If we could choose a sampling method, in all likelihood we could easily reverse their article's conclusions. Of course the real world is not always so malicious, but it sometimes is. And the burden of proof in empirical research always remains with the researcher.[113]

Eisenberg et al.'s study relies on quantitative evidence. But the rule that empirical research must be replicable applies with equal force to studies relying on non-numerical evidence. In many, perhaps most, instances, legal academics conducting these sorts of investigations rarely provide even a tracing of how they collected the evidence. The Sklansky essay on new originalism[114] and Black's on state regulation of political parties[115] are exemplary but there are scores of other doctrinal studies that are equally negligent in providing the reader with guidelines sufficient to replicate the analysis. We rarely learn:

---

[112] For definitions and explanations of these samples, see infra p. 76.

[113] Eisenberg, Garvey, & Wells' research is part of the Capital Juror Project (CJP), a study of how jurors in 14 states reach decisions capital cases. Although the CJP's principle investigator provides a detailed description of the study's general sampling design (see William J. Bowers, "The Capital Jury Project: Rationale, Design, and Preview of Early Findings," (1995) 70 Ind. L.J. 1043) and the authors cite him four separate times, their citations do not have anything to do with the sampling procedure itself (see footnotes 32, 53, 55, and 66 on the Milgrom experiments, and comparing their results to the multi-state data set, respectively), so that even the more careful reader might never have thought to consult the Bowers text. Moreover, the particular sampling procedures varied across states in a manner not desribed by any previous work we could identify, and thus, even though Bowers describes the CJP's general sampling design, we must rely on Eisenberg, Garvey and Wells to learn what was done in South Carolina (also see Eisenberg, Garvey, & Wells, "Deadly Confusion: Juror Instruction in Capital Cases," 79 Cornell L Rev 1 (1993)). Unfortunately, they do not do explain what they did (for an example of what they might have done, see Marla Sandys' description of the CJP's Kentucky data, "Cross-overs-Capital Jurors who Change their Minds about Punishment: A Litmus Test for Sentencing Guidelines," (1995) 70 Ind. L. J.1183). Subsequent work by Eisenberg, Garvey and Wells also does not tell us much about the way their South Carolina study was conducted (see Theodore Eisenberg, Steven P. Garvey, & Martin T. Wells, "Forecasting Life and Death: Juror Race, Religion and Attitude Toward the Death Penalty 30 J Legal Stud 207 (2001), Eisenberg, Garvey, & Wells, "The Deadly Paradox of Capital Jurors." 74 S. Cal. L. Rev. 1 371 (2001), Eisenberg, Garvey, & Wells, "But Was He Sorry? The Role of Remorse in Capital Sentencing," 83 Cornell L. Rev. 1599 (1998), Stephen P. Garvey, "Aggravation and Mitigation in Capital Cases: What do Jurors Think?" 98 Colum. L. Rev. 1538 (1998), and Garvey, "The Emotional Economy of Capital Sentencing," 75 N.Y.U.L. Rev. 26 (2000)).

[114] Sklansky, The Fourth Amendment and Common Law, supra note 8.

[115] Black, "Developments in the State Regulation of Major and Minor Political Parties," supra note 86.

(1) How authors canvassed the relevant case law and what precisely was the population from which they sampled.

(2) How authors selected their cases and how many they read.

(3) How authors distinguished "key" or "a few…exemplary cases"[116] from those that are not central or not typical.

We could raise similar questions about analyses of "legislative" or "framers" intent that continue to populate the law reviews. Take Engel's 1999 work, "The McCulloch Theory of the Fourteenth Amendment," which argues that the U.S. Supreme Court, in its decision in *City of Boerne v. Flores,*[117]

> went astray in focusing upon the judicial branch as the ultimate interpreter of the Fourteenth Amendment. While the Court may retain the last word, the judicial reading obscures the Framers' conviction that it would be Congress, and not the courts, that would be the first reader, and primary enforcer, of the Fourteenth Amendment.[118]

To make this claim, Engel relies (in large part) on historical evidence culled from the congressional record of debates over the Civil War Amendments. While there is nothing otherwise wrong with this strategy, it is—at least in the way that Engel deploys it—entirely nonreplicable. He, like many of those conducting these sorts of analyses, never tells readers how he surveyed the congressional material, how many of the debates he read, and whether the material cited in the article represents "key" events, a "few exemplary" passages, or a systematic sample of some specific kind (*i.e.,* selected with known and uniformly applied rules).[119]

Why is such documentation a requisite step in conducting empirical research, regardless of whether the work is qualitative or quantitative in nature? There are two answers to this question, with the first centering on the ability of outsiders to evaluate the research and its conclusions. In a broad sense, the point of the replication standard is to ensure that a published work stands alone so that readers can consume what it has to offer without any necessary connection with, further information from, or beliefs about the status or reputation of the author. The replication standard keeps empirical inquiry above the level of ad hominum attacks on or unquestioning acceptance of arguments by authority figures.

To see how failure to obey the rule impinges on our ability to evaluate work on its own merit, return to the Eisenberg et al. study of juror responsibility.[120] Based on their interviews with a "random sample" of individuals who served as jurors in capital cases, the investigators reach the following conclusion:

> [W]e find that most jurors accept responsibility for the capital sentencing decision, though a significant minority do not. Most jurors understand and acknowledge the primary role they play in sentencing a defendant to life or death. However, beliefs that cannot easily be changed limit the degree to which jurors view themselves as causing the

---

[116] Schuck & Elliot, To the Chevron Station, supra note 13, at 1060.

[117] 521 U.S. 507 (1997) (striking down the Religious Freedom Restoration Act, 42 U.S.C. 2000bb (1994)).

[118] Steven A. Engel, "The McCulloch Theory of the Fourteenth Amendment: City of Boernes v. Flores and the Original Understanding of Section 5," 109 Yale L J 115 (1999). This work is a "note," a category of law review article usually written by students. As such, some might question why we would be use it as an illustration of problems with articles, usually written by law professors. In our view, it is the content, not the author, that counts. Moreover, as we point out in Section I, the best articles for our expository purposes are those that are especially good, save perhaps the one point we are illustrating. Even if it is true that faculty articles are normally better than student notes, such was not the case here.

[119] For others who have raised these sorts of questions about studies attempting to uncover intent, see supra note 20.

[120] Eisenberg, Garvey, & Wells, Jury Responsibility in Capital Sentencing, supra note 71.

defendant's sentence. Jurors view defendants as primarily responsible for setting of the sequence of events leading to the sentencing decision and do not believe that most death sentences will be carried out.[121]

As we just pointed out, the procedures Eisenberg and his colleagues invoke to make these claims about the real world are not transparent. Recall, for instance, the first sentence of the description of their procedures: "Jurors who sat in forty-three South Carolina murder cases were randomly sampled with a goal of four juror interviews per case."[122] We raised, among other questions, how Eisenberg and his colleagues approached the jurors.

Let us suppose that Eisenberg et al., in response to our concerns, rewrote the description of their procedures, with the first three sentences now reading as follows:

> Jurors who sat in forty-three South Carolina murder cases were randomly sampled with a goal of four juror interviews per case. We began by contacting, in alphabetical order, all jurors in all South Carolina capital cases. We told them that we were conducting a study on whether jurors take responsibility for their decisions; we also told them that we wanted to interview them, with an eye toward completing a fifty-one page survey. If we received no response, we followed up three times. If a juror declined or did not respond, we went to the next juror on the list. We gathered 34 pieces of information from the public record on jurors who did not fill out our survey and performed an analysis that we report below indicating that those jurors we interviewed were similar (in all measurable respects) to those we did not interview.

With this "new" knowledge about how the researchers contacted and selected jurors, would we evaluate and interpret their results differently? In all likelihood, the answer is yes. We might conclude that at least some of their original findings were entirely predictable; or, even more to the point, we might question whether the results tell us anything at all meaningful about the real world given the obvious biases in their "sample." The fact is that even the best empirical research can be inadvertently affected by hundreds of confounding factors; identifying those about which we have knowledge is the least we can do to try to reduce bias.

The same, of course, could be said of Engel's study on the intent of the framers of the Fourteenth Amendment.[123] Suppose the author described his procedure as follows:

> After reading the Supreme Court's decision in *Boerne,* I was distressed. It seemed to me that the justices, like so many constitutional law scholars, went astray in focusing upon the judicial branch as the ultimate interpreter of the Fourteenth Amendment. To prove this argument, I went to congressional debates over the Civil War Amendments. I analyzed only comments and drafts indicating that I was right to feel distressed.

Of course, this statement is ludicrous; no scholar would ever write such words. But do scholars proceed this way without saying so? To answer this question, readers need to know precisely what was done in practice so that we can decide what merits our attention, what is worth further research, and on what makes sense to base public policy. Nothing in Engel's article provides readers with the procedures he used to sort through the vast quantity of historical evidence available to him and, thus, it is impossible to know whether our (hypothetical) description of his modus operandi is not precisely how he proceeded. (And, if he did, we would most certainly construe his findings in a different light.) As it stands, readers have no idea whatsoever how to interpret his results, other than by an (illegitimate) appeal to authority and his reputation. This is precisely what the replication standard is designed to prevent.

---

[121] Id., at 341

[122] Id., at 350.

[123] Engel, The McCulloch Theory of the Fourteenth Amendment, supra note 118.

This takes us to the second reason why we insist that scholars make their procedures public: Those procedures may, and in most instances do, influence the outcomes reported in research. Whatever procedure Eisenberg et al.[124] used to draw their sample led them to select some jurors and exclude others. Since they base their results on juror responses' to interview questions, another selection procedure—which would have excluded some of the jurors they interviewed and included some they did not—could have produced entirely different results. The same holds for Engel's investigation and, in fact, for any study that seeks to make claims or inferences about the real world.

2. Research is a Social Enterprise.

One of the points implicit in our discussion of the replication standard is that the author of the research is entirely irrelevant or, in the parlance of certain humanist schools of thought, "dead." His or her attributes, reputation, or status are unimportant; sentences that begin "I think" or "I believe" are beside the point. What is important is his or her contribution to the scholarly literature, to the communal or social enterprise of learning about the world.

That this guideline now holds for all those conducting empirical research represents a monumental change in thinking. Long ago, academics worked much more in isolation, and during that time produced some brilliant findings about which no one ever learned; others made mistakes that went uncorrected for decades, effectively wasting their entire careers. Progress or a cumulation of knowledge was rare. The reason academics now cluster together in universities is not necessarily because they like each other; it is because their work is much better as a result. These days, the advancement of knowledge depends on an active community of scholars working together in cooperation and competition.

On this much scholars in most units in universities and colleges concur—though not necessarily those in schools of law. While legal academics seem to view teaching as a social enterprise,[125] they apparently do not feel the same about their scholarship. Or, at the very least, telltale signs of their disinterest abound. For one, as we noted above, much, if not most, of legal scholarship violates the replication rule, without which most advantages of an intellectual community are pointless. Law scholars may be fastidious about documenting textual sources of information via the omnipresent footnote,[126] but they are not particularly attentive to the need to document their data procedures, nor, we hasten to note, have they established procedures for ensuring the requisite attention or repositories, private or public, for their data. Such lapses stand in marked contrast to many cognate disciplines. In Political Science, for example, the flagship journal, the *American Political Science Review*, specifies that

> Authors should describe their empirical research in sufficient detail to permit reviewers to understand and evaluate what has been done and, in the event the article is accepted for publication, to permit other scholars to carry out similar analyses on other data sets. For example, for surveys, at the least, sampling procedures, response rates, and question wordings should be given; authors are encouraged to calculate response rates according to one of the standard formulas given in The American Association for Public Opinion Research, *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for RDD Telephone Surveys and In-Person Household Surveys*…For experiments, provide full descriptions of experimental protocols, methods of subject recruitment and selection, information about any payments and debriefing

[124] Eisenberg, Garvey, & Wells, Jury Responsibility in Capital Sentencing, supra note 71.

[125] See infra note 132.

[126] We do not mean to condemn or denigrate the use of footnotes. Indeed, because they connect the extant scholarship to existing literatures, they are one (of the few) sign(s) that legal scholars recognize the importance of developing a community of scholars. We return to this point on infra page 91.

procedures, and other relevant details. It is desirable for articles to be self-contained, and authors should not refer readers to other publications for descriptions of these basic research procedures.[127]

Political Analysis, a leading empirical methods journal, has an even more stringent set of replication requirements:

> Authors of quantitative articles in *Political Analysis* must indicate in their first footnote in which public archive readers can find the data, programs, recodes, or other information necessary to replicate the numerical results in their article. Authors may find the "Publication-Related Archive" of the Inter-university Consortium for Political and Social Research (ICPSR) a convenient place to deposit their data. This replication dataset, along with any other supporting material you wish to submit (such as appendices, supplementary analyses, interactive web tutorials, etc.), will also be published on the *Political Analysis* website. Authors who prefer a period of embargo before public release should consult with the editor. Authors of works relying upon qualitative data are encouraged to submit a comparable footnote that would facilitate replication where feasible. As always, authors are advised to remove information from their datasets that must remain confidential, such as the names of survey respondents.

Moreover, many social scientists either devise their own storehouses for data sets they have created or deposit data and documentation in public archives.[128] These work to ensure compliance with journal policies, such those in the *American Political Science Review* and *Political Analysis,* as well as with a condition the National Science Foundation imposes on all grants: "All data sets produced with the assistance of this award shall be archived at a data library approved by the cognizant Program Officer, no later than one year after the expiration date of the grant." Numerous other journals and granting agencies throughout the physical, natural, and social sciences have similar policies but few exist in the field of law.

Yet another sign of indifference to this guideline is the lack of collaboration in legal scholarship. While one can recognize that empirical research is a social enterprise without collaborating on particular research projects, many branches of the scholarly community acknowledge the value of and rewards in joint enterprises. For example, of the 52 grants awarded by the Political Science division of the National Science Foundation in 2000, 60% (n=31) were given for collaborative projects.[129] An analysis of three leading disciplinary journals, the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*, concludes that "the percentage of multiple-authored articles, in aggregate, has increased seven-fold since the 1950s and almost one-half of the articles are now multiple-authored."[130] A broader study by the National Science Foundation finds that across the fields of science and engineering (S&E), "The proportion of U.S. scientific and technical articles with multiple institutional authors has continued to rise. In 1997, 57 percent of all S&E articles had multiple authors, up from 49 percent a decade earlier."[131] By contrast, only *5%* of

---

[127] Ada Finifter, "Editor's Notes," 94 Am Pol Sci Rev viii (2000).

[128] "The world's largest archive of computerized social science data" is the Inter-university Consortium for Political and Social Research (ICPSR), access to which is available via http://www.icpsr.umich.edu . Micah Altman, Gary King, and Sidney Verba have undertaken a project to computerize and automate access, authentication, documentation, subsetting, conversion, and other aspects of data distribution. For more information, see http://TheData.org .

[129] Data are available at: http://www.fastlane.nsf.gov/servlet. The figure of 52 excludes grants for doctoral work; the term "collaborative" includes grants for "collaborative research," conferences, and infrastructure.

[130] Bonnie S. Fisher, et al. 21 P.S.: Political Science and Politics (1998) 847, available at http://www.apsanet.org/PS/dec98/fisheretal.cfm.

[131] See U.S. National Science Foundation, Indicators 2000 in Science and Engineering, at http://www.nsf.gov/sbe/srs/seind00/access/c6/c6s4.htm.

the 162 articles published in six of the leading law journals (*Chicago, Columbia, Harvard, NYU, Stanford,* and *Yale)* in 2000 were the product of a collaborative effort.[132]

A third sign is the legal community's refusal to subject articles submitted to some of its most prestigious outlets to any form of blind peer review, preferring instead to leave the refereeing task to law students. Most scholars in units outside of law schools are, when they hear of it, astonished at this organizational decision.[133] For they have come to learn that it is easy to fool oneself (or law students, as the case might be) into believing that one has produced an important research result; it is a good deal more difficult to "fool," however inadvertently, a community of experts spending their lives working on related problems. That is why the most prestigious journals in virtually all other academic fields are peer reviewed and often double-blindly reviewed (*i.e.,* neither the author nor the reviewer knows the other's identity) at that. This last feature reduces the possibility that the authors' status or reputation will become a part of the process of evaluating their work. Which is exactly what scholars have accused student editors of weighing heavily in their selection of articles.[134]

We could identify other signs but it is the general point that should not be missed: Lack of recognition that research is a social enterprise is a problem throughout law schools. Since this is highly wasteful of the efforts of legal scholars, we devote the conclusion of this Article to offering suggestions on how law schools and their faculties can begin to move to alignment with the rest of the academy and benefit from building a scholarly community. As we explain, the unique collective norms of the legal community may make it possible to catch up quickly and possibly even exceed the research standards existing in the rest of academia.

3. All Knowledge and All Inference in Research is Uncertain.

Toward the end of their article, Perrin and his colleagues state unequivocally

> our study confirms [that] the exclusionary rule does not effectively deter police misconduct. Indeed, in the area of police deception, it has fostered misconduct.[135]

Fisher's article on the growth of plea bargaining leads him to the following:

> Supported only by the desire of prosecutors to manage their crushing workloads and to gain an occasional effortless conviction, plea bargaining extended no further than the sentencing power of prosecutors….By the middle of the nineteenth century, plea bargaining had stolen to a larger outpost, the on-file form of plea bargaining, left exposed to prosecutors by the procedural fluke that judges could not pass sentence until the prosecutor so moved. There the progress of plea bargaining might have stalled, for the sentencing power of prosecutors reached little further.

---

[132] We conducted this count on December 4, 2000. The figure of 162 excludes book reviews, tributes, case notes, review essays, developments, and articles associated with the Harvard Law Review's annual review of the Supreme Court's term. While these data reveal that law professors do not often coauthor scholarly articles, it is clear that they recognize the value of collaboration—at least with regard to teaching materials. As several legal academics pointed out to us, it (apparently) is the rare treatise or case book that is *not* the product of a collaboration.

[133] See, for example, Wills, "To Keep and Bear Arms."supra note 20; Friedman, Law Reviews and Legal Scholarship: Some Comments, supra note 25, at 661.

[134] See, for example, Bernard J. Hibbitts, Yesterday Once More: Skeptics, Scribes and the Demise of Law Reviews, 30 Akron Law Review 267, 292 (1996); Lindgren, "An Author's Manifesto," supra note 44 ; Nathan H. Saunders, "Student-Edited Law Reviews: Reflections and Responses of an Inmate," 49 Duke L J 1663 (2000); Jordan H. Leibman & James P. White, "How the Student-Edited Law Journals Make their Publication Decisions," 39 J Legal Educ 387 (1989); Banks McDowell, "The Audiences for Legal Scholarship," 40 J Legal Educ 261 (1990). See also Richard A. Posner, "The Future of the Student-Edited Law Review," 47 Stan L Rev 1131 (1995) approving of law review editors taking into account the author's reputation and the prestige of his or her law school.

[135] L. Timothy Perrin, et al., If It's Broken, Fix It: Moving Beyond the Exclusionary Rule, 83 Iowa L Rev 736 (1998).

> Then, in the last quarter of the nineteenth century, judges found themselves confronted by an onslaught of new, and newly complex, civil suits brought on by the ravages of industrial machinery. They saw no choice but to make terms with the new order in the criminal courts. They embraced plea bargaining and turned their considerable sentencing power to its purpose. Sustained now by the two most powerful courtroom patrons, plea bargaining swiftly became the dominant force in criminal procedure. It pushed aside the indeterminate sentence, and it supported those institutions, such as probation and the public defender, that aided its cause. Finally, plea bargaining grew so entrenched in the halls of power that today, though its patrons may divide its spoils in different ways, it can grow no more. For plea bargaining has won.[136]

Shaffer, in his analysis of the reach of the Privileges or Immunities Clause of the Fourteenth Amendment, claims that the evidence points conclusively to the following:

> The Clause does not offer a justification for enshrining any substantive rights other than those clearly and undisputedly emblazoned on the United States' collective constitutional psyche.[137]

Even if all the authors of these quotes followed all other suggestions we offer governing empirical research (which they do not[138]), drawing conclusions with the degree of certainty displayed in their articles would still be unjustified. For a basic premise of all empirical research—and indeed of every serious theory of inference—is that all conclusions are uncertain to a degree. After all, the facts we know are related to the facts we do not know but would like to know only by assumptions that we can never fully verify.

The point is not to qualify every statement—*e.g.,* by changing "I am certain" to "I am fairly certain"—but rather to *estimate the degree of uncertainty* inherent in each conclusion and to report this estimate along with every conclusion. Most statistical procedures come with formal measures of uncertainty.[139] If the research is qualitative in nature or if it is not obvious how to estimate uncertainty, however, one useful course of action is to find the weakest link in the chain of reasoning: the part of the argument that rests on the least empirical evidence or that is most vulnerable to attack. In other words, *identify the "smallest" piece of evidence the researcher has compiled, that, if changed, should cause the reader or the researcher to surmise that the conclusion reached in the study is wrong.* The degree of support that can be mustered for this piece of evidence is one measure of the uncertainty of the conclusions.

---

[136] George Fisher, "Plea Bargaining's Triumph," 109 Yale L J 857 (2000), at 1074-1075.

[137] Derek Shaffer, "Answering Justice Thomas in Saenz: Granting the Privileges or Immunities Clause Full Citizenship Within the Fourteenth Amendment," 52 Stan L Rev 709 (2000).

[138] As for the problems in Perrin, et al., If It's Broken, Fix It, supra note 135, which bases its conclusion on a survey of police officers in Ventura County, California, see Heise, The Importance of Being Empirical, supra note 2, at 831. Sisk, Heise, & Morriss, Charting the Influences on the Judicial Mind, supra note 27, which explores two-hundred and ninety-four U.S. district court cases involving the constitutionality of the Federal Sentencing Guidelines, violates several rules. For an example, see infra p. 40. Fisher, "Plea Bargaining's Triumph," supra note 136, among other problems, fails to specify the target of his inference (see our discussion on infra p. 71). Shaffer, Answering Justice Thomas in Saenz, supra note 137, at 750 premises his conclusion on qualitative evidence "sift[ed]" from the "relevant history" pertaining to the intent of the framers of the Fourteenth Amendment. Despite the author's admission that "the Clause's framing is shrouded and blurred beneath a veil of confusion, obfuscation, and question-begging," he provides no information about his procedures for "sifting" though the evidence. His work thus suffers from the same problem as Engel, The McCulloch Theory of the Fourteenth Amendment, supra note 118. See infra p. 30.

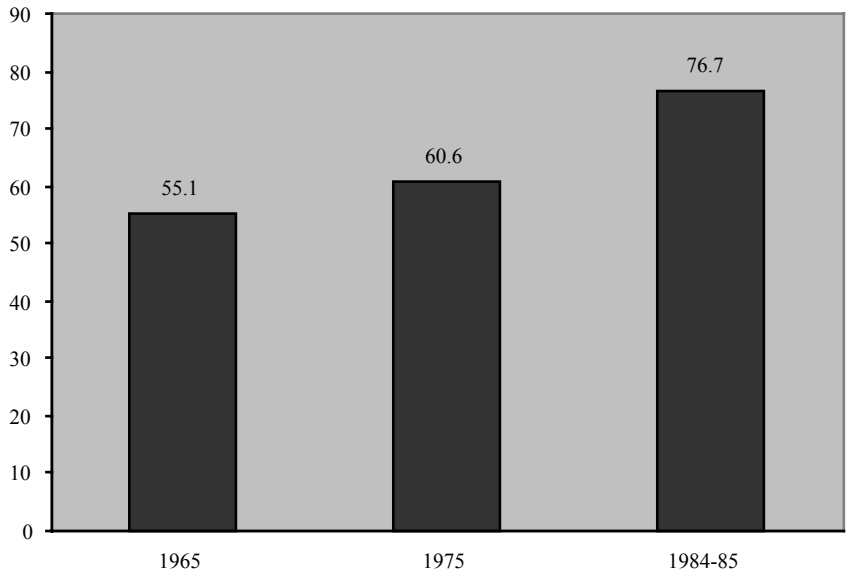
[139] For example, a common measure of uncertainty conveyed in surveys reported in the popular media is the "margin of error," which is (usually) a 95% confidence interval. So, when a newspaper reports the result of a survey—say, that 75% of the respondents prefer dogs to cats with a ±5 margin of error—it is supplying the level of uncertainty it has about the 75% result (here, that the true fraction of people preferring dogs to cats will be captured within the stated confidence interval in 95 of 100 applications of the same sampling procedure).

To see this, consider a study by Schuck and Elliot[140] that seeks to appraise the following piece of conventional wisdom: Courts were "deferential" to administrative agencies in the 1960s, became less so during the "hard look" era of the 1970s, and then settled back to a moderately deferential stance in the 1980s. To assess it, the investigators consider the outcomes in appellate court cases decided in 1965, 1975, and 1984-85. Figure 2 depicts their findings on the percentage of cases courts affirmed (that is, the percentage that deferred to the agency). The data, as Schuck and Elliot note, seem "to contradict conventional wisdom," which would predict a lower rate of affirmances in 1975 than in 1965 or 1984-85.[141] But how certain can we, or they, be of this conclusion?

[140] Schuck & Elliot, To the Chevron Station, supra note 13.

[141] Id., at 1008

Figure 2. Percentage of Affirmances Reported in Schuck and Elliot[142]



To address this question we begin by noting that, although this is a quantitative study, many qualitative factors go into producing its numbers.[143] In particular, the authors made a large number of coding decisions that have varying levels of justification. Let us think about but one result of those decisions, a small piece of evidence, say the 1965 data, and ask what we could conclude if the figure was not 55.1% but 75%. (We chose 75% because public data from the Administrative Office of the U.S. Courts, as Schuck and Elliot themselves point out, match their data for 1975 and 1984-85, but depart by precisely this amount from their 1965 data.[144]) If the number was in fact 75%, we might think the conventional wisdom got it right and Schuck and Elliot were wrong. At the very least, we would question whether the researchers ought reach definitive or even near-definitive conclusions since the uncertainty surrounding the result turns out to depend entirely on the validity of this one coding decision and the single number resulting. This is not a formally quantified measure of uncertainty, but we, the readers, can peruse the authors justification and decide for ourselves.

We can apply the same logic to Schultz and Petterson's examination of how courts have responded to the "lack of interest" defense in Title VII race and sex discrimination cases.[145] Table 3 displays the answer they report.

---

[142] Id., at 1008.

[143] See Herbert Kritzer. 1994. "Interpretation and Validity Assessment in Qualitative Research: The Case of H. W. Perrys Deciding to Decide," Law & Social Inquiry 19, 687-724; Herbert Kritzer. 1996. "Data, Data, Data, Drowning in Data: Crafting The Hollow Core," Law & Social Inquiry 21, 761-804.

[144] Id., at 1009.

[145] Vicki Schultz & Stephen Petterson, "Race, Gender, Work, and Choice: An Empirical Study of the Lack of Interest Defense in Title VII Cases Challenging Job Segregation," 59 U Chi L Rev 1073 (1992).

Table 3 Plaintiffs' Success Rates in Cases Addressing the Lack of Interest Defense by Race and Sex Over Time, as Reported in Schultz and Petterson[146]

Period	Race		Sex	
	% Success	Number of Cases	% Success	Number of Cases
1967-1977	86.0	43	54.5	11
1978-1989	40.0	20	58.1	43
Total	71.4	63	57.4	54

These data, they conclude, reveal a "striking trend"; namely, "race discrimination plaintiffs have become substantially less likely to prevail on the lack of interest argument since the late 1970s."[147] Accordingly, they devote a healthy portion of the article to explaining this "trend."

But is it really so striking? Perhaps not, for we can radically alter the conclusion reached by Schultz and Petterson by (1) assuming that all twelve losses in the 1978-1989 period actually occurred between 1978 and 1980 and (2) moving those losses into a reconfigured data category (1967-1980). Table 4 is thus possible (although contrived via extreme, but still plausible, assumptions) given the information provided in the article. Unfortunately, since the data are not presently available either publicly or from the authors,[148] we do not know for sure. What we do know is that it now appears, as the table shows, that race plaintiffs have become more likely to prevail over time, not less. Thus, the validity of Schultz and Petterson's conclusions, and the degree of certainty that readers should accord their results, rests entirely on our confidence that an alternative aggregation or presentation of their data would not similarly reverse their results.

Table 4 Plaintiffs' Success Rates in Cases Addressing the Lack of Interest Defense by Race and Sex Over Time: A Hypothetical Regrouping of the Schultz and Petterson Data, Consistent with Table 3

Period	Race		Sex	
	% Success	Number of Cases	% Success	Number of Cases
1967-1980	67.2	55	54.5	11
1981-1989	100.0	8	58.1	43
Total	71.4	63	57.4	54

The studies we have considered so far rely primarily on numerical data but the same general lesson applies to those of a more qualitative nature. Recall the Shaffer article—reaching the "certain" conclusion that the framers did not intend the Privileges or Immunities Clause to serve as a device "for weaving new patterns into the text of our Constitution"[149]— and consider one piece of evidence the author invokes to support this claim: the comments made by members of the joint congressional committee who drafted the Clause. Now suppose that the chair of that committee, John Bingham, instead of delivering a speech (as Shaffer summarizes) "elaborating…his…stance that the Amendment as a whole, and the Clause in particular, would expressly secure rights already

[146] Id., at 1097.
[147] Id., at 1098.
[148] Email communication, Vicky L. Schultz to Gary King, on 3 November 2000 (on file with the authors).
[149] Shaffer, Answering Justice Thomas in Saenz, supra note 137, at 750.

tacitly understood to exist but otherwise subject to violation by individual states,"[150] stated the following: "The Amendment as a whole, and the Clause in particular, should expressly secure rights already tacitly understood to exist but otherwise subject to violation by individual states," *as well as new or evolving substantive rights.* Surely that one change might have caused Shaffer to be a bit more circumspect. And just as surely for us, the consumers, the extent to which Bingham might plausibly have made this alternative statement is one way to calibrate the degree of uncertainty in the author's conclusions.

III. Designing Empirical Research: A Dynamic Process Conforming to Fixed Standards

Whether the researcher is relying on numerical or non-numerical evidence, rules exist that can improve each component of the research design—Research Questions, Theories and their Observable Implications, Rival Hypotheses, Measurement and Estimation, and Selecting Observations—and we articulate them in the five Parts that follow. Before so doing, though, two general comments are in order. First, researchers should not regard their design as laying out a singular, mechanical process from which they can never deviate. Quite the opposite: Scholars must have the flexibility of mind to overturn old ways of looking at the world, to ask new questions, to revise their blueprints as necessary, and to collect more (or different) data than they might have intended. It may be that, after amassing the evidence for which the design calls, the scholar finds an imperfect fit among it, the main research questions, and the theory. Rather than erasing months, even years, of work, the investigator certainly ought return to the drawing board, design more appropriate procedures, or even recast the original research question. Indeed, often when researchers find that data turn out to be inconsistent with a hypothesis, they immediately see a new hypothesis that apparently explains the otherwise anomalous empirical results.[151]

Recognizing this new way of looking at the facts is what we mean by being flexible; what we do not mean is that researchers ought make ad or post hoc adjustment of theories to fit idiosyncrasies. We must not fool ourselves into thinking that adjustments made to harmonize theory with data constitute any confirmation of the theory at all. They do not. Confirmation requires vulnerability; continuous post hoc adjustments to theory ensure that the theory is never vulnerable to being wrong. *While using insights from data is a good way to develop theory, investigators should consult a new data set, or different and previously unanticipated testable consequences of the theory in the same data set, before concluding that data confirm their theory.*

This is why we view the conduct of empirical research as a dynamic process of inquiry occurring within a stable structure of rules. Nonetheless, and this takes us to the second point: Investigators should work to improve their blueprints before they actually conduct a full-scale study. Since even the best designs occasionally fall apart after the researcher collects the very first few data, we recommend that legal academics work sequentially: amass that very first piece and consider whether it comports with the research question and the hypotheses. If it does not, they may want to rethink their expectations but, more likely, they will come to see that the type of evidence they intended to collect is not as well suited to answering the research question as they anticipated.

---

[150] Id., at 722.

[151] We adopt and adapt some of the sentiments in this section from King, Keohane, & Verba, Designing Social Inquiry, supra note 1 at 12, 22.

IV. The Research Question

We assume that most readers of this Article have research questions they would like to answer, questions they would like someone else to answer, or an answer to a question they would like to evaluate. Hence, we need not say too much about how legal academics go about developing research questions.[152] What requires explication are the two criteria that better research questions meet: *they contribute to existing knowledge* and *they have some importance for the real world*, each of which is the subject of a section to follow.[153]

As important as these criteria are, we recognize that many questions asked by academics and others about legal phenomena do not meet these standards. On the one hand, this is not particularly problematic: Investigators can conduct rigorous empirical research about any question, no matter how narrow it may be, no matter whether they are the only ones interested in it, no matter if it has virtually no implications for the real world. On the other hand, analysts will be able to answer questions better if they can motivate members of their community—whether other scholars, decision makers, or both—to take an interest in their research. The reason goes back to a point we made in Part I: Empirical research is a social enterprise. An individual researcher is much more likely to accomplish even his or her narrow goals when posing research questions in ways that attract the interest of others. Working on such projects is good career advice too, but one of the reasons this advice is usually given in the first place (and the reason it should continue to be given) is to foster the development of a research community so that we can all better accomplish our goals.

To put it succinctly, academics and others can choose to ignore the two rules that follow; they can focus on a very narrow question that holds interest to few or, even, no others. And, if they choose to do so, complying with the other rules in this Article will help them to answer it better. But following the two below has the added benefit of making the research more reliable and its results, more certain.

A. Contribute to a Scholarly Literature

While many legal scholars already understand the importance of this guideline, we realize that it will cause at least some to bristle. "Isn't it sufficient," they might ask, "that we address real-world problems?" For at least four reasons the answer is no. One—that empirical research is a social enterprise—we need not say too much more about. But the other three require elaboration.

First, finding a way to participate in the social enterprise of scholarship minimizes the chances that informed and, perhaps, even uninformed, readers will question whether the researcher is up on the "state of the art" in the particular area under analysis. In other words, compliance with this rule enhances the credibility of the research. But more importantly, making connections with what has come before helps scholars avoid mistakes, skip arduous reinventions of existing ideas, and find additional observable implications of their theories. Consider a 1998 *New York University Law Review* article claiming "thus far public choice theory has had relatively little to say about judges' behavior in deciding cases…"[154] Given the scores of articles and books predating 1998 that invoked public

---

[152] For more information on the development of research questions in this area, see Lee Epstein, Studying Law and Courts, in Contemplating Courts, ed. Lee Epstein (1995).

[153] We adopt these two suggestions from King, Keohane, & Verba, Designing Social Inquiry (1994), supra note 1, at 15.

[154] Sisk, Heise, & Morriss, Charting the Influences on the Judicial Mind, supra note 27, at 391.

choice theory (or some variant thereof),[155] this statement is misleading to others and probably damaging to the research I this article. In a certain sense, citations should be as irrelevant as the identities of the authors—unless it has a consequence. Which it usually does: Although the authors of this particular piece are well-integrated into the academic community and do not often make the same mistakes in their other work or even in other parts of the same work, on this issue they are effectively working in isolation from a vast literature that could help them accomplish their goals. Rosenberg gives a related example from research produced by law professors on public opinion: "legal academics continually make claims about the ability of the judicial system to affect public opinion, often with an approving cite to Rostow or Bickel. There is an empirical literature on public knowledge of judicial opinions that doesn't support these claims, however—but it is never cited."[156]

This general problem would seem to apply to much research published in the law reviews. Or so some scholars in other fields claim, and often in the strongest terms[157]—with Baer's comment typical: "I swear, if I have to read one more book by a law professor which ignores a whole list of relevant works by political scientists…"[158] Graber, a social scientist with a law degree, concurs: "One would never know from the [legal scholarship]…that there has been a flood of literature in [the social] sciences on constitutional theory, doctrine, history and politics.[159] While, as Graber points out, it would be difficult, if not impossible, for social scientists who write on American constitutionalism to publish work "that displayed the same ignorance of developments in academic law as the giants of academic law routinely display of developments among those political scientists who study constitutionalism,"[160] the converse does not hold. It should. Even more to the point, in

---

[155] A small sample includes William N. Eskridge, Jr., "Reneging on History?: Playing the Court/Congress/President Civil Rights Game," 79 Cal L Rev 613 (1991); William N. Eskridge, Jr., "Overriding Supreme Court Statutory Interpretation Decisions," 101 Yale L J 331 (1991); John Ferejohn & Barry Weingast, "Limitation of Statutes: Strategic Statutory Interpretation," 80 Georgetown Law Review 565 (1992); John Ferejohn & Barry Weingast, "A Positive Theory of Statutory Interpretation," 12 Intl Rev L & Econ 263 (1992); Jack Knight & Lee Epstein, "On the Struggle for Judicial Supremacy," 30 Law & Soc Rev 87 (1996); Frank H. Easterbrook, "Ways of Criticizing the Court," 95 Harv L Rev 802 (1982); Segal, "Separation-of-Powers Games in the Positive Theory of Law and Courts," supra note 63; Walter F. Murphy, Elements of Judicial Strategy (1964); Erin O'Hara, "Social Constraint or Implicit Collusion?: Toward a Game Theoretic Analysis of Stare Decisis," 24 Seton Hall L Rev 736 (1993); Richard A. Posner, "What Do Judges and Justices Maximize? (The Same Thing Everybody Else Does)," 3 S Ct Econ Rev 1 (1993); Maxwell L. Stearns, "Standing Back from the Forest: Justiciability and Social Choice," 83 Cal L Rev 1309 (1995); Pablo T. Spiller & Rafael Gely, Congressional Control or Judicial Independence: The Determinants of U.S. Supreme Court Labor-Relation Decisions, 23 RAND Journal of Economics 463 (1992); Forrest Maltzman & Paul J. Wahlbeck, Strategic Policy Considerations and Voting Fluidity on the Burger Court, 90 Am Pol Sci Rev 581 (1996); Forrest Maltzman & Paul J. Wahlbeck, "May It Please the Chief? Opinion Assignments in the Rehnquist Court," 40 Am J Pol Sci 421 (1996); McNollgast, "Legislative Intent: The Use of Positive Political Theory in Statutory Interpretation," 57 L & Contemp Probs 3 (1994); McNollgast, "Politics and Courts: A Positive Theory of Judicial Doctrine and the Rule of Law," 68 S Cal L Rev 1631 (1995).

For a review of early work, see Lee Epstein & Jack Knight, "Toward a Strategic Revolution in Judicial Politics: A Look Back, A Look Ahead," 53 Political Research Quarterly 625 (2000); for a collection of essays, see Maxwell L. Stearns, Public Choice and Public Law (1997). See also infra note 193, p. 45, and p. 47, where we explain that (and provide examples of how) scholars have made use of positive political theory, which belongs to a particular class of rational choice models, to study judicial decisions.

[156] Rosenberg, Across the Great Divide (Between Law & Political Science), supra note 21, at 268.

[157] Some legal scholars concur. Frank B. Cross, "Political Science and the New Legal Realism: A Case of Unfortunate Interdisciplinary Ignorance," 92 Nw U L Rev 251 (1997), for example, notes that "legal scholarship has been remarkably oblivious to [the] large and mounting political science literature on courts." See also Tracey E. George, "Development of a Positive Theory of Decisionmaking on U.S. Courts of Appeals," 58 Ohio St L J 1635 (1998)

[158] Judith A. Baer, posting on the Law and Courts List Serv, September 23, 1998. Cited in Rosenberg, Across the Great Divide (Between Law & Political Science), supra note 21, at 269.

[159] Mark Graber, posting on the Law and Courts List Serv, July 21, 2000 (on file with the authors).

[160] Id. Stephen L. Wasby, posting Law and Courts List Serv, July 21, 2000 (on file with the authors) followed up with this observation: "There are more law professors, writing in law reviews, who are discussing the factors affecting judges'

this day and age of LEXIS, Westlaw, and other means of Internet accessibility to journals in a broad array of disciplines,[161] and the growing community of (knowledgeable) scholars who study law-related topics, research in isolation is hard to justify.

A second advantage of engaging relevant scholarly literatures is that it decreases the chances of duplicating work already done, of "reinventing the wheel." This is not to say that scholars should necessarily avoid raising the same questions as others, reanalyzing the same data, or pursuing new ways of looking at the same problems or bringing new data to bear on them. It is rather to say that, if they are addressing existing questions, they should take into account the lessons of past studies. Failure to do so is more than wasteful; it also decreases the odds that the "new" research will be as successful as the original because the researcher is, in effect, ignoring the collective wisdom gained from that first piece.

To see how this can happen, consider a study by Perrin et al. in which researchers conduct a survey of police officers in Ventura County, California with the intent of answering the following question: What are the "effects and costs of the exclusionary rule"?[162] To be sure, the authors acknowledge the existing, clearly identifiable body of literature that has bearing on their question.[163] This is all to the good since even perfunctory compliance with our suggestion that research ought contribute to scholarly literature deflects the kind of criticism we leveled at the 1998 *New York University Law Review* study.[164] But it is insufficient. Full compliance requires researchers to take into account the lessons of past studies, both their assets and deficits, in their own endeavor. And it was on this dimension that the Perrin team could have gone much farther. In reporting the results of what was, by their own admission, "plainly the most thorough study of the [exclusionary] rule,"[165] they noted that, Oaks, the author of the study, was loathe to draw comparisons between evidentiary suppression motions in Washington, D.C. and Chicago because he believed that "important differences in the criminal justice systems of the two cities [exist], differences so striking that meaningful comparisons could not be made."[166] But the moral of the Oaks study—that "the same or different characteristics that distinguished the criminal justice systems in Washington, D.C. and Chicago could also distinguish Ventura County from all other American jurisdictions"[167]—was missed by the Perrin team. At the end of their study they had no hesitation about moving beyond

---

votes. Yet in a distressing number of instances, the only citations are to other law review articles—not to the political science literature, which is far more extensive and is original; the law review articles cited are usually at best derivative. What I am saying is that the law professors' failure to read relevant political science 'law and courts' literature does not occur only with respect to matters of constitutional law and constitutional interpretation but is broader."

To be sure, there are exceptions to Wasby's general concern. See, for example, Frank B. Cross & Emerson H. Tiller, "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals," 107 Yale L J 215 (1998); Cross, "Political Science and the New Legal Realism: A Case of Unfortunate Interdisciplinary Ignorance," supra note 157; George, "Development a Positive Theory of Decisionmaking on U.S. Courts of Appeals," supra note 157. But his general claim probably holds more often than it does not.

[161] *E.g.,* J-Stor (http://www.jstor.org) contains the full text of journals in the following disciplines: African American Studies, Anthropology, Asian Studies, Ecology, Economics, Education, Finance, General Science, History, Literature, Mathematics, Philosophy, Political Science, Population Studies, Sociology, and Statistics. Stanford University's HighWire Press now offers 211 journals on line, and OCLC, scores more.

[162] Perrin, et al, If It's Broken, Fix It, supra note 135, at 673.

[163] For examples, see Id., at 678.

[164] Sisk, Heise, & Morriss, Charting the Influences on the Judicial Mind, supra note 27, at 391.

[165] That study is Dallin H. Oaks, "Studying the Exclusionary Rule in Search and Seizure," 37 U Chi L Rev 665 (1970).

[166] Perrin, et al., If It's Broken, Fix It, supra note 135, at 696.

[167] This quote comes from Heise, The Importance of Being Empirical, supra note 2, at 833, who, in his critique of the Perrin et al. study, highlights the problem we mention in the text.

their case, claming that they had confirmed "both the rule's lack of value as a deterrent and the high costs the rule imposes on society and the system."[168]

Finally, following the advice that research engages existing scholarship ensures that someone will be interested in the results. After all, if a body of literature—however slim and underdeveloped it may be—is on hand, it indicates that the question is important to at least some others. By interesting others, researchers benefit those others, as well as increase the chances that another or other investigator(s) will examine their research question, reevaluate their evidence from a new angle, or introduce new evidence of a closely related problem, the result being more certain knowledge about communal concerns.

This, of course, does not mean that scholars ought necessarily ask precisely the same question as others; it simply suggests that their research should contribute to, make connections with, or attempt to interest others in a specific area of inquiry. And that contribution can come in any number of ways: (1) asking a question that the legal community may view as important but that no other scholar has tackled; (2) attempting to settle a question that has evoked conflicting responses; (3) raising an "old" question but addressing it in a unique way; (4) collecting new data on the same observable implications or different implications altogether;[169] or (5) applying better methods to reanalyze existing data.[170]

B. Conduct Research Important to the World

This is a rule about which we need not say too much. For, of all those we set out, this is one that many legal scholars already understand, and may even understand better than some of their counterparts in the sciences and social sciences. Our surveys of law review articles suggest that it is the rare piece of legal scholarship that does not pose a question that has at least a *potential* implication—normative, policy, or otherwise—for the real world.[171] Indeed, the conclusions of many take pains to spell out the nature of those implications, typically in the form of future paths that courts, attorneys, or legislators should follow.

To arrive at these sorts of recommendations, the range of research questions legal academics can (and do) raise is quite broad. Some center on "law in books," asking whether a court reached a decision "correctly" (with "correctly" variously defined as in line with existing precedent, legislative intent, the plain meaning of the text, and so on). Levinson's essay on the Second Amendment supplies an example.[172] In it he asks whether court decisions holding that the Amendment

---

[168] Perrin, et al., If It's Broken, Fix It, supra note 135, at 755.

[169] We take up the subject of observable implications in Part V.

[170] See also King, Keohane, and Verba, Designing Social Inquiry, supra note 1 at 16-17.

[171] Milhaupt & West, The Dark Side of Private Ordering, supra note 77, at 74, is exemplary. Believing that the data support their hypothesis—"organized crime is an entrepreneurial response to institutional shortcoming"—they suggest that governments seeking to combat organized crime should "direct their resources not at crime control per se, but at creating, or facilitating, proper property-rights-enforcement institutions." More generally, see Rubin, The Practice and Discourse of Legal Scholarship, supra note 25, at 1847 (making the point that a "distinctive feature" of legal scholarship is its prescriptive voice, its consciously declared desire to improve the performance of legal decision-makers). Rubin refers primarily to doctrinal studies, many of which, as we note supra p. 2, fall within the purview of our concerns.

Of course it is true that some judges and practitioners do not think that the implications from certain types of empirical work (as we have defined "empirical") are particularly important or relevant to their work. See, for example, Edwards, The Growing Disjunction between Legal Education and the Legal Profession, supra note 51; Posner, Madison Lecture, supra note 11. But this does not mean that implications for the real world do not exist.

[172] Levinson, "The Embarrassing Second Amendment," supra note 20.

establishes only a collective right comport with the text of the Amendment, the historical circumstances surrounding its adoption, and the structure of the Constitution.[173]

Other questions center on "law in action," asking whether a particular court decision or law had the effect its creators or others anticipated. Such was Hightower's research[174] on the impact of *J.E.B.* v. *Alabama*.[175] In his dissent in that case, Justice Scalia fretted that the extension of *Batson v. Kentucky*[176] to sex "will provide the basis for extensive collateral litigation, which especially the criminal defendant (who litigates full time and cost free) can be expected to pursue. While demographic reality places some limit on the number of cases in which race based challenges will be an issue, every case contains a potential sex-based claim."[177] What Hightower asks was whether Scalia's concerns materialized: Did *J.E.B.* generate widespread collateral litigation?

These represent just two of the sorts of questions legal academics raise that fall well in line with the rule that research should be important to various internal and external constituencies. Levinson makes this clear at the end of his article where he writes: "For too long, most members of the legal academy have treated the Second Amendment as the equivalent of an embarrassing relative… That will no longer do. It is time for the Second Amendment to enter full scale into the consciousness of the legal academy."[178] Subsequent events, including a proliferation of law review articles on the Second Amendment[179] and the adoption of his argument by at least one court[180] not only have met Levinson's concern. They also underscore the importance of his question (if not the credibility and certainty of his inferences[181]) for academics, as well as for others in the legal community.

V. Theories and Their Observable Implications

Once a scholar has a research question that engages the scholarly literature and is important in the real world, it is constructive to begin *theorizing* about possible answers, which she or he can, in turn, use to generate *observable implications* (also called expectations or hypotheses). By "theorizing" we mean developing *"a reasoned and precise speculation about the answer to a research question"*;[182] by "observable implications," we mean things that we would expect to detect in the real world if our theory is right.

There is nothing magical or mystical about these activities; in fact, we engage in them every day. After we teach the first few sessions of a class, we might develop a simple theory, say, that the students in it are better than those we taught the year before. Observable implications of this

---

[173] Id., at 643.

[174] Susan Hightower, "Sex and the Peremptory Strike: An Empirical Analysis of J.E.B. v. Alabama's First Five Years," 52 Stan L Rev 895 (2000)

[175] 511 U.S. 127 (1994) (prohibiting gender-based peremptory strikes).

[176] 476 U.S. 79 (1986) (prohibiting race-based peremptory strikes).

[177] *Batson,* at 162.

[178] Levinson, The Embarrassing Second Amendment, supra note 20, at 658.

[179] For a sample, see supra note 20.

[180] See supra note 20.

[181] See supra note 20.

[182] King, Keohane, & Verba, Designing Social Inquiry, supra note 1, at 19. We highlight this definition because legal academics seem to define "theory" differently than those in other disciplines. Friedman, Law Reviews and Legal Scholarship, supra note 25, at 668 makes this point when he writes, "In legal scholarship, 'theory' is king. But people who talk about legal 'theory' have a strange idea of what 'theory' means. In most fields, a theory has to be testable; it is a hypothesis, a prediction, and therefore subject to proof. When legal scholars use the word 'theory,' they seem to mean (most of the time) something they consider deep, original, and completely untestable." Our view of the role of theory in empirical research comports with those used in "most [other] fields."

"theory" are easy enough to summon: we might expect the students to perform unusually well on examinations, to write especially cogent essays, or to say notably smart things in class.

Theorizing in scholarship is not all that much different, though it can and does take many different forms. Some theories are in fact simple, small, or tailored to fit particular circumstances, and these abound in the law reviews. In his study of how judges apply the two-step *Chevron* test for judicial review of agency decisions, for example, Kerr offers a "contextual" theory.[183] That theory, at least as he framed it, is quite specific: Judges continue to use "traditional" factors in adjudicating *Chevron* cases, rather than the two-step test.[184] Along the same lines, legal academics often develop theories about the intent of the framers over particular laws ranging from the Freedom of Information Act[185] to the National Labor Relations Act[186] to the Foreign Trade Zones Act[187] to the Sherman Act[188] or particular constitutional provisions including those on impeachment,[189] the Establishment Clause of the First Amendment,[190] the Search and Seizure and Warrant Clauses of the Fourth Amendment,[191] and the Thirteenth Amendment.[192]

Other theories are grander in scope, seeking to provide insight into a wide range of phenomena. An increasingly common one in legal scholarship is positive political theory (PPT), which consists of "non-normative, rational choice theories of political institutions."[193] Via PPT researchers have

---

[183] Kerr, Shedding Light on Chevron: An Empirical Study of the Chevron Doctrine in the U.S. Courts of Appeals, supra note 27, at 7.

[184] See supra note 101.

[185] 5 U.S.C. 552 (1994). See, for example, Martin E. Halstuk, "Blurred Vision: How Supreme Court FOIA Opinions on Invasion of Privacy have Missed the Target of Legislative Intent," 4 Communication Law and Policy 111 (1999); Eric J. Sinrod, "Freedom of Information Act Reponse Deadlines: Bridging the Gap Between Legislative Intent and Economic Reality," 43 Am U L Rev 325 (1994).

[186] 29 U.S.C. 151-168 (1994). See, for example, Ellen J. Dannin, "Legislative Intent and Impasse Resolution Under the National Labor Relations Act," 15 Hofstra Labor & Employment Law Journal 11 (1997); Eric M. Jensen, "The NRLA's "Guard Exclusion: An Analysis of Section 9(b)(3)'s Legislative Intent and Modern-Day Applicability," 61 Ind L J 457 (1986).

[187] Foreign Trade Zones Act of 1934, Pub. L. No. 73-397, 48 Stat. 998, (codified as amended at 19 U.S.C. 81a-81u (1988)). See, for example, William G. Kanellis, "Reining in the Foreign Trade Zones Board: Making Foreign Trade Zone Decisions Reflect the Legislative Intent of the Foreign Trade Zones Act of 1934," 15 Journal of International Law & Business 606 (1995).

[188] See, for example, Robert H. Bork, "Legislative Intent and the Policy of the Sherman Act," 9 J L & Econ 7 (1966).

[189] Especially U.S. Constit, art. I, §3, cl. 6 ("The Senate shall have the sole Power to try all Impeachments."); art. I, §2, cl 5 ("The House of Representatives…shall have the Sole Power of Impeachment). See, for example, Jonathan Turley, "Senate Trials and Factional Disputes: Impeachment as a Madisonian Device," 49 Duke L J 1 (1999); Lori Fishler Damrosch, "Impeachment as a Technique of Parliamentary Control over Foreign Affairs in a Presidential System?," 70 U Colo L Rev 1525 (1999).

[190] See, for example, Kristin J. Graham, "The Supreme Court Comes Full Circle: Coercion as the Touchstone of an Establishment Clause Violation," 42 Buff L Rev 147 (1994).

[191] See, for example, Akhil Reed Amar, "Fourth Amendment First Principles," 107 Harv L Rev 757 (1994); Anthony Amsterdam, "Perspectives on the Fourth Amendment," 58 Minn L Rev 349 (1974); Morgan Cloud, "Searching through History, Searching for History," 63 U Chi L Rev 1707 (1996).

[192] See, for example, Jacob tenBroek, "The Thirteenth Amendment to the Constitution of the United States," 39 Cal L Rev 171 (1951); Douglas L. Colbert, "Challenging the Challenge: Thirteenth Amendment as Prohibition Agains the Racial Use of Peremptory Challenges," 76 Cornell Law Review 1 (1990); Akhil Reed Amar & Daniel Widawsky, "Child Abuse as Slavery: A Thirteenth Amendment Response to DeShaney," 105 Harv L Rev 1359 (1992).

[193] Daniel A. Farber & Philip P. Frickey, "Foreword: Positive Political Theory in the Nineties," 80 Georgetown L J 457 (1992). More precisely, at least as it has been invoked by legal scholars, PPT belongs to a class of non-parametric rational choice models as it assumes that goal-directed actors operate in *strategic* or *interdependent* decision making context. Seen in this way, it is quite akin to what has been called "the strategic account" in the social-scientific literature. On this account, (1) social actors make choices in order to achieve certain goals, (2) social actors act strategically in the sense that their choices depend on their expectations about the choices of other actors, and (3) these choices are structured by the

sought to address a long list of diverse research questions—from why U.S. Supreme Court justices grant certiorari to petitions cases and deny others[194] to whether the policy preferences of other political organizations (*e.g.,* the legislature and executive) influence judicial decisions[195] to what circumstances lead lower courts to deviate from precedent established by higher courts[196] to why jurists create and maintain (and attorneys now follow) particular rules, norms, and conventions.[197]

At the other end of the spectrum is legal inquiry that does not contain much in the way of theory at all. Some scholars skip it all together as did Manz in his study of Justice Benjamin Cardozo's use of citations.[198] After raising questions about citation practices,[199] Manz moves straightaway to the data; only after presenting some of the results does he begin to theorize about possible explanations. Likewise, in his examination of whether the framers believed that impeachment is a criminal proceeding,[200] Melton, the author, puts forth no theory. Rather he turns directly to the evidence, ultimately using it to make an empirical claim about the framers' beliefs. This strategy is fine, of course, but only if we recognize that "theories" developed in this way have the status of a good hypothesis for which the scholar has yet to provide evidence. In other words, since Melton developed his theory from the same evidence he used to evaluate it, the theory was not vulnerable to being proven wrong. Hypotheses consistent with evidence in the literature play an essential role in scholarship, but we must not confuse them with theories that have empirical support.

Other scholars supplant theory with a review of the relevant literature (or doctrine). Such is typical of work that asks whether a law or court decision had its intended (or unintended) effect. Rather than offering a theory of "effect" or "impact," the researcher reviews other studies, reports, and essays—impressionistic or otherwise—that address whether, say, *Mapp v. Ohio*[201] deterred police

---

institutional setting in which they are made. See, for example, Epstein & Knight, The Choices Justices Make, supra note 63 .

[194] See, for example, Caldeira, et al., Sophisticated Voting and Gate-Keeping in the Supreme Court, supra note 59; Boucher & Segal, "Supreme Court Justices as Strategic Decision-Makers: Aggressive Grants and Defensive Denials on the Vinson Court," supra note 63; Charles M. Cameron, Jeffrey A. Segal, & Donald R. Songer, Strategic Auditing in a Political Hierarchy: An Informational Model of the Supreme Court's Certiorari Decisions, 94 Am Pol Sci Rev 101 (2000).

[195] See, for example, Eskridge, Reneging on History?, supra note 155; Eskridge, "Overriding Supreme Court Statutory Interpretation Decisions," supra note 155; Spiller & Gely, Congressional Control of Judicial Independence, supra note 155; Segal, "Separation-of-Powers Games in the Positive Theory of Law and Courts," supra note 63.

[196] See, for example, Evan H. Caminker, "Why Must Inferior Courts Obey Superior Court Precedent?," 46 Stan L Rev 817 (1994); Donald R. Songer, Jeffrey A. Segal, & Charles M. Cameron, "The Hierarchy of Justice: Testing a Principal-Agent Theory of Supreme Court-Circuit Court Interactions," 38 J Politics 673 (1994); McNollgast, "Politics and Courts: A Positive Theory of Judicial Doctrine and the Rule of Law," supra note 155; Cross & Tiller, "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals," supra note 160.

[197] See, for example, Lewis Kornhauser, "Adjudication By A Resource-Constrained Team: Hierarchy and Precedent in a Judicial System," 68 S Cal L Rev 1605 (1995); Knight & Epstein, "On the Struggle for Judicial Supremacy," supra note 155; Maxwell L. Stearns, Constitutional Process: A Social Choice Analysis of Supreme Court Decision Making (1999).

[198] William H. Manz, "Cardozo's Use of Authority: An Empirical Study," 32 Cal W L Rev 31 (1995).

[199] "Why and to what extent do the citation practices of individual judges actually differ? Are differences idiosyncratic, the product of the cases assigned, or do they reflect the beliefs and intellectual background of the individual judge? Does the authority utilized by a light citer differ in any way from that of a heavy citer? Do the citation practices of a judge of great reputation differ considerably from those of her less famous colleagues, and if so why? Will a liberal jurist use more or less authority than a conservative?" Id., at 32.

[200] Buckner F. Melton, Jr., "Federal Impeachment and Criminal Procedure: The Framers' Intent," 52 Md L Rev 437 (1993).

[201] 367 U.S. 643 (1961).

misconduct[202] or the Federal Rule of Civil Procedure 68 induced pre-trial settlements[203] and, then, uses those reviews to generate expectations.[204]

Theories thus come in many types, levels of abstraction, and substantive applications. Each of these distinctions may be more or less consequential depending of the goal and purpose of the research. But, regardless of those considerations and even of the kind of theory they invoke, researchers should recognize that they can help themselves comply with many of the other rules we discuss in this Article by making theories more useful—a goal they can accomplish by (1) invoking theories that produce observable implications and then (2) extracting as many implications as possible and (3) delineating how they plan to observe those implications.

1. Invoke Theories that Produce Observable Implications

Good theories come with a guide to developing observable implications about the phenomenon they seek to describe or explain. Only by evaluating those observable implications—comparing the theoretical implications with some relevant empirical observations—can we learn whether the theory is likely to be correct.

Observable implications often take the form of claims about the relationships among variables that we can, at least in principle, observe. By "variables," we mean characteristics of some phenomenon that vary across instances of the phenomenon: for example, the race of a person, the outcome of a Supreme Court case. Earlier we discussed "causal" variables—those that we think lead to a particular outcome, such as the existence of *Miranda* causing fewer confessions—and "dependent variables"—those outcomes we are trying to explain, such as the rate of confessions. Causal variables fall under the general rubric of "independent" (or "explanatory") variables—those that may help account for the outcome (others falling under the same category are "control" variables, which we discuss below).

To see how the process of moving from theory to observable implications can work, consider research by Eskridge that invokes positive political theory to understand how justices on the U.S. Supreme Court interpret federal statutes.[205] Under his account, justices have goals, which, according to Eskridge, amount to seeing their policy preferences written into law,[206] but realize that they cannot achieve them without taking into account the preferences and likely actions of other relevant

---

[202] See, for example, Oaks, "Studying the Exclusionary Rule in Search and Seizure," supra note 165; Perrin, et al., If It's Broken, Fix It, supra note 135; Myron W. Orfield, Jr., "The Exclusionary Rule and Deterrence: An Empirical Study of Chicago Narcotics Officers," 54 U Chi L Rev 1016 (1987); William C. Heffernan & Richard W. Lovely, "Evaluating the Fourth Amendment Exclusionary Rule: The Problem of Police Compliance with the Law," 24 University of Michigan Journal of Law Reform 311 (1991); Comment, "Effect of Mapp v. Ohio on Police Search-and-Seizure Practices in Narcotics Cases," 4 Colum J L & Soc Probs 87 (1968)

[203] Anderson & Rowe, "Empirical Evidence on Settlement Devices: Does Rule 68 Encourage Settlement?," supra note 103.

[204] Exemplary of this approach, across a range of issues, is Jesse H. Choper, "Consequences of Supreme Court Decision Upholding Individual Constitutional Rights," 83 Mich L Rev 1 (1984).

[205] See, for example, Eskridge, Reneging on History?, supra note 155; Eskridge, Overriding Supreme Court Statutory Interpretation Decisions, supra note 155.

[206] Eskridge is not alone: Many positive political theory (PPT) accounts of judicial decisions assume that the goal of most justices is to see the law reflect their most preferred policy positions. See, for example, Epstein & Knight, The Choices Justices Make, supra note 63 ; Spiller & Gely, Congressional Control or Judicial Independence, supra note 155. But this need not be the case. Under PPT, strategic actors—including justices—can be, in principle, motivated by many things. As long as the ability of a justice to achieve his or her goal, whatever that may be, is contingent on the actions of others (as PPT suggests), his or her decision is interdependent and strategic. For an example of a PPT account of judicial decisions in which justices are motivated by jurisprudential principles, see Ferejohn & Weingast, "A Positive Theory of Statutory Interpretation," supra note 155.

actors—including congressional gatekeepers (such as chairs of relevant committees and party leaders), other members of Congress, and the President—and the institutional context in which they work.

To develop observable implications from this account, Eskridge uses pictures of the sort we display in Figures 3a and 3b.[207] In each, we depict a hypothetical set of preferences over a particular policy, say, a civil rights statute. The horizontal lines represent the (civil rights) policy space, here, ordered from left (most "liberal") to right (most "conservative"); the vertical lines show the preferences (the "most preferred positions") of the relevant actors: the President, the median member of the Court, of Congress, and of the key committees and other gatekeepers in Congress that make the decision over whether to propose civil rights legislation to their respective houses.[208] Note we also identify the committees' indifference point "where the Court can set policy which the committee likes no more and no less than the opposite policy that could be chosen by the full chamber."[209] To put it another way, because the indifference point and the median member of Congress are equidistant from the committees, the committees like the indifference points as much as they like the most preferred position of Congress; they are indifferent between the two.

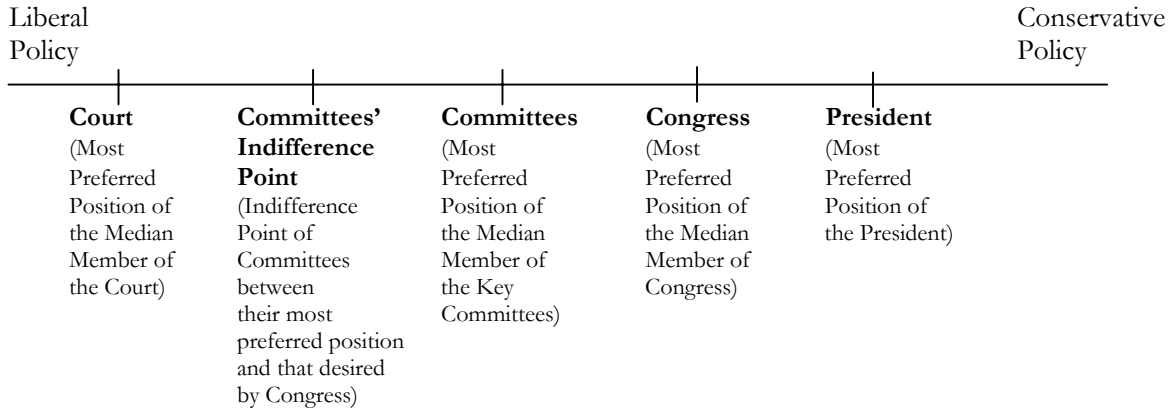
[207] The term of art for these pictures, in the way that Eskridge (and we) uses them, is spatial models. Such help scholars to investigate how the decisions of one actor may influence those of another (or others). For good a introduction to spatial models, see Peter C. Ordeshook, Game Theory and Political Theory: An Introduction (1986).

[208]In denoting these most preferred points, we (and Eskridge) assume that the actors prefer an outcome that is nearer to that point than one that is further away. Or, to put it more technically, "beginning at [an actor's] ideal point, utility always declines monotonically in any direction. This...is known as single-peakedness of preferences." Keith Krehbiel, "Spatial Models of Legislative Choice," 13 Legis Studies Q 259 (1988). To derive the observable implications detailed in the text and in Figure 3, we (and Eskridge) also assume(s), that the actors possess complete and perfect information about the preferences of all other actors and that the sequence of policy making enfolds as follows: the Court interprets a law, the relevant congressional committees propose (or do not propose) legislation to override the Court's interpretation, Congress (if the committees propose legislation) enacts (or does not enact) an override bill, the President (if Congress acts) signs (or does not sign) the override bill, and Congress (if the President vetoes) overrides (or does not override) the veto.

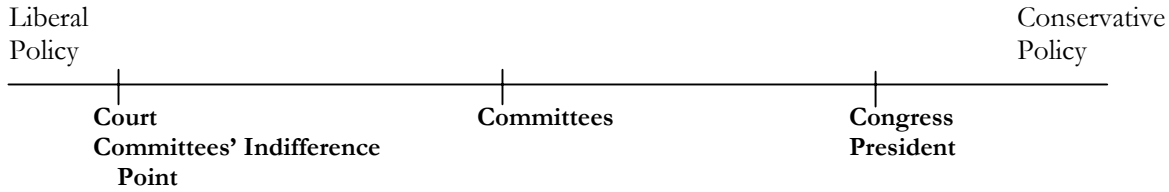
[209] Eskridge, Overriding Supreme Court Statutory Interpretation Decisions, supra note 155, at 381.

Figure 3. Hypothetical Distribution of Preferences[210]

3a. Observable Implication 1: Policy is set on the committee's indifference point.



3b. Observable Implication 2: Policy is set on the committee's indifference point/Court's most preferred position.



As we can see, in Figure 3*a* the Court is to the left of Congress, the key committees, and the President. This means, in this illustration, that the Court favors more liberal policy than do the other political branches. In Figure *3b* the Court also is to the left of the relevant actors but, note, the committees' indifference point is on the Court's most preferred position.

Now suppose that the Court has accepted a case that calls for it to interpret the civil rights statute. How would the Court proceed? From Eskridge's theory the following observable implication emerges: Given the distribution of the most preferred positions of the actors, the *3a* Court would not be willing to take the risk and interpret the statute in line with its most preferred position. It would see that Congress could easily override that position and that the President would support Congress. Rather, under Eskridge's theory, the best choice for justices interested in seeing the law reflect their policy preferences is to interpret the statute near the committees' indifference point. The reason is simple: Since the committees are indifferent between that point and the position preferred by the median legislator, they would have no incentive to introduce legislation to overturn a policy set at their indifference point. Thus, the Court would end up with a policy close to, but not exactly on, its ideal point without risking a congressional backlash. The distribution of preferences in Figure *3b* points to a different observable implication: The Court would set policy in a way that reflects its sincerely held preferences. For if it votes its preferences (which are comparatively liberal) and sets the policy at its most preferred position, the relevant congressional committees would have no incentive to override the Court: Since their indifference point is the same as the Court's most

[210] We adapt this figure from Id.; see also Eskridge, Reneging on History?, supra note 155.

preferred position, they would be indifferent to the policy preferred by the Court. Note that for both implications the theory suggests the main explanatory variables—the preferences of the key actors relative to one another and the dependent variable—the Court's interpretation of a statute.

The theory underlying Eskridge's account is grand in scope but those that are narrower, that scholars develop from, say, reviews of the relevant literature or doctrine, can provide similar guidance. The Hillman study on promissory estoppel, which we described earlier, is exemplary.[211] The author did not develop his own theory but rather relied on the "new consensus" to develop observable implications.

This is all to the well and good, for theories without observable implications are of little use. That is so for several reasons, not the least of which is that, without clear implications, readers cannot know whether there is any empirical support for the theory, or whether it is ever vulnerable to being proven wrong. Doctrinal studies or those theorizing about legislative intent occasionally fall into this trap, with Amar and Widawsky's study of the Thirteenth Amendment illustrative.[212] In it, the authors theorize that the Amendment, both "in letter and spirit," "speaks to the horror of child abuse with remarkable directness."[213] To be sure, this is an interesting theory and one the researchers describe in some detail and attempt to support with reference to the legislative history of the Amendment, judicial interpretations, and so on. What they do not do, however is provide observable implications that would ultimately help readers determine whether the theory rests on solid ground.[214] To accomplish this crucial task, the authors need to ask themselves the following: If our theory is right and the "letter and spirit" of the Amendment speaks with "directness" to child abuse, then what implication*s* or prediction*s* about real-world behavior follow from it?[215]

B. Extract as Many Observable Implications as Possible

As our emphasis on the plural suggests, scholars should not stop with one or two implications or predictions; they should instead develop as many as possible even if they are only indirectly connected to the specific hypothesis of interest. To begin to see the logic behind this claim, consider studies on the effect of *Miranda v. Arizona*[216] that start with the following (if usually implicit) theory: *Miranda* led to a reduction in the number of confessions obtained by police. Such a theory, to reiterate a point we made earlier, exemplifies those often invoked by legal scholars: It is clear, concrete, and tailored toward a specific phenomenon. But that does not mean it is incapable of generating observable implications; it, in fact, is, with one being rather obvious: We should observe a reduction in confession rates when *Miranda* is enforced. Another might be: In places without a *Miranda*-like ruling or law, we should observe a smaller reduction in confession rates.[217]

---

[211] Hillman, Questioning the "New Consensus" on Promissory Estoppel, supra note 91.

[212] Amar & Widawsky, Child Abuse as Slavery, supra note 192.

[213] Id., at 1360.

[214] We write this fully appreciating that Id. and, in fact, many studies of this sort, as well as doctrinal analyses, may be more concerned with advancing a new perspective or a new way of contemplating familiar problems than they are with providing support for their claims. We have no qualms with these goals; in fact, we applaud them. But, to the extent that the authors of these studies—Amar & Widawsky included—make empirical claims or inferences—they are not exempt from following the rules we offer any more than would be, say, a scholar conducting a wholly quantitative study, the only goal of which is to offer empirical claims or inferences. In fact, it would be inappropriate, if not disingenuous, for us to say to legal academics that our rules apply if their article is almost all empirical but not if it is just a little bit. Empiricism is empiricism whenever, wherever, and however much it is used.

[215] We offer some answers on infra p. 52.

[216] 384 U.S. 436 (1966).

[217] See, for example, James W. Witt, "Non-Coercive Interrogation and the Administration of Criminal Justice: The Impact of Miranda on Police Effecuality," 64 J Crim L & Criminol 320 (1973); Michael Wald, et al., Interrogations in

These observable implications might lead an investigator to locate jurisdictions where law enforcement officials comply more or less fully with *Miranda*, and to collect information on confession rates in both.[218] But the researcher should not stop there; she or he should seek out other implications—those that, yes, would match up with his or her original theory about the impact of *Miranda* on confession rates, *as well as* those that would be consistent with the many other effects *Miranda* might have had on the criminal justice system. *That is because any effects found in areas other than confession rates would increase the plausibility of the original theory.* (This general claim always holds but is especially applicable for this example since, as we know from the large body of literature on the case, many members of the legal community—including scholars, lawyers, and judges and other decision makers—are not solely or simply concerned with whether *Miranda* reduced the number of confessions; they seem more interested in the broader effect, if any, it had on the criminal justice system.[219]).

To understand why we make this claim, imagine a researcher who accepted it and listed, in addition to observable implications extracted from the researcher's specific theory—*e.g.,* If my theory is correct, we should observe a reduction in confession rates when *Miranda* is enforced—those implications flowing from a broader take on *Miranda's* effect on the criminal justice system, including:

(1) Trial court judges granting *Miranda*-based motions to exclude confessions[220]
(2) Appellate court judges reversing convictions on *Miranda*-based issues[221]
(3) Courts, at all levels, expending too much "valuable" time on *Miranda* cases[222]
(4) Law enforcement officials seeing a reduction in clearance and conviction rates[223]
(5) Defendants failing to waive their *Miranda* rights[224]

Other implications are easy enough to summon[225] but assume, for a moment, that these were the ones on which an investigator focused. Further assume that the researcher, following all other advice

---

New Haven: The Impact of Miranda, 76 Yale L J 1519 (1967); Richard J. Medalie, et al., Custodial Police Interrogation in Our Nation's Capital: The Attempt to Implement Miranda, 66 Mich L Rev 1347 (1968); Younger, "Interrogation of Criminal Defendants—Some Views on Miranda v. Arizona," 35 Fordham Law Review 255 (1966-67).

[218] We realize that this example opens the door for possible selection bias; *i.e.,* compliance may be lower in areas where confession rates are already very high. We discuss this problem in Part VIII.

[219] See, for example, Richard H. Seeburger & R. Stanton Jr. Wettick, "Miranda in Pittsburgh—A Statistical Study," 29 U Pitt L Rev 1 (1967); Special Comm. on Criminal Justice in a Free Society, Criminal Justice Section, ABA, Criminal Justice in Crisis 28-29 (1988); Miranda v. Arizona, 384 U.S. 436 (1966) (White, J. dissenting); Executive Summary of the Office of Legal Policy, U.S. Dep't of Justice, Truth in Criminal Justice Series, Report No. 1, Report to the Attorney General on the Law of Pretrial Interrogation (1986), reprinted in 22 U. Mich. J.L. Ref. 437 (1989).

[220] Peter Nardulli, "The Societal Costs of the Exclusionary Rule Revisited," 1987 U Ill L Rev 223 (1987); Gerald M. Caplan, "Questioning Miranda," 38 Vand L Rev 417 (1985).

[221] Thomas Y. Davies, "Affirmed: A Study of Criminal Appeals and Decision-Making Norms in a Calfornia Court of Appeal," 1982 Am Bar Found Rsrch J 543 (1982); Karen L. Guy & Robert G. Huckabee, "Going Free on a Technicality: Another Look at the Effect of the Miranda Decision on the Criminal Justice Process," 4 Criminal Justice Research Bulletin 1 (1988)

[222] Fred E. Inbau & James P. Manak, "Miranda v. Arizona: Is It Worth the Cost?," 24 Cal W L Rev 185 (1988)

[223] Seeburger & Wettick, Miranda in Pittsburgh, supra note 219; Robinson, "Police and Prosecutor Practices and Attitudes Relating to Interrogation as Revealed by Pre- and Post-Miranda Questionnaires: A Construct of Police Capacity to Comply," 1968 Duke L J 425 (1968); Witt, Non-Coercive Interrogation and the Administration of Criminal Justice, supra note 217.

[224] Medalie, Custodial Police Interrogation in Our Nation's Capital, supra note 217.

[225] Some of these could even come from theories that, at first blush, have little relationship to questions about the impact of *Miranda.* One that comes readily to mind is agency theory, which assumes that value conflicts are pervasive in organizations, that the outcomes of these conflicts reflect the power of the contestants, and that the details of

we offer, found that all five held. If that were the case, we would have more confidence in the claims made by some scholars that *Miranda* had an effect on the American legal system than we would had the researcher considered only one implication, such as a reduction in confessions even if she or he found that one to hold. More relevant here: The original goal of assessing the effect of *Miranda* on confession rates would be better served by this strategy. The reasoning is that if *Miranda* is powerful enough to have an effect on these rates, it should also evince some of the other observable implications listed above; conversely, if some of these do not hold, the conclusion about confession rates would have less support. Seen in this way, the strategy of seeking to "maximize leverage"[226]—that is, identifying the largest number of observable implications possible, even if the immediate purpose is a narrower inquiry—can be very powerful.

We can say the same of research that relies on non-numerical forms of evidence. Return to Amar and Widawsky's study of the Thirteenth Amendment.[227] Though the authors offer no explicit hypotheses, surely their theory—that child abuse is a form of slavery outlawed by the Thirteenth Amendment—lends itself to several testable predictions; for example, if the term "slavery" is broad enough to cover child abuse, then we might expect to find traces of that sentiment in the historical legislative and judicial records. From this idea flow many observable implications; to wit, (1) statements made by authoritative law makers (*e.g.,* members of the legislative majorities)[228] indicating that they intended the word "slavery" to encompass more than the "'peculiar institution' of southern chattel slavery;"[229] (2) early interpretations of the Amendment produced by courts demonstrating that judges understood the term slavery to cover children subject to abusive parental behavior; and (3) definitions provided in dictionaries of the day supporting a broad conception of "slavery."

These are just three examples; readers and researchers may take issue with some but could no doubt develop others. For now, we emphasize the more general lesson: For all theories researchers should ask, "What are their observable implications?" and, in turn, *list all the possibilities—even if only a small subset of them are actually observed in the course of the research.*[230] The more implications scholars identify, the more powerful and useful their theory. And, the more of these implications they can evaluate against real data, the more confidence we can have in their conclusions.

C. Delineate How Implications Can Be Observed

Because theories and their implications are typically comprised of concepts, researchers must, to begin to assess their theories and related implications, delineate how they can observe them in the real world. For example, in order to assess an observable implication of Eskridge's theory—that the Court will interpret a civil rights law consistent with its preferences if the Court and other relevant actors share the same vision of civil rights policy—we need a clear definition of civil rights laws (*e.g.,* we could define them narrowly as only those that specifically claim to be civil rights laws or we

---

organizational design and operating procedures ("the rules of the game") determine power. See, for example, Songer, Segal, & Cameron, The Hierarchy of Justice, supra note 196; Terry M. Moe, The New Economics of Organization, 28 Am J Pol Sci 739 (1984). From this theory we could develop numerous observable implications pertaining to the effect of *Miranda*, for the theory gives us leverage in understanding under what circumstances, say, a federal circuit court will (and will not) deviate from precedent established by the U.S. Supreme Court, police officers will (and will not) defy their superiors, and so on.

[226] King, Keohane, and Verba, Designing Social Inquiry (1994), supra note 1, at 29-31.

[227] Amar & Widawsky, Child Abuse as Slavery, supra note 192.

[228] Richard S. Kay, "Adherence to the Original Intentions in Constitutional Adjudication: Three Objections and Responses," 82 Nw U L Rev 226 (1988).

[229] Amar & Widawsky, Child Abuse as Slavery, supra note 192, at 1359.

[230] King, Keohane, and Verba, Designing Social Inquiry, supra at 1, at 30.

could define them broadly as all laws, regardless of their original purpose or intent, that courts have used to protect civil rights). That is because civil rights legislation—just as are so many phenomena of interest in legal research, such as "compliance," "legitimacy," and "efficiency"— is a concept requiring clarification so that we can observe it. This process of "clarification" is sometimes called "operationalizing," "operationally defining," or, more simply, "defining" the concepts.

This should not, on the one hand, be a terribly onerous task. If researchers carefully and clearly lay out their theories, then it should be easy for them to develop clear, measurable definitions of the concepts contained in the implications of those theories, as well as for readers to judge whether they have done a good job. For this reason, specifying the theory with sufficient precision so that readers can see how authors measure their implications is a critical part of theorizing. Return to Eskridge's work on statutory interpretation. Under his theory, the preferences of the median member of the Supreme Court, as well as those of other political organizations (*e.g.,* the Senate, the House of Representatives, key committees) help to explain particular policy outcomes. Defining the "median member" of the Supreme Court is straightforward; it is the justice in the middle of the distribution of members of the Court along a defined policy dimension.

On the other hand, difficulties often arise in identifying this justice in practice. Perhaps the most common problem occurs when researchers have not clearly specified their theory. Revesz's examination of voting patterns of judges in the U.S. Court of Appeals for the District of Columbia in environmental cases provides an example.[231] In part, he seeks to assess the following theoretical statement:

> Some commentators have…maintained that judges simply vote according to their policy preferences. In environment cases, the allegation goes, judges appointed by Republican Presidents vote principally for laxer regulations and judges appointed Democratic Presidents for more stringent regulation.[232]

Note the conceptual difficulty: Does the theory require judges to vote in accord with their own policy preferences or in line with the party of the president who appointed them? This may seem a distinction without meaning but, because a "policy preference" is not the same as a "partisan affiliation," it is not. It would be possible to operationalize "policy preferences" in any number of ways: whether the judge affiliates with the Democratic or Republican party, whether the judge is a liberal or conservative, or, yes, whether the president appointing the judge affiliated with the Democratic or Republican party. The "partisanship of the President" normally will be less ambiguous and will be defined simply as the president's political party membership. But since the theory conflates the two, we face the difficulty of deciding whether the implication to be observed is "policy preferences" or "partisanship." (What, it turns out, Revesz actually means here is that partisanship could serve as a *measure* of policy preferences.[233] We discuss this altogether different issue in Part VII.)

Revesz's is a quantitative study but similar problems emerge in qualitative studies in which authors do not carefully specify their theories. Take Williams's 1999 essay, which posits that legislative history of the Securities and Exchange Act of 1934 permits the SEC to require corporations to file social, and not just financial, disclosures.[234] But what does Williams mean by the "legislative history"? Is she using the term in "a very broad sense," to signify "the entire circumstances of a statute's creation (and evolution)" or "something much narrower—the

---

[231] Richard L. Revesz, "Environmental Regulation, Ideology, and the D.C. Circuit," 83 Va L Rev 1717 (1997).

[232] Id., at 1717.

[233] See Id., at note 6. See also infra note 260.

[234] Cynthia A. Williams, "The Securities and Exchange Commission and Corporate Social Transparency," 112 Harv L Rev 1199 (1999).

institutional progress of a bill to enactment."[235] Because of the lack of precision on this dimension, we, the readers, cannot automatically envision a clear measure of "legislative history," one that we would know to be consistent with the original theory.[236] The result, in turn and as we suggest below, impinges on our ability to judge whether any measures she invokes to tap that implication of the theory appropriately tap her concepts. If she takes a narrow view of legislative history, then she might need a measure that summarizes all or some of the 20 materials identified on, say, the Hetzel et al. list, moving in time from committee reports to floor debates to recorded votes.[237] A broader view would require a measure that went back even further, to the circumstances surrounding the introduction of the bill.[238]

How ought Williams go about making choices with regard to the delineation of the observable implications of her theory and how ought we judge whether she has made good ones? The general rule here is to develop working definitions that minimize loss from concept to the definition. This follows from the fact that we cannot observe directly the concepts that flow from our theories, even though it is those very theories that we want to assess. Hence, the closer researchers can come to clarifying concepts so that they can measure them empirically, the better their tests will be.

VI. Controlling for Rival Hypotheses

At the onset of this Article, we noted that scientists who seek out all evidence against their "favored" theory are following the rules of inference, that researchers who maximize their vulnerability, and the different areas and data sets in which they could be proven wrong, are operating in accord with the best traditions of empirical scholarship. Nothing we have written since changes this basic premise. Quite the opposite: Since only by posing sufficient challenges to their theory (and its observable implications) are researchers able to make the strongest possible case for it, scholarship that treats theories as clients in need of the best defense is highly problematic.

To see why, consider Bufford's "empirical" study of chapter 11 bankruptcy cases, in which he posits that "relatively modest judicial case management can squeeze a substantial amount of delay out of [these] cases within the context of the present bankruptcy law."[239] Bufford goes on to lay out one observable implication of this claim—judges who adopt "fast track," a particular model of case management, reduce delays—that he assesses by examining the docket of one judge before and after she invoked fast track. When this investigation reveals that fast track "shortened by 24.1% the time to confirmation of a chapter 11 plan in a typical case,"[240] the researcher claims victory; his theory, he believes, is correct. But Bufford's declaration is premature. For he fails to take into account competing explanations of delay reduction as they may pertain to the particular judge under analysis or to the general phenomenon. It is entirely possible that fast track was not the only change the judge made during the period under analysis or even that attorneys, realizing that the judge had altered her management practices, altered theirs (*e.g.,* failing to file particular suits or taking them elsewhere). Without a consideration of these and many other alternative explanations, the author has not supported his theory with the force it deserves. In fact, his neglect makes it look all the weaker.

---

[235] William N. Eskridge, Jr. & Philip P. Frickey, Cases and Materials on Legislation (1995).

[236] A reader, or future researcher, could of course further specify the theory and then derive observable implications, but subsequent empirical work may then be testing a different theory altogether.

[237] Otto Hetzel, Michael Libonati, & Robert Williams, Legislative Law and Process (1993), at 438

[238] Eskridge & Frickey, Cases and Materials on Legislation, supra note 235.

[239] Samuel L. Bufford, "Chapter 11 Case Management and Delay Reduction: An Empirical Study," 4 Am. Bankr. Inst. L. Rev. 85 (1996).

[240] Id.

Avoiding this problem requires researchers to comb the existing literature for and to think hard and imaginatively about explanations that do not square with the theory they are offering. (The former is easy to do;[241] the latter is not, and, as such, is one reason why all investigators need a community of scholars to help.) Moreover, authors should alert readers to the fruits of those exercises—any existing rival explanations—and, ultimately, build them into their research.

This last step is critical for, if scholars ignore competing explanations, their work will suffer from what is known as "omitted variable bias," making any causal inferences they reach suspect. We have more to say about avoiding this type of bias momentarily but note the general implication of this claim: In selecting variables for their study, scholars cannot stop with those that flow directly from the observable implications of their theory. *They will in all likelihood have to incorporate variables designed to control for the implications of other theories that do not necessarily square with theirs (i.e., rival explanations or hypotheses).*

Collecting data on variables that support the positions of potential critics is perhaps the optimal way scholars maximize vulnerability, and ultimately, inoculate themselves from criticism. The goal is not to destroy the position of a detractor, since both the researcher and the critic could be right, but to ensure the absence of omitted variable bias. This is especially critical in work seeking to make causal inferences. The Bufford study makes this clear, as does Ramos' attempt to assess the effect of affirmative action plans invoked by law reviews to select the members of their staff.[242] Based on the results of a survey of law review editors, Ramos concludes that the absence of such an affirmative action program "effectively excludes minorities from membership on a large number of law reviews."[243] His evidence? 38% of the 78 law reviews without affirmative action programs lack minority members, but all 6 law reviews with affirmative action programs have minority members. In other words, he relies on one explanatory variable (the presence/absence of an affirmative action plan for the selection of law review members) to make a causal claim. The problem with this approach is akin to the one we identified in Bufford's research; namely it ignores many other potential explanations for the dependent variable (minority membership on law reviews)— chief among them minority representation in the law school writ large. Surely, we might expect that the more minorities in the school's population, the greater their numbers on law review staffs. Worse still, this rival explanation may be causally prior to the author's: It is entirely plausible that the more minorities in a law school, the more likely the existence of an affirmative action program for the selection of law review members.

Because Ramos omits potential rival explanations, he cannot be confident that his favored variable is actually doing the work. But avoiding omitted variable bias does not mean that he or other analysts must incorporate variables representing every conceivable alternative explanation. Rather, the requirements are easily formalized: Researchers should control for (*i.e.,* hold constant) a potential confounding variable *only if the rival variable meets one of the following conditions*:

---

241 That LEXIS-NEXIS, OCLC, J-Stor, and others, see supra note 161, provide journals in searchable electronic form makes combing the literature a far easier task than it was just a mere decade ago.

242 Frederick Ramos, "Affirmative Action on Law Reviews: An Empirical Study of its Status and Effect," 22 U. Mich. J.L. Ref. 179 (1988). Ramos and Bufford, Chapter 11 Case Management and Delay Reduction: An Empirical Study, supra note 239, are just two of many examples illustrating omitted variable bias. Indeed, legal scholars themselves have pointed out the problem as it has manifested in various essays appearing in the law reviews. See, for example, Schulhofer, "Miranda's Practical Effect: Substantial Benefits and Vanishingly Small Social Costs," supra note 40, on studies on the effect of *Miranda* on confession rates; Warren, "The Untenable Case for Repeal of Chapter 11," supra note 81, on Bradley & Rosenzweig's, "The Untenable Case for Chapter 11," supra note 80 analysis of causes in the rise in bankruptcy filings; and Theodore Eisenberg, "Measuring the Deterrent Effect of Punitive Damages," 87 Georgetown L J 347 (1998) on W. Kip Viscusi's, "The Social Costs of Punitive Damages Against Corporations in Environmental and Safety Torts," 87 Georgetown L J 285 (1998) examination of whether punitive damges meet their objectives.

243 Ramos, "Affirmative Action on Law Reviews," supra note 242 , at 198.

(1) It is related to (correlated with) the key causal variable.,
(2) It has an effect on the dependent variable.,
(3) It is causally prior to (*e.g.,* preceding in time) the key causal variable.

In the case of the Ramos study, at least one potential rival explanation—the number of minorities in the law school—appears to meet all three of these conditions. That number is probably related to the existence (or not) of affirmative action programs in law review staff selection; it may affect the number of minority students on law review staffs; and it is causally prior to (*i.e.,* a cause of) the creation of affirmative action plans for law review selection. Since it meets each of the three conditions for omitted variable bias, Ramos' causal inference is biased. And, since they do not account for the effect of the rival explanation, his results are indeterminate: They are consistent with a strong casual effect, a negative effect (whereby affirmative action programs reduce minority participation on law review staffs), or no relationship at all.

As uninformative as omitting a rival variable would be, measuring it and merely showing that it is separately associated with the dependent variable is insufficient. What needs to be done is to examine the effect of affirmative action programs on the minority composition of law review staffs *controlling for* the effect of the number of minorities in the law school. A simple way Ramos could do this is by conducting his study using a set of law schools that have nearly the same fraction of minority students.[244] In this way, the rival variable would be held constant; and, since a constant cannot cause (or even be correlated with) a variable, the research design guarantees that it cannot confound the relationship between the key causal variable and the dependent variable.

If selecting observations so that rival explanations are held constant is infeasible, investigators can use statistical methods to "statistically hold constant" control variables; similarly, they can conduct experiments in which they "physically hold constant" these variables. One way or another, however, they must control for rival variables that meet all three conditions for causing omitted variable bias. They cannot ignore them, and they cannot stop by merely demonstrating that no separate association exists between one explanatory variable and the dependent variable.

On the flipside, if any of the three conditions for omitted variable bias do *not* apply, then controlling for the rival variable will not merely be useless in evaluating the effect of the key causal variable; it can waste valuable data in estimating irrelevant quantities, as well.[245] Even worse, if the third condition (that the rival variable is causally prior to the key causal variable) does not hold, and researchers control for the variable anyway, they will introduce large biases. The dependent variable in Ramos' study is the number of minorities *actually* selected to serve on each law review staff. Suppose we control for the number of minority members those charged with selecting new members *intended* to appoint five minutes before they announced their decision. Clearly this "intentions" variable will predict our dependent variable almost exactly and we should not control for it. If we held "intentions" constant by selecting all law schools where the decision makers had approximately the same intentions, the actual level of minority representation would be almost identical for every law school in the analysis. This would lead us to conclude that affirmative action

---

[244] We realize that this particular alternative design is not necessarily optimal given Ramos' concerns but it does provide an example of one way of controlling for this omitted variable.

[245] Such occurs too often in the law reviews, with authors incorporating virtual laundry lists of variables into their analyses but failing to explain, theoretically or otherwise, their importance. Lederman, Which Cases Go to Trial?, supra note 27, at 327 provides a *possible* example. In attempting to explain why some tax cases go to trial, she includes the following variables: 'STAKES,' 'APPEALS,' 'JUDGETYPE,' 'DECADE,' 'BACKGROUND,' 'PARTY,' 'TAXPAYER,' 'REGION' and 'COUNSEL.' It is only later in the article that we learn what some of these variable labels mean. For example, it turns out that PARTY is is the political party of the President appointing the judge.

programs have no causal effect, even if their effect is actually quite large. So, while this possible rival variable, "intentions," meets conditions (1) and (2), it must be excluded from the analysis because it fails to meet condition (3).[246]

VII. Measurement and Estimation

Once scholars have identified the variables for inclusion in their study, they must *measure* those variables and derive *estimates*. Imagine that we wanted to determine whether attending a law school with a better reputation leads to a higher starting salary upon graduation. To assess this hypothesis, we must translate the variable "reputation of a law school" into some precise indicators of reputation. This is the act of *measurement*: comparing an object of study (*e.g.,* a real world event, subject, or process) with some standard, such as exists for quantities, capacities, or categories. We typically measure height by comparing an object to a standard such as feet or meters; we might measure achievement by comparing student test scores to a percentage correct scale; we can measure temperature in inches of mercury in a thermometer. We could measure the reputation of a law school by asking a sample of potential employers to tell us which law school they would turn to first for new lawyers or using published rankings such as from *U.S. News and World Report*.

While *measurement* involves how to record each individual datum, *estimation* involves marshalling a whole set of measurements (or "data") to learn about a quantity of interest. Suppose we learn about the reputation of law schools by administering a survey to a random sample of employers. Each respondent's answer to the survey question is a measurement whereas the average of all the responses is one way to estimate the average beliefs about law school reputations among all employers in the United States.

What this very simple example is designed to illustrate is that the process of moving from an observable implication, to observing many instances of the implication (*i.e.*, measurement) to estimation is a critical step in empirical research. That is because, as we suggested above, we can never actually draw comparisons among variables as immediately conceptualized in a theory; we cannot compare "reputations." All we can do is compare the readings of reputation we obtain from some measure of them. This means that our comparisons and, ultimately, our answers to research questions are only as valid as the measures we have developed. If those measures do not adequately mirror the concepts contained in our theories, the conclusions we draw will be faulty.

Accordingly, in this section, we expend a considerable number of pages on rules for evaluating measures and estimates, and we offer suggestions on how to improve each. Since individual measures can involve very hard work, many subparts, and indeed inferences that are as difficult as estimation, the distinction between the two is often arbitrary. We retain it because it is often convenient, as it is here, to put aside some issues to focus on more important ones.

A. Measurement

Measurement, as we just noted, involves comparing some aspect of reality with a standard, such as exists for quantities, capacities, or categories. For example, if we define inter-circuit "conflict" as did Lawless and Murray in their study of the U.S. Supreme Court's certiorari decisions in bankruptcy cases[247]—as *genuine* conflict between at least two circuit courts—then we must develop a measure that captures the definition as precisely as possible. To Lawless and Murray, that measure is whether

---

[246] Gary King, ```Truth' is Stranger than Prediction, More Questionable Than Causal Inference," 35 Am J Pol Sci (1991), at 1047-1053.

[247] Lawless & Murray, "An Empirical Analysis of Bankruptcy Certiorari," supra note 58.

an "originating circuit court expressly stated its disagreement with the holding of another circuit court regarding any of the 'issues presented' in the petition."[248] If a petition met this criterion, conflict existed; if it did not, conflict did not exist.

From even this brief example, a clear disadvantage of measuring phenomena is apparent: Everything about the object of study is lost except the dimension or dimensions being measured. This is true of almost all measurement schemes. Summarizing "George W. Bush" by saying he is 5 feet 10 inches tall obviously leaves out an enormous range of information, just as does claiming that conflict exists only when a court of appeals notices it. And yet measurement allows us to put many apparently disparate events or subjects on the same dimension, making it far easier to comprehend at least one aspect of the phenomenon under study. Instead of understanding intercircuit conflict in 200 cert. petitions by looking at the set all at once, we can greatly simplify the task by summarizing it with 200 numbers. Even more to the point, understanding the real world always requires a certain level of abstraction, and so measurement of some kind plays a central role in empirical research. The key is that we abstract the right dimensions for our purposes, and that we measure enough dimensions of each subject that we capture all the parts that are essential to our research question.

This is as true in quantitative empirical research as it is qualitative work. In the former, researchers typically assign numerical values to their measures. Separation of powers studies, of the sort Eskridge conducts, offer an example. If researchers define the "median member of the Court" on a particular policy dimension as that justice who is in the middle of the distribution of members of the Court on civil rights policy, then they must identify that justice and, typically, attach a numerical policy preference score to the justice's policy position. Doing so requires researchers to develop a measure of the policy preferences of justices or to invoke an existing one, such as those supplied by Segal and Cover[249] or derived from Spaeth's U.S. Supreme Court Judicial Data Base.[250]

Although numerical summaries can be convenient and concise, and are by definition precise, measurement need not involve numbers—as is often the case in qualitative research. Categorizations, such as "tall," "medium," and "short," or "Catholic," "Protestant," and "Jewish," are reasonable measures that can be very useful, assuming researchers sufficiently define the standard for measurement so that they (or others) can unambiguously apply it. To see this, return to Williams's study of the SEC. What she must measure is whether the legislative history of the Securities and Exchange Act of 1934 enables the SEC to require social disclosures.[251] Let us assume that she defines legislative history in the "narrow" sense—to include materials tracking the law from, say, committee to enactment. If this is how Williams proceeded, she would then need to measure whether, at each stage in legislative process, the materials (*e.g.,* committee reports, floor debate, and so on) support her proposition or do not. Measurement here is likely to take the form of a categorization—the report or statement "Supported" or "Did Not Support" the power in question. In developing this categorization or measure Williams, of course, would need to be explicit about what she would count as a statement in support or not in support. That is what we mean when we say that researchers must define the standard for categorization or measurement.

How should scholars evaluate their measurement methods? Virtually all efforts to do so involve assessments on two critical dimensions – *reliability* and *validity*.

---

[248] Lawless & Murray, An Empirical Analysis of Bankruptcy Certiorari, supra note 59, at 117.

[249] For a description of these scores, see infra p. 63

[250] See supra p. 16 and note 61.

[251] Williams, "The Securities and Exchange Commission and Corporate Social Transparency," supra note 234.

1. Reliability

Reliability is the extent to which it is possible to replicate a measurement, reproducing the same value (regardless of whether it is the right one) on the same standard for the same subject at the same time. If any one of us stepped on the same bathroom scale 100 times in a row and if the scale were working reliably, it would give us the same weight 100 times in a row—even if that weight is not accurate. (In contrast, a scale that is both reliable and valid will give a reading that is both the same and accurate 100 times in a row.)

In other words, in empirical research we deem a measure reliable when it produces the same results over and over again regardless of whom or what is actually doing the measuring. Say Williams developed the following measure to assess whether floor statements over the SEC Act supported her theory that the SEC has the power to require social disclosure: if the speaker claimed that the SEC has broad powers to define disclosure, then the speaker supported it. Also suppose that she then classified a statement made by Senator Duncan Fletcher, then-Chair of the Senate Committee on Banking and Currency, as supporting her thesis but another researcher, *using her measurement procedure as described in her article*, did not classify the same statement as supportive. This would provide some evidence that her measure is unreliable.

Why should unreliable measurement procedures concern us? A major reason is that they might provide evidence that the researcher, however inadvertently, has biased a measure in favor of his or her pet hypothesis. Assume Williams, on her measure, categorized 90% of all materials in the legislative record as supporting her theory but another researcher, using her same procedures, found only 10% supportive. We would have reason to believe that the measure was not only unreliable but that the analyst wielded it in a biased way as well.

That is why, when researchers produce measures that others cannot replicate, it is the researchers' problem; they, not the replicators, must take responsibility. But what specifically has gone wrong for the researchers? A major source of unreliability in measurement is vagueness: If researchers cannot replicate a measure it is probably because the original study did not adequately describe it. Return to the Revesz study,[252] and recall that the author is interested in whether judges' policy preferences affect their dispositions of environmental cases. To measure policy preferences, he uses the party of the appointing President—surely a reliable measure (though not necessarily a valid one, as we suggest in the next section); as for dispositions, he creates a simple dichotomy, whether the court reversed or not. While he tells us that he treats "remands" as "reversals,"[253] he does not report how he characterizes the dispositions listed in Table 5 (under "Value Label")—all of which, according to the U.S. Court of Appeals Data Base (a public data base containing scores of attributes on cases decided in the courts of appeals between 1925 and 1996[254]), have occurred in the Nation's circuit courts.[255]

---

[252] Revesz, Environmental Regulation, Ideology, and the D.C. Circuit, supra note 231.

[253] Revesz, Environmental Regulation, Ideology, and the D.C. Circuit, supra note 231, at note 29

[254] These data are available at: http://www.ssc.msu.edu/~pls/pljp/ctadata.html

[255] Chief Judge Harry T. Edwards Harry T. Edwards, "Collegiality and Decision Making on the D.C. Circuit," 84 Va L Rev 1335 (1998), in a critique of the Revesz article, raises a similar objection:

…there is the question of the methodology Revesz used to code and classify the cases that provided the raw data for the study. In the technical terminology of agency review, a panel of the court may grant or deny a petition for review of agency action. When it grants a petition for review, the panel may vacate the action under review. Such a vacation may or may not be accompanied by an express remand to the agency for further action or explanation. While remand often includes vacation of an order, it need not do so. The court may remand without vacating an order, thus leaving the order in place pending further explanation. The court may even dismiss the petition for review but still remand.

Table 5. Possible Dispositions in Cases Decided by the U.S. Courts of Appeals

Value	Value Label
0	stay, petition, or motion granted
1	affirmed; or affirmed and petition denied
2	reversed (including reversed & vacated)
3	reversed and remanded (or just remanded)
4	vacated and remanded (also set aside & remanded; modified and remanded)
5	affirmed in part and reversed in part (or modified or affirmed and modified)
6	affirmed in part, reversed in part, and remanded; affirmed in part, vacated in part, and remanded
7	vacated
8	petition denied or appeal dismissed
9	certification to another court

If another researcher attempts to replicate (or *backdate* or *update*) Revesz's study, should that researcher characterize an "affirmed in part, vacated in part, and remanded" as a reversal or an affirmance? Neither the researcher nor we can answer that question because Revesz does not tell us. Rather, in this circumstance, we would have to make judgment call, which may or may not be the same one as Revesz made. This detracts from the reliability of his measure.

As a rule, then, human judgment should be removed as much as possible or, when judgment is necessary, the rules underlying the judgments should be clarified enough to make them wholly transparent to other researchers. The key to producing reliable measures, in fact, is to write down a set of very precise rules for the coders (*e.g.*, those who are reading the case, noting the value of the disposition, and typing that value into a computer software package) to follow—with as little as possible left to interpretation and human judgment. This list should be made even if the investigator codes the data himself or herself, since without it, others would not be able to replicate the research (and the measure). Along these lines, an important rule of thumb is to imagine that the researcher had to assign a first-year law student the task of classifying each case by its disposition, and that the only communication permitted between the researcher and the student was through a written appendix to the article detailing the coding scheme. This is the way to conduct research and how it should be judged.

To see this process in action, let us return Revesz's study. As a first step, even though Revesz was interested solely in whether a disposition affirmed or reversed the lower court, he might have

---

In other words, there is considerable nuance in the options available to the reviewing court. Revesz ignores this range of nuances, possibly because taking account of their subtlety would have made it more difficult for him to draw sweeping conclusions.

Revesz also makes a technical error that may skew his data: he simply treats all remands as "reversals" of agency action, ignoring the possibility of remand without vacation of the order, under which remand is explicitly not accompanied by reversal. Such remands without vacation do occur; and they do not fit Revesz's metric. The existence of such remands without vacation underscores the complexity of the review process, which is in some ways a dialogue between the court and the agency. It is precisely this complexity and dialogic character that Revesz misses by referring to case outcomes as "affirmances" or "reversals."

Richard L. Revesz, "Ideology, Collegiality, and the D.C. Circuit: A Reply to Chief Judge Harry T. Edwards," 85 Va L Rev 805 (1999) responds to this criticism by (1) reiterating why he treated remands as reversals and (2) asserting that the Chief Judge's "distinction between remands that vacate the agency's decision and those that do not may be of current importance, but it was not during the period of study." To us, this response misses the point. The long and short of it, as we explain in the text, is that Revesz would have been far better off had he coded dispositions as finely as did, say, the U.S. Court of Appeals Data Base research team.

been better off starting with all possible dispositions as reported in the U.S. Court of Appeals Data Base or some other authoritative source. (Table 5 displays those in the data base.)

To be sure, the researcher should know which values of the variable DISPOSITION ought count as a "reverse" and which should count as an "affirm"; and we should require him to specify that (*e.g.,* values 2, 3, 4, 6,7=reverse). But starting out with the more detailed values has two clear advantages. First, whoever is coding the data will make fewer errors. Think about it this way: If Revesz tells the coder in advance to report values 2, 3, 4, 6, and 7 as "reversals," the coder must take two steps: identify the disposition and, then, identify whether it is a reversal or affirmance. But, if Revesz simply has the coder identify the disposition, then the coder has only one step to take. Since every step has the possibility of introducing error, researchers should seek to reduce them. A second advantage comes when Revesz turns to analyzing his data. Because he has now coded the variable DISPOSITION quite finely, he will be able to ascertain whether any particular coding decision affects his conclusions. For example, suppose that he counts value 6 as a "reverse," even though the Court affirmed in part. Since this represents a judgment on his part (though one he should record, thereby enabling others to replicate his measure) and since the converse coding (counting 6 as an "affirm") is plausible, he will be able to examine the effect of his judgment on the results.

Next, researchers must supply clear coding instructions—those that they or anyone else could follow without having to consult with them. The following, from the U.S. Court of Appeals Data Base, provides an example:

> This variable records the disposition by the Court of Appeals of the decision of the court or agency below; i.e., how the decision below is "treated" by the appeals court. That is, this variable represents the basic outcome of the case for the litigants. [The variable takes the following values (see Table 5 above), which the coder should take verbatim from the Court's opinion.][256]

Finally, researchers ought recognize that even with these explicit instructions, errors in coding will occur. That is because they or the coder may incorrectly record a value or misread the Court's decision. What they should do is attempt to estimate this error by conducting a reliability analysis. A simple way to accomplish this is to draw a random sample (perhaps 10% of the cases in the study) and have another researcher recode them. This was the approach taken by those who created the U.S. Court of Appeals Data Base:

> To check the reliability of the coding, a random sample of 250 cases was selected from the 15,315 cases in the data base. This sample of 250 cases was then independently coded by a second coder and the results of the two codings were compared. Three measures of reliability are reported, [including]…the simple rate of agreement (expressed as a percentage) between the code assigned by the first coder and the code assigned by the second coder is reported.

Law review studies rarely undertake even this simple sort of reliability analysis, but authors could easily do so—and, more to the point, ought want to do so. For, assuming that researchers follow the procedure we set out above, they are likely to attain satisfactory results, which in turn will lead them to have more confidence in their studies. This held for the compilers of the Court of Appeals Data Base on their disposition variable: The rate of agreement between the coders was 95.2%.[257]

---

[256] The U.S. Court of Appeals Data Base, Documentation for Phase 1, at 101, available at: http://www.ssc.msu.edu/~pls/pljp/ctadata.html.

[257] While an inter-coder agreement of 95.2% seems high, whether it is high enough depends on the use to which the measure will be put. For example, this level of reliability will make researchers unable to detect differences on this variable smaller than about 5%. If greater sensitivity is needed in the results, then a better measurement procedure should be sought.

2. Validity

Earlier we noted that a bathroom scale was reliable if one stepped on it 100 times in a row and obtained the same value. This is all to the good but does not necessarily mean the scale is valid. If one's true weight is 150 and the scale, even 100 times in row, reports 125, we would not think much of the scale. It is this concern with accuracy that validity implicates. Validity is the extent to which a reliable measure reflects the underlying concept being measured. A scale that is both reliable and valid displays the weight of 150, 100 times in a row; a scale that displays a weight of 125, 100 times in a row is reliable but not valid.

Just as a bathroom scale can be reliable but not valid, so too can measures that scholars invoke.[258] Consider Cross and Tiller's investigation into the effect of various factors, including the the policy preferences of the judges, on U.S. Court of Appeals decisions.[259] The authors measure policy preferences in the same way as did Revesz— by the party of the President who appointed them. Undoubtedly, this is a measure that would produce high inter-coder agreement: If the coders had a list of the party membership of every President—a list on which we would all agree—and knew which President appointed a particular judge, no judgment calls are required. We would have a perfectly reliable measure. But does this measure accurately capture the underlying concept of "policy preference"? Revesz thinks so, deeming this a "proxy…likely to be fairly good"; Cross and Tiller apparently agree.[260] For some purposes, they may be entirely right but, unfortunately, at least for the use to which they put it, many scholars would take issue with their inference. They might point to an assumption underlying the Revesz/Cross and Tiller measure—namely, all Republican Presidents are conservatives and all Democratic Presidents are liberal—and argue that data show otherwise. On Segal's measure of presidential liberalism, for example, Bill Clinton is ideologically closer to Richard Nixon than to Lyndon Johnson.[261] Or as Giles and his colleagues write, "presidents of the same political party vary in their ideological preferences. Eisenhower is not Reagan. Indeed, the empirical record demonstrates that the voting propensities of the appointees of some Democratic and Republican presidents do not differ significantly."[262] They also might suggest that

---

[258] This raises the question of whether a measure can be valid but not reliable. The simple answer is that for such a measure we would not raise the question of validity; we would develop a different measure. To see why, return to the example of the bathroom scale. If one stepped on it once and obtained an accurate reading but then stepped on it again and obtained an inaccurate reading, one might conclude that the scale was "broken" and replace it with a new model.

The idea behind this question is better addressed in the context of unbiasedness and efficiency (in that a researcher would generally prefer a measure that is slightly biased but much less variable than another that is unbiased but very variable), concepts we introduce shortly

[259] Cross & Tiller, "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals," supra note 160.

[260] Revesz, Environmental Regulation, Ideology, and the D.C. Circuit, supra note 231, at note 6. In response to Chief Judge Edwards', Collegiality and Decision Making on the D.C. Circuit, supra note 255, at 1347, criticism of this measure Revesz, Ideology, Collegiality, and the D.C. Circuit, supra note 255, at 824 provides additional justification for it: Since "information is not available to researchers" that would enable them to construct a direct measure of ideology, "there is no alternative but to rely on some proxy for ideology. The proxy I used has been employed in a large number of studies before mine, many of which I cite in my article." But see infra p. 67.

We can only assume that Cross & Tiller, "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals," supra note 160, agree with Resevez since they provide no rationale for their measurement choice.

[261]Bill Clinton's score is 55, Nixon's is 38.5, and Johnson's is 77.7. For information on how Segal constructed his scores, see Jeffrey A. Segal, Richard J. Timpone, & Robert M. Howard, Buyer Beware? Presidential Success through Supreme Court Appointments, 53 Political Research Quarterly 557, 560 (2000).

[262]Micheal Giles, Virginia Hettinger, & Todd Peppers, An Alternative Measure of Preferences for Federal Judges, Political Research Quarterly (2001, forthcoming), at 5.; see also Donald R. Songer & Susan Haire, "Integrating

another of Revesz's/Cross and Tiller's assumptions—namely, all Presidents are motivated to appoint judges who reflect their ideologies—does not match up with various studies delineating other presidential motives.[263] Finally, they would argue that the measure neglects an important institutional feature of the appointment process—namely, senatorial courtesy—which may have the effect of constraining the President from nominating a candidate to the lower federal courts who mirrors his ideology.[264]

In short, ensuring that a measure is reliable says nothing about the degree to which it is valid, and so we need to ensure validity as well and simultaneously with reliability. Unfortunately, assessing validity is more difficult than assessing reliability. Since all conclusions about the world are uncertain,[265] we need to identify what valid means when we know that no measurement procedure will yield the truth exactly in all applications. That is, we must recognize that even the concepts of "right" and "wrong" are probabilistic categories.

To define validity in a world where all decisions and comparisons are uncertain, scholars have developed a number of criteria that they have put to use in quantitative and qualitative empirical research. We consider three—facial validity, unbiasedness, and efficiency— with the caveat that no one is always necessary and the collection is not always sufficient, even though together they are often helpful in understanding when a measure is more or less valid.

The first and most basic is face validity. *A measure is facially valid if it is consistent with prior evidence*, including all quantitative, qualitative, and even informal impressionistic evidence. Facial validity is not a casual armchair judgment of plausibility but instead requires a careful comparison of the new measure with prior evidence.

Because most people recognize that not all Democratic or Republican presidents are alike, aspects of a measure of the policy preferences of judges that relies on the party of the president who appointed them may not pass this test. But consider another measure that might—one developed by Segal and Cover.[266] To derive it, the researchers content-analyzed newspaper editorials written between the time of justices' nomination to the U.S. Supreme Court and their confirmation. Specifically, and as they tell it,

> We trained three students to code each paragraph [in the editorial] for political ideology. Paragraphs were coded as *liberal, moderate, conservative,* or *not applicable*. Liberal statements include (but are not limited to) those ascribing support for the rights of defendants in criminal cases, women and racial minorities in equality cases, and the individual against the

---

Alternative Approaches to the Study of Judicial Voting: Obscenity Cases in the U.S. Courts of Appeals," 36 Am J Pol Sci 963 (1992)

To provide but one example, Susan B. Haire, Martha Anne Humphries, and Donald R. Songer, "The Voting Behavior of Clinton's Courts of Appeals Appointees," 84 Judicature at 278, report that: "In contrast to Reagan-Bush appointees, the Clinton appointees offered substantially more support to the liberal position in civil rights claims. When voting on criminal and economic issues, [however], Clinton judges generally adopted positions that were strikingly similar to those taken by judges appointed by moderate Republican [Presidents]."

[263] See, for example, Sheldon Goldman, Picking Federal Judges (1997), arguing, on the basis of detailed archival work, that presidents seek to advance one or some combination of three agendas—personal, partisan, or policy—when they make judicial nominations. Personal agenda refers to using the nominating power to please a friend or associate; partisan agenda means using nominations as vehicles for shoring up electoral support for their party or for themselves within their party; and policy agenda is about using nominations to enhance the substantive policy objectives of an administration.

[264] Giles, Hettinger, & Peppers, An Alternative Measure of Preferences for Federal Judges, supra note 262, at 5.

[265] See supra p. 34.

[266] Jeffrey A. Segal & Albert D. Cover, "Ideological Values and the Votes of U.S. Supreme Court Justices," 83 Am Pol Sci Rev 557 (1989). We realize that Segal and Cover designed this measure to tap the preferences of Supreme Court justices, and not court of appeals judges. We discuss an alternative approach to measuring the preferences of lower federal appellate judges on p. 67.

government in privacy and First Amendment cases. Conservative statements are those with an opposite direction. Moderate statements include those explicitly ascribe moderation to the nominees or those that ascribe both liberal and conservative values.[267]

They then measure judicial policy preferences by subtracting the fraction of paragraphs coded conservative from the fraction of paragraphs coded liberal and dividing by the total number of paragraphs coded liberal, conservative, and moderate. The resulting scale of policy preferences ranges from -1 (unanimously conservative) to 0 (moderate) to +1 (unanimously liberal). The first column of numbers in Table 6 displays the results of their efforts.

Table 6. Assessing Policy Preferences of Supreme Court Justices Appointed Since 1953[268]

Justice	Segal/Cover Score[269]	Actual Voting
Brennan	1.00	79.5%
Fortas	1.00	81.0
Marshall	1.00	81.4
Harlan	.75	43.6
Goldberg	.50	88.9
Stewart	.50	51.4
Warren	.50	78.7
Ginsburg	.36	61.5
White	.00	42.4
Whittaker	.00	42.9
Breyer	-.05	59.6
O'Connor	-.17	35.3
Kennedy	-.27	37.0
Souter	-.34	57.1
Stevens	-.50	63.6
Powell	-.67	37.4
Thomas	-.68	25.3
Burger	-.77	29.6
Blackmun	-.77	52.9
Rehnquist[270]	-.91	21.6
Scalia	-1.00	29.7

It is easy to see why many scholars deem this measure facially valid. To be sure there are some exceptions (*e.g.,* Warren seems more liberal than his score; Thomas seems more conservative than his) but the overall results it yields comport with scholarly impressions of the justices. Brennan and Marshall, generally regarded as liberals, receive scores of 1.00; Scalia and Rehnquist, generally regarded as conservatives, receive scores of –1.00 and -.91 respectively.

---

[267] Id., at 559.

[268] Data Sources for the Segal/Cover Scores: Segal and Cover, Ideological Values and the Votes of U.S. Supreme Court Justices," supra note 266, at 560; Segal, et al., Ideological Values and the Votes of U.S. Supreme Court Justices Revisited, supra note 62, at 816; Data Source for Actual Voting: Lee Epstein, et al. The Supreme Court Compendium (2001), supra note 62, at Table 6-2.

[269] The Segal/Cover values are from 1.00 (most liberal) to -1.00 (most conservative); they were derived from content analyses of newspaper editorials prior to confirmation. Actual voting is the percentage of liberal votes cast by the justice over the course of his or her career in civil liberties cases (through the 1998 term).

[270] Rehnquist received the same score as an associate justice and as chief justice.

This is a quantitative measure but those developed for qualitative research are easy enough to subject it to similar tests of face validity. Consider a study by Gerry, seeking to address the question of whether state courts interpret the U.S. Constitution differently than the federal courts.[271] To answer it he focuses on lower federal and state court reactions to a U.S. Supreme Court decision, *Nollan v. California Coastal Commission*;[272] to tap differences in interpretation, he develops several measures of judicial reasoning—one of which is the degree of deference the lower court appears to give to the government's action. After reading each opinion, he codes it as (A) a "highly deferential" reading of *Nollan*, "in which courts demand only that government actors meet the most minimal standards of means-ends scrutiny;" (B) an "intermediate deference" to a government action; or (C) those "nondeferential" interpretations of *Nollan* "that require reviewing courts to conduct probing inquiries into the government's conduct in regulating land use."[273] If we are familiar with the cases in Gerry's study, we can approach face validity in much the same way we did for Segal and Cover's scores, that is, by asking ourselves whether the overall results yielded by his measurement scheme sit comfortably with our prior knowledge of the reasoning used in the cases.

In addition to being facially valid, measures should be approximately *unbiased*. We say *a measurement procedure is unbiased if it produces measures that are right on average across repeated applications*, that is, if we apply the same measurement procedure to a large number of subjects, sometimes the measure will be too large and sometimes too small, but on average it will yield the right answer. Suppose we asked a 100 people to step on a bathroom scale. Our scale would be unbiased if each person who was measured was reported to be a bit too heavy or a bit too light, but the errors in reported weights that were too large were about the same size and number as the errors in reported weights that were too small. An example of a biased procedure would be to ask subjects for their (self-reported) weights. In all probability, some would give accurate answers, or answers that were right on average; but others would respond to the social situation and underestimate their weights. Since the underestimates would not be canceled out by a similar set of overestimates, the result would be a biased measure.

This example highlights an important point: Often the quickest way to create a biased measure is to develop a procedure that relies in a biased way on responses from the population under analysis. Suppose we asked justices on the U.S. Supreme Court whether their vote in *Bush v. Gore*[274] turned on their political preferences. Just as surely as people underestimate their weight, the justices would say no, that they vote on the basis of some neutral principle(s).[275] More generally, as it turns out, asking

---

[271] Brett Christopher Gerry, "Parity Revisited: An Empirical Comparison of State and Lower Federal Court Interpretations of Nollan v. California Coastal Commission," 23 Harv J L & Pub Pol 233 (1999).

[272] 483 U.S. 825 (1987).

[273] Gerry, Parity Revisited, supra note 271, at 273.

[274] 531 U.S. ___ (2000).

[275] We need not guess here. The day after the Court issued its decision in *Bush,* Justice Clarence Thomas, in response to a question about what role the justices' party affiliations played in their decision, responded "zero." He went on to say that "I plead with you that whatever you do, don't try to apply the rules of the political world to this institution." Neil A. Lewis, "Justice Thomas Speaks Out on a Timely Topic, Several of Them, in Fact," New York Times (December 14, 2000), at 17A . Along similar lines come responses from judges to scholarly writings suggesting that they reach decisions on grounds other than neutral principles. See, for example, Edwards, Collegiality and Decision Making on the D.C. Circuit, supra note 255, at 1359 (writing that "The main reason I am astonished by the hypotheses advanced by scholars [that votes reflect policy or partisan considerations]…is that my colleagues and I dissent so rarely from the opinions of panels on which we sit… [E]ven where there was dissent, the dissent only occurred along presumed "party" lines around half of the time. This is, in my view, extremely strong prima facie evidence of consensus among judges about the correct judgment in a given case") and Patricia M. Wald, "A Response to Tiller and Cross," 99 Colum L Rev 235 (1999) (asserting that "even in the comparatively rare instances where a judge's personal convictions have maneuvering room, they do not automatically trump the facts of the case, relevant law, and constitutional constraints on a judge's discretion").

someone to identify his or her motive is one of the worst methods of measuring motive.[276] People often do not know, or cannot articulate, why they act as they do. In other situations, they refuse to tell, and in still others, they are strategic both in acting and in answering the scholar's question. This is obvious from the example of asking justices about how they reach decisions, and it needs to be better understood by legal scholars who too often rely on this general measurement procedure. [277] Consider Miller's analysis of whether factors such as local bias lead attorneys to file in federal, rather than state, court when concurrent jurisdiction exists.[278] We have more to say about this study later but, suffice it to note for now, to assess the factors he deems relevant, Miller surveys attorneys. This strategy suffers from the problem that attorneys may have incentives to hide their sincere preferences from a researcher who they know will make policy recommendations about the need for concurrent jurisdiction; for example, lawyers who wish to retain it may provide what they believe to be the most "legitimate" rationale for concurrent jurisdiction rather than the that reflects their sincere preferences.

Creative measurement procedures need to be developed for cases like this. So, instead of (or sometimes in addition to) asking respondents to answer research questions directly, it is usually better to look for *revealed preferences*, which are consequences of theories of motive that are directly observable in real behavior. Translating this to Miller's study, he had no need to ask attorneys why they filed in federal (or state) court. Since he knew where they filed, he could have posited a series of factors and assessed them in line with the rules we outline in this Article.

Of course, even if researchers devise "creative" measures they still need to judge whether those measures are unbiased and have face validity. Typically that judgment can come about only if they have an existing measure, or develop a new one, to which they can compare the original. We can only know if our bathroom scale over- or under-estimates weight if we compare its readings to those of another scale, which we know to be correct. Likewise for the Segal and Cover scores. To see if they are biased we would want another measure of judicial policy preferences, with the justices' revealed behavior—say, the percentage of liberal votes they cast in civil liberties cases—providing one. We could thus compare the scores and votes (see Table 6) to determine the extent, if any, of bias.[279]

This approach is not limited to quantitative research. Suppose that Williams or any other scholar studying the legislative history of a particular bill developed the sort of categorization scheme we denote above: whether a document or speaker supported a specific interpretation of the bill or not. If Williams ends up classifying every speech and document as supportive, and previous scholars had concluded otherwise, we might suspect bias in the measurement procedure. We would thus advise her to take the same type of step we commend to assess the Segal and Cover scores: Develop another measure to use for comparison purposes. In this case, it might be useful to gather all

---

[276] See, for example, Richard Nisbett & Tim Wilson, Telling More than We Know: Verbal Reports of Mental Processes, 84 Psych Rev 231 (1977); Wendy M. Rahn, Jon A. Krosnick, & Marijke Breuning, Rationalization and Derivation Processes in Survey Studies of Political Candidate Evaluation, 38 Am J Pol Sci 582 (1994).

[277] This is not to say that legal academics fail to appreciate the problem; some, in fact, do, as the following, in Mann, "Explaining the Pattern of Secured Credit," supra note 110 at note 33, suggests: "I…recognize that any attempt to use interviews to evaluate complex environments is subject to the problem that the interview subjects mat not be able to explain the motivations for their actions to the interviewer." But it is to say that mere acknowledgement is insufficient; that scholars, as we suggest below, must take remedial action.

[278] Neal Miller, An Empirical Study of Forum Choices in Removal Cases under Diversity and Federal Question Jurisdiction, 41 Am U L Rev 369 (1992).

[279] In fact, scholars have undertaken this task, and found high correlations between the scores and the justices' votes. See Segal & Cover, Ideological Values and the Votes of U.S. Supreme Court Justices, supra note 266; Segal et al., Ideological Values and the Votes of U.S. Supreme Court Justices Revisited, supra note 62, at least in some areas of the law, see Lee Epstein & Carol Mershon, "Measuring Political Preferences," 40 Am J Pol Sci 260 (1996).

scholarly analyses of the bill at issue in Williams' study, the Securities and Exchange Act of 1934. Her interpretation need not be the same as that offered by others, but she should explain any differences, provide reason(s) why previous studies are wrong, and analyze the reason(s) (perhaps stemming from different measurement procedures) prior research came to the conclusions that it did.

Related to unbiasedness is a third important criterion for judging validity, *efficiency*. Efficiency helps us choose among several unbiased measures, with the basic idea being to choose the one with the minimum variance. For example, if we had access to two bathroom scales that were each unbiased but one had smaller errors in any one measurement, we would choose that scale. Efficiency, in other words, indicates the degree of reliability for unbiased measures.

To see the implications for empirical research, return to the Cross and Tiller study.[280] While the authors use the party of the President to tap the political preferences of circuit court judges, many other measures are possible.[281] Giles and his colleagues offer one,[282] a measure that takes into account the political preferences of the President (via common space scores for Presidents developed by Poole[283]), as well as the preferences of Senators who may be involved in the appointment process through senatorial courtesy (via common space scores for Senators offered by Poole and Rosenthal[284]). We know from scholarly research that both the Cross/Tiller and the Giles measures, on average, are able to predict how judges vote, that is, they will yield the same answer for many judges (they are, in other words, unbiased). But we also know that, sometimes, for some judges, the predictions produced by Cross/Tiller's measure will be way off.[285] That is because the measure is inefficient. It disregards information—important information, as it turns out—about the appointment process: When a senator is of the same party as the president and the vacancy is from the senator's state, the senator can exert considerable influence on the selection of judges. His or her influence will sometimes produce a more liberal judge, and sometimes a more conservative one, such that on average no bias is introduced. But omitting information such as this is the definition of inefficiency. And it is for this reason that we would prefer the Giles measure to Cross/Tiller's.

More generally, whenever researchers are confronted with two unbiased measures, they should normally select the more efficient one. For individual applications of the measurement procedure based on more information *(e.g.,* Giles's) yield measures that cluster more narrowly around the true answer than does the one based on less (*e.g.,* Cross/Tiller's). The result is that any one application of the measure with more information will be likely to yield an answer closer to the truth than any one application of the measure with less.

This holds for quantitative research as our examples thus far illustrate, as well as for more qualitative work, including research that seeks to identify whether the legislative history of a law supports a particular interpretation of that law. Reconsider Williams's study of whether the legislative history of the Securities and Exchange Act of 1934 enables the SEC to require some

---

[280] Cross & Tiller, "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals," supra note 160.

[281] Recall that Revesz, "Environmental Regulation, Ideology, and the D.C. Circuit," invokes the same measure. Hence, the concerns we express in this paragraph apply equally to his research.

[282] Giles, Hettinger, & Peppers, An Alternative Measure of Preferences for Federal Judges, supra note 262.

[283] Keith T. Poole, "Recovering a Basic Space from a Set of Issue Scales," 42 Am J Pol Sci 954 (1998). These scores are based on the CQ Presidential Support Roll Call Votes. The scores, along with a detailed description of them, are available at: : http://voteview.uh.edu/

[284] Keith T. Poole & Howard Rosenthal, Congress: A Political-Economic History of Roll Call Voting (1997). These scores are based on all non-unanimous roll call votes cast each year. The scores and a detailed description of them are available at: http://voteview.uh.edu/

[285] See, for example, Giles, Hettinger, & Peppers, An Alternative Measure of Preferences for Federal Judges, supra note 262.

degree of social disclosures.[286] Suppose that Williams measures legislative history by content analyzing speeches, and only speeches, delivered on the floor of Congress. Undoubtedly, if we had some way to know the "truth"—exactly what members of Congress intended—Williams's approach would sometimes reflect it and sometimes it would not. If speeches revealed congressional intent, as they sometimes do, Williams would get it right; if speeches were delivered for other reasons, such as to appease constituents, she might not. But this does not necessarily mean that her measure is biased, for the concept of unbiasedness is that the measure does not work in a particular direction. If legislators make speeches to explain their positions to constituents rather than to articulate legislative history, it would not be clear that such speeches support a systematically higher or lower degree of social disclosure than would be accurate. Now suppose that another researcher comes along and measures intent by considering speeches, committee reports, testimony, and the like. Even with all the additional information, this approach too will not yield the right answer every time because some of the same confounding factors (*e.g.,* legislators making speeches to appease constituents) will affect the measure. But, again, if both these measurement strategies are in fact unbiased—right on average, but not necessarily right for any one application—we would prefer the second because it is based on more information; it is more efficient than the first.

We could go on, for the scholarly literature on measurement is immense and we have barely begun to summarize it. Indeed, entire fields of study are devoted to measuring psychological well-being, health, income, education, happiness, survey responses, intelligence, and numerous others. We write this not necessarily to exhort law professors to immerse themselves in this vast literature but rather to register a cautionary note (*i.e.,* we have just scratched the surface on this topic) and, more important, to emphasize that their studies need not occur in a vacuum. If the research calls for measuring income, talk to an economist. If it requires a measure of democracy, it is easy to find political scientists and sociologists who have spent their careers working on the subject.

B. Estimation

Estimation is measurement writ large. The idea is to divide the inference task into *two* steps. In the first, researchers take measures about a single thing, that is, they ultimately make an inference about a unit of observation. Doing so requires them to decide on a standard of measurement, on a specific measure, on how to select observations on that measure, and, then, on the method for drawing those observations. Those wishing to study legislative intent, for example, must select a measure of legislative intent that is reliable and valid (*e.g.,* floor speeches, committee reports, and so on). Next, as we discuss in Part VIII, they must specify how they plan to select their observations so that they are able to make an inference about the population; for example, will they take into account all speeches, draw a random sample of speeches, create a sample of those speeches that are visible to them, or invoke some other method? Finally, they must collect and code those observations. For a study of legislative intent, this may amount to determining whether each speech (the unit of observation) supports or does not support a particular interpretation of legislative intent.

In the second step, the researcher also makes an inference but at the level of a population rather than at a single unit of observation. This involves identifying a quantity of interest in the population and invoking a set of measures to estimate that quantity. Suppose researchers draw a random sample of speeches delivered on the floor of Congress (with each speech constituting the unit of observation). What probably does not interest them much is whether a particular speech supported a particular interpretation of legislative intent. What they want to know is whether, if they had collected all speeches, would those speeches (on average) support a particular interpretation of

[286] Williams, "The Securities and Exchange Commission and Corporate Social Transparency," supra note 234.

legislative intent. They would then use a set of measures to estimate this quantity, the mean, in the population

To evaluate good estimates, we use two of the same criteria we discussed above for measures—unbiasedness and efficiency. To these, we add one more—*consistency*. The idea here is that as we include more measurements of more subjects in our estimate, we should get closer and closer to the truth. In the example above, the more and more speeches researchers are able to collect the more the mean in their sample will approximate the mean in the target population.

Examples of these are difficult to find in the law reviews. The problem, we suspect, lies rather in the failure of scholars to document precisely the procedures that led them to the estimates. We see evidence of this in quantitative legal research but it seems to occur with greater frequency in qualitative empirical work—such as doctrinal analyses or investigations of intent—in which scholars rarely specify how they determined whether court decisions or legislative materials, on average, support the interpretation they offer.

This is unfortunate because making such estimates is often exactly what authors of these sorts of studies wish to do and, in fact, do. But failure to explain their procedures makes their inferences way too uncertain. Consider Kramer's[287] attempt to counter critics of Wechsler's "The Political Safeguards of Federalism," which argued that "the existence of the states as governmental entities and as the sources of the standing law is in itself the prime determinant of our working federalism, coloring the nature and the scope of our national legislative processes from their inception."[288] On this account, states do not need to look to the U.S. Supreme Court for protection from congressional incursions into their power; actually, the Court is "on weakest ground when it opposes its interpretation of the Constitution to that of Congress in the interest of the states."[289] Wechsler's critics counter by arguing the Founders "wanted and expected the Supreme Court to protect the states from overreaching by Congress."[290] In his effort to sort through these competing claims, Kramer could have mined the historical and case records in accord with the rules inference and, particularly relevant here, estimated the degree to which, say, earlier Court decisions, on average, supported Wechsler's or the critics' assertions. But these were not the steps he chose to take. Rather, the essay is replete with "estimates"—*e.g.,* the justices "did *nothing* to restrict national power vis-à-vis the states during the entire antebellum period"[291]—at which we are unsure of how Kramer arrives.[292]

The same problem afflicts Lin's qualitative study of "traditional" narratives invoked by courts in cases involving same-sex adoptions—narratives that make it difficult for gays and lesbians to establish families. Like Kramer, the author offers numerous estimates—*e.g.,* "Courts *often* embrace the misconception that lesbian and gay parents, through interaction with their children will somehow cause the children to become homosexual"—that he supports with little more than string citations.[293]

---

[287] Larry D. Kramer, "Putting the Politics Back into the Political Safeguards of Federalism," 100 Colum L Rev 215 (2000).

[288] Herbert Wechsler, The Political Safeguards of Federalism: The Role of the States in the Composition and Selection of the National Government, 54 Colum L Rev 543 (1954)

[289] Id.

[290] Kramer, Putting the Politics Back into the Political Safeguards of Federalism, supra note 287, at 227.

[291] Id., at 228; our emphasis

[292] We realize that some sort of scholarly consensus may exist over such "estimates," but that is not the issue here; it is rather whether or not the estimates are way off the mark—a question we cannot answer because we do not know precisely how the researcher arrived at them.

[293] Timothy E Lin, Social Norms and Judicial Decisionmaking: Examining the Role of Narratives in Same-Sex Adoption Cases, 99 Colum L Rev 739, 775 (1999); our emphasis

Does this mean that Kramer's and Lin's estimates are necessarily way off? No. But, because neither specified the procedures that led to them, it is impossible to know whether they are or are not. We simply cannot say, with any degree of certainty, if the estimates the researchers offer support their conclusions.

VIII. Selecting Observations

As we suggest above, a crucial bridge between measurement and estimation is the selection and collection of observations. The question here takes the following form: How should researchers select observations to include in their studies? Suppose a scholar wants to understand, via an analysis of court decisions, why judges depart from the federal Sentencing Reform Act guidelines. How should she decide on which cases to collect data? What about the researcher who wants to study whether police comply with *Miranda* by interviewing police? How ought he determine which police to interview?

There are many rules to answer these questions. Below we review four that are, regardless of whether the research is qualitative or quantitative, essential to reaching valid inferences: (1) identify the population of interest; (2) collect as much data as is feasible; (3) record the process by which data come to be observed; and (4) collect data in a manner that avoids selection bias.

A. Identify The Target Population

When we collect data to make inferences, a critical step is to identify the target *population* (or "population of interest"). *This is all subjects, cases, countries, or other units in a specified time frame abut which the researcher would collect information if time and resources were unconstrained.* If the goal is to estimate the average age in the United States, then the population of interest includes all human beings *presently* living in the United States, where the investigator clearly and precisely defines the concepts of "human beings," "United States," "living," and "age." It should be possible, in principle even if not in practice, to collect data on all members of this population, and the definition of the population should be sufficiently clear that no ambiguity exists as to who is included and excluded.

This may seem a rather simple task to execute, but various examples in the legal literature suggest otherwise. Consider Friedman's detailed examination of the history of "countermajoritarian difficulty,"[294] a term Bickel coined to reflect the "problem" of allowing unelected judges to strike down legislation passed by elected representatives.[295] In developing that history, Friedman attempts to refute a piece of conventional wisdom; namely, criticism of the use of judicial review by the Court ran in "one straight arrow from the time of *Lochner* through the New Deal,"[296] with the nature of those critiques taking similar forms. To accomplish this, Friedman details a few specific critiques that arose during the Populist/Progressive Era and the New Deal period. For the former he points to, among others, a comment made by Theodore Roosevelt: "Here the courts decide whether or not…the people are to have their will;" for the New Deal period he cites a line written to Franklin Roosevelt by a "correspondent:" "Nine OLD MEN whose total age amounts to about 650 years, should have additional help."[297]

---

[294] Alexander M. Bickel, The Least Dangerous Branch of Government: The Supreme Court at the Bar of Politics (1962), at 16.

[295] Barry Friedman, "The History of the Countermajoritarian Difficulty, Part Four: Law's Politics," 148 U Pa L Rev 971 (2000).

[296] Id., at 985.

[297] Id., at 986.

Friedman's goal thus is to reach a descriptive inference, that is, to use this sample of quotes to make a general claim about the population. But what is the population? All criticisms leveled by anyone against the Court? Those made to or by politicians? In the media? Recorded in history books? And what is the time frame of the eras of interest? Does the New Deal period begin with Roosevelt's election or with his first economic proposals? Does it end with World War II, FDR's death, or something else? It is difficult to answer these questions from Friedman's examples and narrative because he never specifies the precise target of his inference. Empirical scholarship requires less ambiguity, since without clearly identifying the target population, evaluating the quality of the inferences and uncertainty of his conclusions becomes impossible. Even though Friedman may be unable (for whatever reason) to investigate every single element in his population, he should be clear about what that population is and, and, at least in theory, be able to identify all its members. The general point is critical but mundane: A researcher can accomplish a goal more easily if the goal is clearly identified.

Friedman's is a qualitative study but the same holds for quantitative work, such as Veilleux's investigation into what she hypothesizes are the major causes of a decline in the proportion of stays of execution granted in the federal courts: changes in interpretation regarding the abuse of writ doctrine and increased attention to the interests of states.[298] To investigate these explanations, that is, to make a causal inference, she examines a sample of stay decisions *published* between *1981 and 1995* by the U.S. Supreme Court, the U.S. Courts of Appeal for the Fifth and Eleventh Circuits, and all federal district courts located in those circuits.[299] This much she tells us; what she neglects to define, however, is the relevant population about which she wants to make the casual inference. Is it *all* stays granted by all federal courts since the *beginning of time*? It is all *published* stays? Is it all *published* stays since *1981*?

The latter seems a reasonable one but we should not have to guess—as we must in this case as well as in Fisher's article attempting to explain the entrenchment of plea bargaining. At the onset, Fisher writes the following:

> We need to follow the course of plea bargaining's ascent to learn the source of its strength. I will tell this story as it unfolded in America, for although the earliest instances of plea bargaining may well have happened elsewhere and although plea bargaining in time would spread across the common-law world and beyond, it triumphed here first. Within America, I will focus on Massachusetts…
>
> Within Massachusetts, I will focus mainly on its largest county – Middlesex… where I practiced as a prosecutor, an experience that left me familiar with the ways of its courts and perhaps more aware of the ways in which things have changed.
>
> Within Middlesex County, I will look most closely at the middle tier of the county's judicial system, which had jurisdiction over all but the most serious crimes.[300]

But what is the actual target of his inference? Based on his description and our reading of his essay, it could be any of the following: plea bargaining in (1) the world (2) the United States; (3) Massachusetts; (4) Middlesex, the largest county in Massachusetts; (5) the middle tier of the county's judicial system; or perhaps somewhere else. The list of possibilities is seemingly endless but it should not, and need not, be; the author should make known the target of every inference.

---

[298] Nicole Veilleux, "Staying Death Penalty Executions: An Empirical Analysis of Changing Judicial Attitudes," 84 Georgetown L J 2543 (1996). This is both a quantitative and qualitative study. The author collects numerical data on stays of execution granted and denied, and qualitatively explores explanations for increases in the fraction denied.

[299] Id., at 2551.

[300] Fisher, "Plea Bargaining's Triumph," , supra note 136**,** at 861-863.

B. Collect As Much Data as Feasible

Whether descriptive or causal, inference—learning about facts we do not know by using facts we know—requires some facts. Knowing *more* of these facts (along with the rules given here) should make for better inferences. So, to return to the examples above, if Friedman wants to make claims about the sorts of criticisms of the Court made during the Progressive/Populist and New Deal eras, that is, to use criticisms about which he has learned to make reach conclusions about the population of criticisms, he should collect as many criticisms as possible. Likewise for Veilleux and Fisher.

Simply, when an opportunity exists to collect more data, we should generally take advantage of it. We should also judge empirical research by how much information the researcher brings to bear on the inference at issue. If a scholar bases his or her inferences on relatively little information, then any conclusions will be especially uncertain. If, however, he is able to marshal a massive quantity of information, then answers to the research questions posed may even be certain enough to change the course of legal scholarship or to recommend public policies that affect many people.

Since all observations are uncertain, and all sources of observations have perhaps different types of measurement error, our advice about collecting more data does not only or necessarily mean to collect more of the same type, such as increasing the number of observations. Indeed, it can be especially useful to collect data of many different types from many different sources.

In a certain sense, this recommendation takes us back to our prior advice about listing all observable implications of a theory, even those the researcher lacks the time and resources to observe. To see this, reconsider one of the explanations Veilleux offers for the increase in denials of stays of execution: federal court deference to the states. If this theory holds, then the researcher could develop observable implications that transcend the specific legal area under analysis, that is, if federal courts defer to states on matters of who to execute, then they may be deferring in other areas as well. Veilleux could then measure and collect observations on those other areas, thereby following the rule that more data are better.

But this rule goes beyond merely listing all the observable implications; it also commends that researchers undertake broad searches for diverse types of data even over a particular implication. If Friedman is interested in the types of criticisms of the New Deal Court made by a wide range of actors, including citizens, scholars, politicians, and others, then the sorts of data he or others could collect are equally wide ranging: anthropological, ethnographic, historical, archival, survey research, aggregate data, in-depth interviews with a few people, cross-court comparisons, and even cross-country comparisons, to name just a few.

This noted, we do not mean to imply that scholars ought spend years collecting data for every individual research project; we recognize that many other constraints—personal and professional, in addition to scientific—quite reasonably affect research decisions. If offered the choice, however, researchers should almost always take the data. If they have an easy way of collecting data even partially relevant to their project, they should do so. If the procedures for gathering data for another project can be slightly amended to be relevant to the researchers' without much trouble, then do it. Since a major task confronting empirical researchers is to make inferences, basing them on more data in a appropriate way will not hurt.

C. Record the Process by which Data Come to be Observed

Regardless of how researchers go about selecting their observations, *valid inferences require information about the data-generation process*.[301] A study that gives insufficient information about the

[301] See King, Keohane, and Verba, Designing Social Inquiry (1994), supra note 1, at 23.

process by which the data come to be observed by the investigator cannot be replicated and, thus, stands in violation of the rule we articulated in Part II. Equally important, it breaks the assumed link between the facts we have and the facts we would like to know, and, as such, is of no use in making inferences about the population. Finally, as we explain below, only by knowing the process by which they obtained the data can researchers determine whether bias afflicts their inferences.

The list of law review articles that do not follow this recommendation is long. Many, such as Friedman's,[302] Lin's,[303] and Kramer's,[304] are qualitative. But quantitative studies are not immune from violating this critical rule of inference. In Part II we pointed to two, Mann's examination of secured credit[305] and Eisenberg, Garvey, and Wells's[306] study of jury sentencing in capital cases. But there are many other offenders. Anderson and Rowe's investigation into how various rules might induce settlement between parties to a lawsuit provides an interesting example.[307] To address this question, the researchers sent a computer diskette (containing an "interactive" litigation simulation program) to about 1,310 attorneys. (We say "about" because it is impossible to specify, from their description, the number in their original sample. What they do tell us is that 131 lawyers completed the simulation, which represents a response rate of "approximately 10%."[308]) But how did they draw this sample? All they say about their data-generation process is that they started with a list of lawyers (how many, we do not know) they obtained from the American Inns of Court Foundation and "sent survey materials to practitioner members of selected Inns, chosen for geographical and city-size dispersion."[309] From this description, we might surmise that certain biases exist in their ultimate sample of 131 (*e.g.,* a bias toward lawyers with enough time and knowledge to run a computer simulation) but, because the researchers provide virtually no information about the data-generation process, we cannot rule out many others.

This makes any inference they reach about whether certain rules induce settlement among lawyers valid for the *ultimate sample* of 131 lawyers (that are of little interest in and of themselves), but not necessarily valid the *intended sample* of 1310 lawyers to whom the survey was sent, and probably invalid for the target *population* of all lawyers. The only link among the population and the ultimate observed sample is the process by which the data come to be observed. All inferential methods that seek to learn about the population with data from a sample require knowledge of this process. Without this knowledge, we are left making unjustified theoretical assumptions about easily knowable facts and without substantive conclusions that are far more uncertain than necessary.

For another example, consider Mills's research on state versus clinical strategies designed to help battered women, in which she hypothesizes that particular types of state interventions—such as mandatory arrest and prosecution—may do more harm than good for the women.[310] To assess this claim, Mills conducts what she calls a "clinical analysis," comparing the strategies that social workers and other clinicians use when dealing with women who were subjected to domestic violence and those invoked by the state.[311] This comparison takes the form of an examination of the courses of

---

[302] Friedman, "The History of the Countermajoritarian Difficulty, Part Four: Law's Politics," supra note 295.

[303] Lin, Social Norms and Judicial Decisionmaking, supra note 293.

[304] Kramer, Putting the Politics Back into the Political Safeguards of Federalism, supra note 287.

[305] Mann, "Explaining the Pattern of Secured Credit," , supra note 110.

[306] Eisenberg, Garvey, & Wells, Jury Responsibility in Capital Sentencing, supra note 71.

[307] Anderson & Rowe, "Empirical Evidence on Settlement Devices: Does Rule 68 Encourage Settlement?," supra note 103.

[308] Id., at 526.

[309] Id.

[310] Linda G. Mills, Killing Her Softly: Intimate Abuse and the Violence of State Intervention, 113 Harv L Rev 550 (1999).

[311] Id., at 556.

action prescribed in particular books and articles written by clinicians, on the one side, and of select state practices, on the other. (We stipulate, for purposes of isolating one methodological issue to discuss at a time, that these sources are representative of actual empirical practice, even though we might reasonably question this and, in fact, probably should subject them to a separate empirical inquiry.)

Since Mills does not specify the target of her inference at least with regard to clinical strategies, let us suppose that it is every course of action ever prescribed by clinicians in books and articles. By the same token, since she does not reveal how she chose the particular books and articles for her analysis—in fact, the only thing we readers know for sure is that she did not collect every course of action ever prescribed in writings by clinicians—let us further assume that the observations she analyzed in her study represent the product of one of two selection mechanisms that are very common in law reviews: the researcher exercises complete discretion over what observations to include (sometimes called purposive sampling) or the researcher chooses observations because they are convenient or visible (called convenience or haphazard sampling). If either is how Mills proceeded, it is entirely possible that bias of one form or another infects her sample and, thus, any inferences she hopes to make.[312] Such biases might manifest themselves in a sample of observations by inappropriately supporting the conclusion she wanted to reach (selecting clinical writings that she believes help women) or, in the more subtle (but no less pernicious) ways we describe momentarily. Now suppose that Mills collected data on all the elements in her population—every course of action ever prescribed in writings by clinicians. (We continue to stipulate that these writings represent empirical practice.) Obviously the same sort of biases would not be present; by including all strategies suggested by clinicians the "sample" would not, for example, include only those that supported her thesis.

D. Ascertain the Process by which the Potential Observations are Generated

Just as researchers can introduce bias in their studies when they draw unrepresentative samples, the world that creates the set of *potentially observable data* can also bias inferences if it differs systematically from the target population. Hence, even if researchers follow our advice and collect all the elements in the population, they may be unable to reach valid inferences about that population.

Of particular concern is when the world, and not the investigator, invokes a selection rule such that those items that somehow make their way into the available population are correlated with the dependent variable (*Y*), even after taking into account the explanatory variable (*X*). Unfortunately, this occurs quite often in the legal world. Among the most prominent examples are studies that base their inferences exclusively on *published* opinions, rather than on the full population of *published* and

---

[312] A related selection procedure—one that leads the researcher to bias the sample against the hypothesis (or claim) of interest—might also produce bias. Lisa Bernstein, "The Questionable Empirical Basis of Article 2's Incorporation Strategy: A Preliminary Study," 66 U Chi L Rev 710, 714-715(1999) seems to provide an example along these lines. To investigate the extent to which "customary practices" "actually exist as to most aspects of contracting relationships in merchant communities," the author examines four industries—hay, grain and feed, textiles, and silk. It appears that she selects these industries because they are ones that "in an early stage of their development were roughly characterized by conditions favorable to the emergence of customs"—such as being "close knit." In other words, she seems to be biasing her sample against the hypothesis of interest (or, at least the claim she later makes); namely, that "merchant transactors do not…have similar views about the meaning of common contractual terms… She does so out of the (apparent) belief that, if she does not find the existence of customs among merchant transactors in industries in which conditions are favorable toward the development of such customs (*e.g.,* "close knit ones) then she will not find them in industries in which conditions are unfavorable. The problem here is that—if, for whatever reason, customs are harder to detect in "close knit communities" than in other sorts—she may have inadvertently biased her sample *for* the hypothesis of interest, not against it.

unpublished opinions. Such was Kerr's effort, attempting to explain judicial review of agency decisions in the post-*Chevron* era.[313] Among the explanations he offers, as we noted earlier, is a contextual theory: Judges continue to use "traditional" factors in adjudicating *Chevron* cases, rather than the two-step *Chevron* test; in other words, they defy the U.S. Supreme Court. To assess this explanation, Kerr collected data on every circuit court decision *published* between 1995-1996 that applied *Chevron.* Now, if all Kerr wants to do is reach inferences about *how circuit court judges dealt with Chevron in opinions they published between 1995-96* (*i.e.,* if this is target of his inference), then his approach is more than acceptable; it actually comports with our advice about collecting all elements in the population. If, however, he attempts to make more general claims about how *Chevron* has fared in the nation's courts of appeals—as he does in the conclusion of his article—his inferences are unfounded. That is because the legal world has conspired against him in a way for which he did not compensate. That world—judges, really—invokes a selection rule that may be correlated with *Y*, given *X*: the rule governing publication of opinions. While this rule commends that judges publish only those opinions that are "of general precedential value,"[314] a rather large body of literature suggests that the rule is sufficiently vague to permit circuit court judges to publish or not as they see fit.[315]

Suppose then, as the literature also suggests, that court of appeals judges make strategic use of their discretion,[316] for example, publishing opinions that follow precedent established by U.S. Supreme Court (*e.g., Chevron*) and failing (with the goal of avoiding reversal) to publish those opinions that do not. If this is so, then any inferences Kerr reaches solely on the basis of published opinions will be biased, and that bias is in a predictable direction: He will *overestimate* the effect of *Chevron.* Which is exactly what Kerr may have done. Based on his data, he concludes that circuit

---

[313] Kerr, Shedding Light on Chevron: An Empirical Study of the Chevron Doctrine in the U.S. Courts of Appeals, supra note 27, at 7.

[314] We should be clear: "'Unpublished'," as Richard S. Arnold, "Unpublished Opinions: A Comment," 1 Journal of Appellate Practice and Process 219 (1999) writes, "does not mean 'secret'." "All opinions are public, in the sense that they are available to the public. Anyone may walk in off the street, pay the appropriate fee, and get a copy of any opinion or order of a court of appeals." What unpublished means rather is "that the opinion is not to be published in a book, a printed medium. It means that the opinion is not mailed (or otherwise transmitted) to West Publishing Company or any other legal publisher with the intention that it be printed in a book commercially available."

In 1999, 78.1% of all court of appeals decisions went unpublished. In only the First Circuit were more opinions published than unpublished (45.4%). Administrative Office of the U.S. Courts, "Judicial Business of the United States Courts: 1999 Annual Report of the Director," (1999), Table S. Data available at: uscourts.gov/judbus1999/supps.html

The movement toward unpublished opinions begin in 1964, when, the Judicial Conference of the United States recommended that the United States Courts of Appeals authorize publication of "only those opinions which are of general precedential value...." Judicial Conference of the United States, "Report of the Proceedings of the Judicial Conference of the United States," (1964), at 11. Eight years later, in 1972 the Conference endorsed a recommendation by the Federal Judicial Center and directed the circuits to devise plans to limit publication. By the mid-1970s all circuits had done so. To be sure, the specifics of the rules they adopted differ. But they are all are premised on the idea that judges should "generally seek to publish only cases of general precedential value." Kirt Shuldberg, "Digital Influence: Technology and Unpublished Opinions in the Federal Courts of Appeals," 85 Cal L Rev 541 (1997); see also Donald R. Songer, "Criteria for Publication of Opinions in the U.S. Courts of Appeals: Formal Rules versus Empical Reality," 73 Judicature 307 (1990); William L. Reynolds & William M. Richman, "An Evaluation of Limited Publication in U.S. Court of Appeals: The price of Reform," 48 U Chi L Rev 573 (1981).

[315] See, for example, Richard A. Posner, The Federal Courts (1985), at 122; Shuldberg, Digital Influence, supra note 314; Songer, Criteria for Publication of Opinions in the U.S. Courts of Appeals, supra note 314; Lauren K. Robel, "The Myth of the Disposable Opinion: Unpublished Opinions and Government Litigants in the United States Courts of Appeals," 87 Mich L Rev 940 (1989). See also *Anastasoff v. United States*, 223 F.3d 898, vacated as moot, 235 F.3d 1054 (8th Cir. 2000) (en banc).

[316] See, for example, Robel, The Myth of the Disposable Opinion, supra note 315 ; Songer, Criteria for Publication of Opinions in the U.S. Courts of Appeals, supra note 314; Songer, "Criteria for Publication of Opinions in the U.S. Courts of Appeals: Formal Rules versus Empical Reality," supra note 314; Donald R. Songer, "Nonpublication in the United States District Courts: Official Criteria versus Inferences from Appellate Review," 50 J Politics 206 (1988).

court judges have not continued to use traditional factors in adjudicating *Chevron* cases but have relied on *Chevron:* "Oddly, the best guide for predicting judicial outcomes under *Chevron* is probably the [*Chevron*] test itself."[317] Perhaps he is correct but, given the bias induced in his analysis from a selection rule that may be highly correlated with the dependent variable, his conclusions are unsupported and may be precisely the converse of the empirical reality.

E. In Large-n Studies, Draw a Random Probability Sample

If circumstances prevent researchers from collecting data on all members of the population but they have the resources to collect a large number of observations, they ought draw a random probability sample— a sample in which each element in the total population has a known (and preferably the same) probability of being selected.

Before we explain the advantages of this selection strategy, let us be clear about what it entails: Random probability sampling involves identifying the population of cases and selecting a subset according to known probabilistic rules. To do this, each member of the population must be assigned a selection probability, and selection into the observed sample must be done according to these probabilities.[318] Random selection thus is not haphazard selection or selection by convenience; it follows very specific rules and, in the vast majority of studies, will occur only if the researcher intentionally chooses to invoke it.

Several different forms of random probability sampling exist. In *equal probability sampling* all observations in the population have an equal chance of being included in the study. Consider the researcher who wants to understand, via an analysis of court decisions, why judges depart from the federal Sentencing Reform Act guidelines. Suppose she had a list of all 1,000 cases implicating the guidelines set out in the Act (with 1,000 representing a hypothetical figure) and wishes to draw a sample of 100 or 10%. Equal probability random sampling involves assigning every case the same probability of selection and selecting only 100. One way to do this would be to draw a set of 100 numbers from what is known as a "uniform distribution on the integers 1 to 1000". This process is equivalent to writing the numbers 1 to 1000 on poker chips, mixing them all up in a barrel, randomly choosing one, writing down the number, throwing the chip back in, and repeating the process until the investigator obtained a list of 100 numbers.[319]

A potential problem with this approach comes when the researcher has in mind a key causal variable. Suppose that we are interested in the causal effect of the political party of federal district court judges on the probability of departing from the sentencing guidelines, such that we posit that Republicans are more likely to depart (upwards) than Democrats. If resources permit us to select, say, 100 cases and we did so at random with equal probability of selection, we might by chance have a sample with no Republicans, thereby making causal inferences impossible. Even if we end up with many more Democrats than Republicans, our causal inference would be inefficient (have higher variance) than one with equal numbers of each.

To guard against the inefficiencies of chance occurrences, scholars often use the technique of *stratified random sampling*. The idea is to draw separate equal-probability-of-selection random samples

---

[317] Kerr, Shedding Light on Chevron: An Empirical Study of the Chevron Doctrine in the U.S. Courts of Appeals, supra note 27, at 7.

[318] Collecting all observations is of course a special case of random selection with a selection probability of 1.0 for every element of the population.

[319] An alternative would be to assign every one of the 1000 cases a 0.1 probability of selection, draw 1000 numbers from what is known as "a uniform distribution on the unit interval," each of which is a number between 0 and 1, and select each case if the number drawn is less than 0.1 and reject it otherwise. Drawing random numbers involves looking at a table of random uniform numbers or using a specially designed computer program.

within each category of another variable. In the present example, stratifying by the key explanatory variable would be especially useful since it guarantees a fixed number of observations (presumably an equal number) within categories of Republican and Democratic judges and, hence, can be used to maximize the efficiency of an estimator within the constraints of a fixed-sample size. All the researcher would need do is, first, stratify the cases according to the political party of the judge (that is, create two lists, one of cases decided by Republican judges, the other, by Democrats); second, assuming that the researcher wants a sample of 100, with an equal number of cases decided by Republicans and Democrats, she would draw an equal probability sample of 50 from each stratum.

These types of random probability samples noted, why do we advise using them if researchers have a large number of observations in the population? After all, throughout this Article we have counseled that "more data are better," and so why do we suggest ignoring any information we have about potential observations to be selected and selecting according to some random number generator guaranteed to be ignorant of all this auxiliary information? The main reason is that *random selection is the only selection mechanism in large-n studies that automatically guarantees the absence of selection bias.* That is because when we use random sampling we are, by definition, assuring the absence of any association that may exist between selection rules and the variables in our study. We already made this general point with regard to the Mills' study, but let us be more specific here. If Mills' had invoked a selection rule that led her to choose only state strategies that appear harmful to women and only clinical strategies that are favorable to women (*i.e.,* selecting observations that supported her theory), we would say that she used a selection rule that biased her sample in favor of her theory. And that she did so by selecting her observations on the basis of her dependent variable (*i.e.,* the strategy for dealing with battered women): Choose only those values of the dependent variable ("harmful" state strategies and "beneficial" clinical strategies) that support my thesis.

Under a selection rule that drew randomly from the population, Mills would not end up picking only observations favorable to her thesis, either intentionally or unintentionally. When appropriately applied, random selection prevents bias except by chance, and a large *n* means that the chance is exceptionally small. Of course, this does not mean that her rule and random selection would lead to different outcomes. Her selection rule might have not been biased, but since the exact data-generation process was not identified in her article, we do no know.

To put it more generally: Unless the researcher collects all observations in the population and that population itself was created in a unbiased manner by the world, random sampling is the only selection rule that safeguards against choosing observations in which the selection rule is related or correlated with the dependent variable or indeed *any* variable, except by chance. Indeed, no matter how carefully a selection rule is designed, when it is based on human knowledge it may inadvertently be related to the outcome variable being studied and so may introduce bias. So, while it is true that selecting observations randomly (in lieu of using whatever knowledge we have about the data to make the selection) violates the fundamental rule that "more data are better," we are willing to live with some information being discarded in order to avoid the inadvertent introduction of selection bias.

To our recommendation of using random selection strategies in large-*n* studies, we issue two caveats. One—that it applies to large *n* studies only—we take up in the next section. The other is simply this: Just because researchers use the verbiage of science or introduce some element of randomness in their studies does not mean that they have insulated their study from selection bias. We already have seen that, without additional precautions, a random sample from a set of *potentially* observable data will not yield unbiased inferences if the potentially observable data systematically differ from the population.

Unfortunately, the law review literature is replete with articles that demonstrate the same point. Exemplary is Miller's analysis of what particular factors (with those particular factors serving as the

independent variables in his study) cause attorneys to file a case in a federal or state court (with the choice of forum his dependent variable) when concurrent jurisdiction exists.[320] Given perceived caseload pressures in the federal courts, this is a serious matter—and one that has drawn the attention of the American Law Institute (ALI), Congress,[321] numerous scholars, and Chief Justice Rehnquist, who appointed a committee to study it.[322] Perhaps not so surprisingly these sources disagree over the continuing need for concurrent jurisdiction. The Chief Justice's committee, for example, concluded that a major rationale for its existence—a fear of local bias in state courts—is no longer valid; ALI expressly disagreed with this conclusion; and the scholarly verdict is mixed. Without doubt, then, Miller's attempt to sort through the various reasons could have elevated scholarly and public policy debates had he done what he claims to do—reach valid inferences about why attorneys file in one forum over the other based on a "random" sample of attorney responses to a survey.

What he does instead is the following: (1) Attempt to obtain a list of all state cases that could have been filed in federal court—an effort that failed because state courts apparently do not keep this information; (2) Obtain instead a list of all 1987 removal cases (n=18,860)—those in which the plaintiff chose to file in state court but the defendant had removed to federal court; (3) "Randomly select" (through some undisclosed means) 600 of these cases; (4) Send surveys to the 1,092 attorneys involved in those 600 cases; and (5) Receive and analyze responses from only 482 attorneys (a 44.1% usable response rate).

Why does this procedure, which does seem to rely on random sampling, fail? The problem is that Miller selects cases in a manner related to his dependent variable, the forum choice made by the attorney: He surveys plaintiff attorneys who want to be in state court and defense counsel who want to be in federal court. This means, first, that his study produces a biased descriptive inference. He cannot say much about the population of all attorneys confronted with a choice between filing in state and federal court from a sample focusing only on plaintiff attorneys who made one of two possible choices and defense counsel who made one of two possible choices. It also means that all causal inferences he makes are biased; we can even specify the direction of the bias. By only looking at defense attorneys who do not want to be in state court, Miller may, for example, overestimate the importance of factors such as local bias and by looking only at plaintiff attorneys who want to be in state court he may underestimate the effect. (To see this, think about a study that included defense counsel who did not ask to have their cases removed and thus, may not see a local bias, and plaintiff attorneys who filed in federal court and, thus, may perceive such a bias.) To be sure, Miller recognizes the problem: "The sample, of course, is biased since it excludes attorneys appearing in cases filed and not removed from state court."[323] But that did not stop him from reaching causal inferences and using those to make a series of policy recommendations.

Whether any legal or political organizations will adopt them, we do not know. What we do know, simply because Miller introduces an element of randomness into his study, does not prevent the resulting sample from being biased. And that until he corrects it or another researcher designs a study that does not select on the dependent variable, we should consider any inferences suspect.

F. In Small-n Studies, Avoid Selection Bias Without Random Selection

---

[320] Miller, "An Empirical Study of Forum Choices in Removal Cases under Diversity and Federal Question Jurisdiction," supra note 278.

[321] See Id. for a review of the activities of these groups.

[322] See Id., at note 46.

[323] Id., at 448.

As we note above, our recommendation to use random selection strategies applies to large *n* studies only. The reason a large *n* is useful is that it makes correlations by chance extremely unlikely, in fact less and less likely as *n* increases. But random selection does not help to avoid selection bias with a very small *n* since correlations between the selection rule and the dependent variable can occur by chance, even with fairly high probabilities.[324] This is a non-trivial issue because without random selection, there exists no single method to which researchers can automatically turn to avoid the problem of fooling themselves.

So the question becomes: How should researchers proceed if they are conducting small-*n* studies? The answer is that they must guard against biases they may inadvertently introduce into a study when they substitute random selection for some form of intentional choice. Achieving this requires them to *design a method of selection so that the selection rule is not related to the dependent variable.*

This is often quite difficult, of course, since the values of the dependent variable are typically unknown prior to any sampling. But, fortunately, various methods exist for overcoming it, that is, for selecting observations to include in small-*n* studies. Consider a scholar who wants to understand the degree to which police officers currently working in the State of Illinois comply with *Miranda v. Arizona* and hopes to do so by observing their behavior. Surely he could identify all police officers in the state and draw a random sample of those he will include in the study. But let us assume that he, like many of us, has constraints on his time and ability to travel such that he can only observe police officers in the station closest to his home. In other words, he has intentionally, and not randomly, selected his sample, thereby risking the possibility that the station he has chosen may be different than all others in the way it enforces *Miranda.*

How can he minimize this risk? Ideally, the researcher should collect more data; but, given constraints, one good approach is to identify a measurement strategy that is easy or inexpensive to apply across a wider range of observations, one that complements the detailed examination of the neighborhood police station. In terms of this example, perhaps the investigator could draw a random sample of, say, 100 police stations in Illinois and discern from public records the number of *Miranda*-based motions filed by defense attorneys involving officers in those stations. No doubt these data will be different and less valid indicators than those the researcher collects from his detailed case study;[325] but equally true is that they may help him to determine whether his station is representative of the others (*e.g.,* a ranking of 3 out of 100 on motions filed would reveal something very different than, say, a ranking of 50 out of 100). Another productive strategy might be to conduct the analysis from public records first and, on that basis, choose a more representative police station in which to conduct the detailed study. With this sort of information in hand the researcher could be more (or less, as the case might be) certain of the inferences he reaches from the smaller, intentionally drawn sample

---

[324] King, Keohane, and Verba, Designing Social Inquiry (1994), supra note 1, at 124-128 provide an example in which random selection with a small number of observations produces bias with a two-thirds probability.

[325] Although several studies have used this sort of measure (e.g., the number of confessions suppressed owing to *Miranda*) (see, e.g., Floyd Feeney, et al., Arrests without Conviction: How Often They Occur and Why (1983); Nardulli, "The Societal Costs of the Exclusionary Rule Revisited," supra note 220) to assess the impact of *Miranda,* legal scholars have pointed out problems with it; e.g., Paul G. Cassell, "Miranda's Social Costs: An Empirical Assessment," 90 Northwestern Law Review 387 (1996), at 393, 395: "Analysis of numbers of suppressed confessions tells us only about what happens to cases *when police obtain confessions*. It tells us nothing about cases in which police *fail* to obtain confessions because of the *Miranda* rules…[O]ne cannot simply tote up the number of 'lost confessions' by looking at a law enforcement bulletin or court docket."

IX. Concluding Suggestions: Developing an Infrastructure to Support Empirical Research

Rules matter. This is a claim that most scholars but especially legal academics accept without much debate. To the extent that conducting proper empirical research and, in particular, reaching valid inferences, depends on following rules, scholars have sufficient incentives to learn and to apply those we have outlined.

At the same time we recognize that—however much individual researchers want to carry out good research, however much they want to contribute credibly to policy debates, however much they want to speak authoritatively to their colleagues—following the rules may be difficult or, in some instances, nearly impossible without a sufficient research infrastructure. By "sufficient," we mean an infrastructure that supports, encourages, and enhances the ability of scholars to carry out empirical research and of lawyers, judges, and students to consume it.

To this end, we have designed, and now explicate, a series of recommendations centering on how law schools and the legal community writ large can facilitate the development of such infrastructure. None of them, we hasten to note, calls for building it from scratch. Quite the opposite. While few law professors conducting empirical research now seem to have much facility with the rules of inference, the methods and norms of empirical research, or the criteria with which to evaluate such work, the legal profession has developed high standards to govern many other aspects of their academic lives. Law schools, at least from our vantage point, appear highly organized, efficient, and well-funded, and most seem collegial and congenial. They put a remarkable amount of emphasis on satisfying their multiple constituencies, and they focus on matters such as curriculum and teaching far more so than most departments in the Arts and Sciences. The comparison to those in the social sciences is stark: political scientists, economists, sociologists and so on can go their entire careers without meeting collectively to discuss matters of pedagogy. Even the students in law schools are better organized than their counterparts in traditional academic departments.

What this means to us is that opportunities for quickly and significantly improving the research infrastructure in law schools are substantial. The norms and institutions already exist to do so, and our recommendations follow from them. In fact, owing to the strong norms, if law schools heed even some of our recommendations, not only might they be able correct the unfortunate state in which empirical legal research now finds itself. They also may be able to leapfrog other academic disciplines—even ones that have been doing superior empirical work but are nonetheless not as unified around a clearly identifiable community, with norms that support the enterprise.

What follows are recommendations that evidence seems to indicate can make this happen or, at least, start the process. We group them into five categories. The first set are geared toward three actors: (1) law school students, (2) law school faculty, and (3) judges and lawyers; the second set focus on two issues of interest to the entire community: (4) law reviews and (5) data archiving and documentation.

One final note: In order to convey our ideas as clearly as possible, we lay them out with a certain degree of specificity. But we surely do not regard them as the only way to proceed. Indeed, these recommendations are based only on our *hypothesis* that implementing them would improve empirical analyses in the law. We obviously are not certain that any of our ideas will work as intended at any particular law school, and we have conducted no analyses to evaluate them. Such studies surely should be done. At the same time, decades of experience, in dozens of academic disciplines, at hundreds of universities, and by thousands of scholars, give us some confidence in the general direction of these proposals. As outsiders, we necessarily have less confidence in whether we have

appropriately adapted these general principles to the culture of legal scholarship and law schools and, in any event, proposals of this sort should always be adapted further to the unique local conditions at individual schools. Whatever the ultimate fate of these ideas, we hope they help stimulate a vibrant discussion of ways to improve empirical research in the community of legal scholars.

A. Offer Courses in Empirical Research for Law School Students

Perhaps more than most units in the university, law schools are reputed to be responsive to the interests and needs of their students. In contrast to, say, academic Ph.D. programs, which rarely train their students to do what they will spend a large fraction of their careers doing—teaching—law schools devote resources and faculty to clinical instruction, moot court competitions, and other programs that help students develop the skills that they will need to practice law; in contrast to the academic disciplines, which typically allow only faculty to edit their journals, law schools entrust students—who presumably benefit in nearly uncountable ways from the experience[326]—to run some of their most prestigious publication outlets; and, in contrast to many of their academic graduate counterparts, which dedicate limited amounts of time and money to placing their students in jobs, law schools typically establish sophisticated apparatus to facilitate all aspects of the job search, from preparing resumes to interviewing strategies.

Why this norm has come about is not material. What is important, law schools can further enhance it by incorporating into their curriculum at least one course on empirical research—a course that would cover quantitative and qualitative approaches to research design and evaluation. It certainly should be required for students serving on the school's law review (a subject to which we return in Section D below) and probably for all others as well.[327]

We offer this recommendation not because all students will necessarily be conducting empirical research of their own. Law schools are not primarily designed for training future law professors, and so most students will never be preparing law review articles. It is rather because they will need the skills to evaluate such research, whether for a client, a senior member of their law firm, or a judge, whether in a criminal or civil suit. This is true today and it may become even more so as judges increasingly make demands on lawyers to meet particular legal standards, to question experts, or to back up specific claims with credible empirical support.

In addition to meeting the needs of students and the legal community, training students in the standards and norms of empirical research has at least two happy byproducts. First, again given the increasing demand for data, students with these skills will be more marketable than those without them (a trend we encourage in Section C below). This is good news for students who take and law schools that offer an empirical class; after all, high placement rates help attract better applicants—

---

[326] For some of those benefits, see Saunders, Student-Edited Law Reviews, supra note 134, at 1670; Wendy J. Gordon, "Counter-Manifesto: Student Edited Reviews and the Intellectual Properties of Scholarship," 61 U Chi L Rev 541 (1994); Leo P. Martinez, "Babies, Bathwater, and Law Reviews," 47 Stan L Rev 1139 (1995); John T. Noonan, "Law Reviews," 47 Stan L Rev 1117 (1995); Martin, The Law Review Citadel: Rodell Revisited, supra note 19, at 1099. Even critics of the current system acknowledge its value to students. See, for example, Roger C. Cramton, "The Most Remarkable Institution: The American Law Review," 36 J Legal Educ 1 (1986).

[327] Whether law schools should make this a part of the first-year curriculum or reserve it for second- and third-year students depends on their individual needs and goals. One relevant comparison is to graduate programs in the social sciences, which typically encourage students to take tool-oriented courses (*e.g.,* research design, methods, foreign languages, game theory) as early as possible in their academics careers. Also worth noting, the strategy of trying to cover the rules of inference in existing substantive courses is useful, but not sufficient: the study of inference, research design, and empirical methods constitutes a distinct and clearly-delineated field of scholarly inquiry, and is best taught in a separate course.

both of which can lead to increases in the school's status. Second, faculty will benefit enormously. Offering empirical courses will require law schools to hire a scholar trained in empirical methodology who, in turn, could serve as a resource for faculty—and one that they may have been unable to obtain but for curriculum needs.

We have more to say about the infrastructure requirements of faculty below but let us first address two obvious concerns: Who might this methodologist be and from what academic fields should he or she come? Starting with the first, surely she or he should be a dedicated scholar and teacher well versed in the rules of inference and the norms and standards for conducting empirical research but more than that is necessary. The selected methodologist also should be able to teach students and faculty how to analyze *their* data and, thus, possess technical skill sets.

This methodologist could, on the one hand, hail from any number of academic disciplines. Because empirical research in law has methodological problems that overlap with those in Biology, Chemistry, Economics, Medicine and Public Heath, Political Science, Psychology, and Sociology, methods can be adopted from those other disciplines to the study of the law. On the other hand, in virtually every discipline that has begun to develop a serious empirical research program, scholars discover methodological problems that are unique to the special concerns in that area. Each new data source, as it turns out, often requires at least some adaptation of existing methods, and sometimes the development of new methods altogether. There is bioinformatics within Biology, biostatistics and epidemiology within Medicine and Public Health, econometrics within Economics, chemometrics within Chemistry, political methodology within Political Science, psychometrics within Psychology, sociological methodology within Sociology, and so on.

Thus, to encourage serious, enduring, and continually improving empirical research, the legal community ought foster the development of a subfield of methodology within law. To accomplish this, law schools, to be sure, should hire scholars who have deep training in empirical methods in whatever discipline they obtained their degree. But they also should select a methodologist who has, or at least has an interest in developing, an understanding of the kinds of problems that interest, and the sorts of data available to, legal scholars. Certain academic disciplines regularly turn out Ph.D.s who fit this description (*e.g.,* Economics and Political Science). And law professors can help themselves out by inculcating in these methodologists an even greater appreciation of their concerns. This could come about through co-teaching courses, which would work to the benefit of students and faculty; it also might evolve via collaborative research—a subject we discuss in more detail in the next section. Either way, new scholarly links would be created. Some new methods for law would then be developed by, say, the political scientist or economist, which is fine given the increasingly interdisciplinary nature of law, but others would be developed locally. And eventually, law schools would not need to contract out methodological concerns; the field—the empirical methodology of legal scholarship—would flourish on its own.

B. Enhance Opportunities for Faculty to Conduct High-Quality Empirical Research—and to Disseminate it Quickly

When it comes to their research, law professors seem to have developed a norm of timeliness. Perhaps they are interested in weighing into current policy debates, seeing their arguments worked into legal briefs, ensuring that a court decision or an act of Congress does not render their ideas moot, obtaining tenure, or attaining some other goal but, whatever the explanation, they are concerned—perhaps more concerned than most other academics—with getting their ideas and results out as quickly as possible.

We take no issue with this norm. In fact, scientifically valid input into current debates about public policy can make highly important and dramatically influential contributions. Legal scholars

have proven time and time again that they are uniquely situated to take on this task at least with regard to speed; where they have failed, time and time again, is with the high quality aspect of the task. This is unfortunate *because they can do both*. That is, they can conduct first-rate research that they can create and disseminate rapidly. For, even if time, information, and resources are limited, there are ways to produce credible results.

To help faculty accomplish this, we offer two sets of recommendations for law schools to follow. The first centers on fostering the development of the skill sets necessary for their professors to do high quality research so that they can respond correctly in the time allowed; the second is aimed at building an infrastructure to allow law professors to produce credible research results as quickly as possible. These suggestions will enable scholars to produce valid scientific inferences given their limited time and resources (*i.e.,* correctly judging their uncertainty), but we do not stop there. Although time will always be limited, we suggest ways that resources can be redirected and marshaled so that the degree of uncertainty in scholarly conclusions can be greatly reduced, even for research that needs to or ought be disseminated quickly.

1. Help Build Methodological Skills

Learning and understanding the rules we discuss in this Article are, we believe, necessary steps for law professors to take. Yet, at the same time, we acknowledge that these steps alone are insufficient, and that legal academics will require additional training to implement the rules we have offered and to master skills associated with the analysis of data—whether of the qualitative or quantitative variety. How can they develop them?

Individual faculty can proceed in any number of ways, with three rather obvious. First, they can take an empirical research course. We say this fully appreciating that law professors typically do not audit their colleagues' classes. But they should know that in many cognate disciplines, scholars—whether tenured or untenured, whether beginners or senior scholars, whether to brush up on their skills or learn entirely new ones—regularly take technical courses. And they do so without shame; actually, in many academic departments attendance in a methods class can lay claim to bragging rights as it indicates both to colleagues and graduate students a desire to stay au courant.

Second, faculty members can obtain training at institutes, with the Inter-university Consortium for Political and Social Research (ICPSR) among the most prominent. Located at the University of Michigan, the ICPSR provides a summer training program for faculty and students in interested in empirical analysis. Courses range from the introductory (*e.g.,* Introduction to Statistics and Data Analysis) to the technical (*e.g.,* Advanced Topics in Maximum Likelihood Estimation), from the general (*e.g.,* Introduction to Computing) to the very specific (*e.g.,* Nonrandom Selection in Aging and Retirement Studies), from the highly theoretical (*e.g.* Mathematical Models: Game Theory) to the unabashedly empirical (*e.g.* Quantitative Methods and African Studies).[328] And though only a few are geared specifically to law-related issues (*e.g.,* Criminal Justice Data: Integrating Qualitative and Quantitative Studies), legal academics would walk away from the experience considerably more knowledgeable about, at the very least, the vast array of tools available for developing theories and for analyzing qualitative and quantitative data.

Third, law professors can learn on the job by entering into collaborations with a methodologist in the law school or colleagues with an interest in law in, say, one of the social sciences. This is, perhaps, the easiest and most efficient way—and one used quite often in other academic disciplines—for legal academics to develop an appreciation of empirical methods and to learn the skills necessary to carry out such inquiries on their own.

[328] Information about the ICPSR, including its Summer program is available at http://www.icpsr.umich.edu.

Law schools can and should facilitate each of these activities. For faculty members who would like to take an empirical research course and, ultimately, can demonstrate a mastery of the skills (perhaps in the form of a research presentation to the faculty or a published article), their schools could provide some release-time from teaching. For those who would like to attend the ICPSR, they should pay their tuition, as many graduate programs in other fields currently do. And for legal academics interested in entering into collaborations with empirically skilled colleagues, their school should provide incentives to turn interest into action—perhaps in the form of seed grants for the proposed project or other forms of support. However they proceed, law schools must acknowledge that the point of an academic research career is to make the maximum contribution to a scholarly literature. Whether that contribution is single or co-authored should not be matter so long as the contribution is there. Even more to the point, if including coauthorship in one's repertoire can help a researcher generate a larger total contribution—as is the case for scholars in many other fields—then it should be strongly encouraged.

At the very least, law schools should not punish faculty (*e.g.,* denying tenure) for coauthoring articles. This is the source of some concern since collaboration is so rare in legal journals (and so common elsewhere) but it is easy enough to allay. Law schools, as Schuck counsels, must "devise techniques for properly evaluating each individual's contribution to jointly conducted research."[329] The most common method in other disciplines takes the form of simple advice to junior faculty: "Collaborate with more than one scholar; don't collaborate only with senior scholars." Following this advice makes individual contributions easier to evaluate across projects but other sources of information also exist, such as directing outside letters of evaluation at the time of promotion to those familiar with the scholar and his or her research practices and contributions to coauthored projects.

But law schools can go even further, especially if there is a critical mass of faculty interested in producing high-quality empirical research. Perhaps they could hold a monthly seminar on applications and innovations in empirical methodology, inviting prominent scholars (for now, from other disciplines) to lead them. They also ought consider setting up their own on-site summer institute, which might involve hiring experts in empirical research related to law. These experts would serve as teachers and discussion leaders and, with any luck, create opportunities for more long-lasting collaborations. Finally, law schools should explore the possibility of obtaining foundation support to develop empirical research workshops lasting three or four days at a time, to bring perhaps 20 or 30 faculty up to speed quickly. Obviously short sessions are not sufficient, but they can help considerably. Setting them up in combination with creative web-based distance-learning materials, along with assigned reading in preparation for the workshops (and, of course, attending several) would go a long way. And, eventually, if an empirical course were to become a part of the law school curriculum, the workshops would not be needed, as students (*i.e.,* future law professors) would obtain sufficient training.

2. Save Time by Improving Resources

Scholars can conduct serious empirical research no matter how limited the time or resources. But, if both time and resources are highly constrained, they will pay a price in the form of less certain findings (*i.e.,* the less time and resources, the smaller the number of observations that can be collected or the less reliable the measurement procedures that can be used, and hence the greater the inefficiency of the resulting inferences). Since at least some law professors desire to produce

[329] Schuck, Why Don't Law Professors Do More Empirical Research?, supra note 2, at 333.

research results that they can disseminate as quickly as possible while also ensuring that they are as informative as possible, then taking more time is not in the cards. Increasing resources, however, could make a non-trivial difference. We see at least four ways law schools can help.

First, they can ensure that faculty conducting empirical research have computers and software up to the task, along with the technical support they need to use those resources. As a rough calculation, computers should be replaced every three years, and software upgraded approximately every year.[330] Staff support could take many different forms but normally includes network administrators, systems operators, user support personnel, and clerical assistance. Of course, many law schools already have some of these persons in place (in addition to web designers and administrators), and will likely add more as their faculty members increase their dependence on technology for communication and writing. Our suggestion, in addition to maintaining excellence in this area, is to supplement the existing information technology group with experts who can perform specific research tasks. We note two examples here, although finding individuals to cover both might be possible.

One is a specialist in statistical software programs. Many law professors now rely on simple database or spreadsheet programs such as Microsoft Excel. These are fine for exploring data in their raw form, but they are not useful for serious statistical analyses and graphics, and are not even numerically stable for many statistical purposes.[331] The infrastructure should be available so that researchers are not constrained in what analyses they perform by limitations in their software. Fortunately, numerous alternatives (including SPSS, SAS, Stata, Gauss, R) exist. These are more powerful, to be sure, but they are also more difficult to learn and use. Thus, the need for expertise, which can take the form of consultants to faculty who will do the programming themselves, all the way up to assistants who will do the work for the faculty. The other is a specialist in graphic design. Presenting data, again whether qualitative or quantitative is an important skill—and one that many empirical researchers often require but do not have the time or inclination to learn. Having a person on staff with the necessary expertise would be an enormous boon to law professors who, perhaps more often than most, must communicate their research results to laypersons.

Who we do not include on this list is a person skilled in data-analytic techniques. While we have explained that retaining such an individual is critical, he or she or should not be considered part of the information technology group. A methodologist is an academic—in the field of law, a law professor—who focuses on, contributes to the field of, and applies quantitative and qualitative legal methodology. Since statistics and research design are not "merely technical," as is, say, like plumbing, "staff statistician" positions generally do not work in this context. Law schools need *creativity* in methods, not a technician who merely applies existing techniques by rote to legal scholarship—a path that generally leads to the use of methods that do not comport with the needs of researchers. Just as in any other field, methodology is a creative endeavor and cannot be delegated to anyone other than another scholar.

A second way law schools can help their faculty optimize their time is by providing additional person power, in the form of research assistants (RAs), who will enable scholars to collect data as quickly and efficiently as possible. Academic departments accomplish this in various ways—including fellowships, stipends, and course credit for students providing research assistance—all of which would be feasible for, and, to some extent, already exist in many law schools. Merely adopting or extending the model used in academic departments, however, will not work. The problem is that

---

[330] Gerald V. Post, "How Often Should a Firm Buy a PC?," 42 Communications of the ACM 17 (1999).

[331] Micah Altman and Michael McDonald, The Robustness of Statistical Abstractions: A Look Under the Hood," paper presented at the Annual Society for Political Methodology meetings, College Station, TX, 1999 (on file with the authors).

researchers in other fields typically must make commitments to RAs at the start of the semester or academic year, and so finding help at the last minute can be very difficult, if not impossible. Clearly, RAs even in this traditional form would be an improvement, but they will not solve the problem of producing the best results as quickly as possible to answer questions that arise in public debate.

Accordingly, in addition to law professors hiring RAs as many already do, a solution we favor would be for the law schools themselves to employ a number of RAs (along with a portion of the resource people we mention above) and keep them unassigned at the start of the semester. They would then be available to deploy, on request, when a particular policy debate or other matter emerges that requires immediate attention. Undoubtedly such a resource pool would help law professors to enhance the already-existing norm of speed.

Third, and this we already have mentioned, law schools should encourage their faculty to enter into collaborations with scholars who know how to conduct serious empirical research. In addition to the reasons we offered earlier, collaborative empirical work is faster to conduct. Legal academics need not waste precious time learning the details of every possible new skill and instead can rely on coauthors, who presumably would benefit from the substantive expertise law professors bring to the table.

Finally, to conduct empirical research, scholars often require funding: They may need to acquire a particular data set, field a survey, hire interviewers, and so on. So that their faculty can optimize their time, we recommend that law schools, and their associated centers, follow the lead of many other academic units and supply seed money to credible projects. Such funding would enable scholars to conduct pilot studies that they could, in turn, use to inform public policy debates or to demonstrate the worthiness of their research to various outside funding agencies, foundations, and donors. External funding certainly has benefits for individual research projects, but it also has positive implications for law schools. Surely deans would not turn down reimbursements for indirect costs that would flow into their coffers if more of their faculty obtained support from, say, the National Science Foundation's Law and Social Science Program. As a further incentive, law schools might follow the path of research units that pass back some fraction of indirect cost reimbursement to the faculty generating the funds in the first place.

C. Encourage Employers to Hire Students with Empirical Training

At the onset of this Article, we highlighted the calls issued by members of the legal community, especially judges and attorneys, for credible empirical research on a range of topics. We have tried to respond to those pleas by offering rules that scholars could use to improve their research and that consumers could employ to evaluate various studies.

But those consumers, again primarily judges and lawyers, can further their own cause by making special efforts to hire law students with empirical training. This is so for at least two reasons. First, it would encourage law schools to develop the necessary infrastructure, that is, to follow our other recommendations. If, for example, judges began querying applicants for clerkships about their empirical training, this would provide incentive for law faculty to add the necessary courses to their curriculum. Which, in turn, would lead schools to hire a methodologist, facilitating the ability of faculty to produce the properly conducted studies that judges are now demanding.

The second reason is akin to our recommendation that legal academics find collaborators with empirical skills. Just as coauthoring with a methodologist enables faculty to optimize their time, hiring a clerk or associate with empirical skills is an efficient way for judges and lawyers to appraise research without investing resources other than those that they already must (*e.g.,* salary). Judges confronted with conflicting results produced by empirical studies in redistricting, employment, or innumerable other areas of law will have in-house talent to help them sort through the findings.

Ditto for lawyers who must challenge results that do not sit comfortably with theirs. Hired statistical guns, while perhaps not moving into complete obliteration, will become less necessary.

By offering this recommendation, we do not mean to imply that empirical training ought be the only or even most important criterion that lawyers and judges should consider when making hiring decisions. What we are instead suggesting is that if it were to become even one of many, the ripple effect would be extremely beneficial to all members of the legal community—the producers and consumers of research.

D. Move to an Alternative Model of Scholarly Journal Management

In the law world, students run and edit their school's flagship journal (*e.g.,* the *University of Chicago Law Review, Yale Law Journal*), though they often consult informally with faculty before making decisions over particular articles.[332] We have read various accounts of how this norm came about,[333] and we appreciate the tradition. At the same time, certain aspects of it are bothersome, such as its failure to conform to a critical aspect of empirical research, that it not be ad hominum, that the focus be on the work and not the person.[334] Without some form of blind reviewing, separating the person from the product will be difficult. Also problematic is that students (and indeed any one person) may lack the expertise necessary to evaluate the submissions that cover complex and technical areas of the law or employ sophisticated statistical or qualitative methods. Finally, the lack of blind peer review in most of law journals puts legal academics at a distinct disadvantage vis-à-vis the rest of the university. Gary Wills is not the only prominent scholar to express shock upon learning how flagship law reviews operate.[335] As Lawrence Friedman puts it,

> Law reviews are the primary outlet for legal scholars, and the law review system is unique to legal education. People in other fields are astonished when they learn about it; they can hardly believe their ears. What, students decide which articles are worthy to be published? No peer review?…
>
> Secretly, I share their astonishment; and I think the system is every bit as crazy, in some ways, as they think it is. There is, in fact, quite a literature of invective—professors and others railing against the law reviews.[336]

What this quote, and many others we could supply,[337] suggests, the lack of peer review (among other features of this "unique" system) makes it difficult for scholars in other units to take legal

---

[332] Saunders, Student-Edited Law Reviews: Reflections and Responses of an Inmate, supra note 134, at 1683; John G. Kester, "Faculty Participation in the Student-Edited Law Review," 36 J Legal Educ 14 (1986).

[333] E.g., Michael I. Swygert & Jon W. Bruce, "The Historical Origins, Founding, and Early Development of Student-Edited Law Reviews," 36 Hastings L J 739 (1985); John J. McKelvey, "The Law School Review," 50 Harv L Rev 868 (1937); Hibbitts, "Last Writes? Reassessing the Law Review in the Age of Cyberspace," ; William P. LaPiana, Logic and Experience: The Origin of Modern American Legal Education (1994); James W. Harper, "Why Student-Run Law Reviews?," 82 Minn L Rev 1261 (1998).

[334] This concern implicates the way students select articles—and it is one that many scholars have raised. See supra note 134.

[335] See Wills, "To Keep and Bear Arms," supra note 20.

[336] Friedman, Law Reviews and Legal Scholarship, supra note 25, at 661.

[337] See, for example, Arthur D. Austin, "The 'Custom of Vetting' as a Substitute for Peer Review," 32 Ariz L Rev 3 (1989) (claiming that "The use of student edited journals as the main outlet for legal writing is an embarrassing situation deserving the smirks of disdain it gets from colleagues in the sciences and humanities"); Lindgren, An Author's Manifesto, supra note 44, at 535 (noting that "In some other parts of the academy, legal journals are a joke. Scholars elsewhere frequently can't believe that for almos all out major academic journals, we let students without advanced degrees select manuscripts"); Rosenberg, Across the Great Divide (Between Law & Political Science), supra note 21, at 270 (in discussing factors that contribute to the intellectual isolation of legal academics, notes "There is…the question of the way in which the work of legal academics is published.").

work seriously—especially since their colleagues do not "count" non-peer reviewed articles when it comes time for tenure, promotions, salary raises, and other perks. That others do not take legal research seriously is, of course, not our point; what is central is that peer review has, at times, important benefits.

As far as we can tell, though, switching wholesale to the full blind peer review model used in academic journals throughout the natural and social sciences is politically infeasible. The large number of journals would create an enormous increase in the workload for law professors serving as anonymous reviewers. Moreover, the work that goes into reviewing is normally accompanied by a prohibition against multiple submissions (so that the editors' and their reviewers' efforts would not be wasted); accordingly, a switch to peer review could also slow publication—an especially undesirable outcome given the norm of speed. Other difficulties have also been identified,[338] but suffice it is to say here that the full-blown version of the traditional blind peer review model does not seem to fit with the norms, needs, and goals of the legal community.

It is for these reasons that law professors have retained their "unique" model but, at the same time, have developed various mechanisms to compensate for perceived deficits in it. Extensive footnoting is one.[339] Another is the ever-present and long list of scholars thanked at the beginning of articles, which may have emerged as a means to testify to the credibility of the research.[340] Of course, name-dropping does not mean that the person named served as even a non-anonymous peer reviewer or in any way approves of the article; nor do elaborate footnotes guarantee the value of the research. But, apparently, law school faculty viewed these as handy responses to the shock expressed by universities when "the family skeleton was exposed. LAW PROFESSORS ARE EDITED BY LAW STUDENTS."[341]

However useful these stylistic additions may have been at the time law professors developed them, they do not compensate for the problems we and others have identified in the lack of blind peer review. We thus propose an alternative model—one that enables law schools to continue the existing norm while enhancing it by taking advantage of some features of the peer review system. Other possibilities exist of course,[342] but this one, which is similar to that many university (book) presses follow, may best fit with the traditions in law. That model would work as follows.

---

[338] Many legal scholars have raised objections to the application of variants of a blind peer review model to their work. See, for example, Hibbitts, Yesterday Once More, supra note 134, at 292; Friedman, Law Reviews and Legal Scholarship, supra note 25, at 665; Posner, The Future of the Student-Edited Law Review, supra note 134, at n.8; Saunders, Student-Edited Law Reviews: Reflections and Responses of an Inmate, supra note 134, at 1677.

[339] Austin, The 'Custom of Vetting' as a Substitute for Peer Review, supra note 337, at 3 argues as much, but the use of elaborate footnotes, as he notes in Footnotes as Product Differentiation, 40 Vand L Rev 1131 (1987), reflects other (albeit related) factors.

[340] See Austin, The 'Custom of Vetting' as a Substitute for Peer Review, supra note 337, at 5. Such "testimony" may be geared as much to law review editors as it is to scholars in other parts of the university. Both Austin and Hibbitts suggest as much, with Hibbitts, Last Writes?, supra note 333, noting that "the number of prominent names the author can drop in an 'acknowledgements' footnote" may affect editors' decisions." See also Austin, "Footnotes as Product Differentiation," supra note 339.

[341] Austin, The 'Custom of Vetting' as a Substitute for Peer Review, supra note 337, at 3.

[342] Indeed, we are well of aware of the numerous proposals scholars and others have offered (some of which law schools have adopted). Many call for various degrees of "tinkering" with the existing model. See, for example, Posner, The Future of the Student-Edited Law Review, supra note 134, at 1136, suggesting that the law reviews, among other things, obtain blind reviews on "every plausible" nondoctrinal manuscripts and "forswear line-by-line editing;" Saunders, Student-Edited Law Reviews: Reflections and Responses of an Inmate, supra note 134, at 1682, making several suggestions, including that law reviews diversify their memberships and "encourage informal faculty involvement." Others call for major overhauls, such as moving to a faculty symposium model, see Lindgren, An Author's Manifesto, supra note 44, at 1127; faculty-edited journals, see Richard A. Epstein, Faculty-Edited Law Journals, 70 Chicago-Kent Law Review 87 (1994) and David M. Richardson, Improving the Law Review Model: A Case in Point, 44 J Legal Educ 6

•Students would continue to serve as they do now, as law review editors and members. But law schools would expand editorial boards to include faculty members.

•As they receive manuscripts, students—like university press editors—can reject manuscripts for whatever reasons they think valid, just as they do now. But for any manuscript that they deem potentially publishable, they must obtain at least one blind peer review (that is, the reviewer does not know the identity of the author and the author does not know the identity of the reviewer). The reviewer should be an expert in at least some aspect of the subject or methods in question. In most situations, this means a law professor (ideally but not always from another law school), although occasionally it may mean a student who has written a thesis on a related subject or, possibly, a Ph.D. in another area. The key is that editors turn to a reviewer based on the reviewer's expertise, not status.

•After receiving the external evaluation, students would be free to reject the manuscript. But, if they would like to publish the essay, they must bring the anonymous peer review and any internal student evaluations to the editorial board for final approval. If desired, the student law review editor may assign someone to write a response to the anonymous review, solicit a response, or even a revised version of the article from the author, and include this material in the information that the editorial board reviews.

Whatever the exact procedure, the important point is that the law review would publish only articles that have (1) been reviewed by at least one external expert in a double blind (or at least single blind) peer-review setting and (2) attained the approval of the editorial board. The editorial board would serve as a check on the student law review editor, but in the vast majority of cases expectations would be clear enough so that the editorial board would support the law review editor's decision. Indeed, what happens at most university presses, and what would in all likelihood happen if this model were appropriately adapted to law, is that the editorial board operates to empower the editor, a position that would be as autonomous as it is now. The new system would make it easier for the editor to say "no" to senior faculty members who may hold some influence on their future careers ("I'm sorry, the editorial board did not approve your article…."), and it would add substantial credibility to the decision-making process, to the prestige to the law review, and to the scholarly value of its content. Student editors are already aware of some of these advantages, as their practice of consulting faculty informally attests.[343]

In offering this model, we recognize that following it—or some variation that schools adapt locally— may add to the burden of students and faculty. Students, with faculty guidance, must begin to develop a pool of external referees; deans will need to persuade faculty to sit on editorial boards; and faculty will, on occasion, be asked to serve as manuscript evaluators. Moreover, law professors, accustomed to relatively rapid turnaround time, will (perhaps) have to wait slightly longer for decisions on manuscripts sent out for review.

To us, none of these costs are terribly onerous or problematic. Reviewer lists, as faculty editors have suggested, are easy enough to develop; and this may be even more so in law given the annual

---

(1994); and even self publishing on the World Wide Web Hibbitts, Last Writes?, supra note 333. Our proposal falls between to these two extremes and, as such, has the virtue of preserving some of the best features of the existing model while offering a corrective to its more serious deficits.

[343] Providing student editors with a voice in the selection of editorial board members can further enhance their autonomy and guarantee a smooth working relationship between the board and the editors.

AALS Directory of Law Teachers, a type of resource that many other disciplines do not possess. Moreover, once the law review establishes its initial list, the next set of editors need only build on it. (As an aside, the process of compiling such information alone can be quite informative, helping students and faculty to learn about entire fields of research.) Faculty serving on editorial boards will rotate over time, thereby ensuring that no single professor is saddled with the task for long periods. If necessary, law schools can compensate faculty board members with rewards ranging from release time to seed money for their next project. Finally, our experience in Political Science, a discipline (like most) in which journals rely on peer review, is that competent editors can turn around manuscripts in two or three months. And that is with obtaining at least *three* external reviews for *all submitted articles,* not simply for those they deem publishable. They manage to accomplish this by, among other strategies, taking advantage of the speed of email to contact potential reviewers, sending out manuscripts in electronic form, and enforcing strict deadlines on the submission of referee reports. These and others are all feasible for law reviews.

Even more to the point, the benefits of this alternative model outweigh any of these inconveniences. Both students and faculty accrue an advantage of which scholars in other disciplines are only too well aware: Reading and making assessments over manuscripts, while a chore, is a great way to learn about the state of the literature, and to do so even before publication. This is one of the reasons why scholars in other disciplines are willing to take on the burdens associated with reviewing and serving on editorial boards. Moreover, our alternative model provides a mechanism—the editorial board—to facilitate faculty-student interaction, to break down the hierarchy that seems more severe and entrenched in law schools than in many other academic programs. To us, the only relevant hierarchy in an academic discipline is based on knowledge, and sometimes students have this knowledge and faculty do not. Indeed, while the opinions of outside experts can help discern whether the article in question is right, sometimes the person who knows more than most anyone else about a subject is a student who has researched it; sometimes the person with the best idea about a research topic is someone who has not been "biased" by years of operating within the standard paradigm, and this too sometimes may be a student. Having faculty and students make joint decisions thus has enormous benefits for all involved. It invites students to become to be part of the academic community, to be socialized into a world where learning never stops, where expertise is shared, where the norms of the free exchange of academic information are inculcated, and where new ideas are developed.

The benefits associated with this model for law schools—both for their standing in their universities and in the legal community—are clear: No longer will deans and their faculty be embarrassed by the "family skeleton;" they will be able to say credibly that all published law reviews have followed at least *a* model of peer review. This will improve their status in other parts of the university, which—regardless of the ranking of the law school—often look at them as intellectually impoverished. Externally, it may help as well. The review process will filter out at least some of the "junk" law professors themselves accuse their own journals of publishing,[344] with the result being better publication outlets. We also can imagine law schools using the model as a vehicle for improving their reputations. Suppose that, of the top twenty-ranked law schools, ten adopt the model and ten do not. This would provide an opening for the ten adopters to advance their positions, for once it becomes common knowledge that their flagship journal is peer reviewed, deans elsewhere will have incentives to push their faculty to publish in them. Over time, the peer review journals will become better and better while others will be discounted, just as they are in other academic disciplines. The same logic, of course, holds for all other law schools, regardless of their current place in the pecking order.

[344] See supra notes 41 to 52.

Surely, in commending this model, we recognize that it does not overcome many of the liabilities scholars associate with the existing law review selection process. But it does have substantial advantages over the present model, while retaining those elements that legal scholars seem to find attractive.

E. Develop Standards for Data Archiving

One of the strongest norms in legal publishing is the norm of textual documentation: Law review editors and authors are, to a greater extent than most others in academia, obsessed with footnotes. We realize that this norm has come under attack from many quarters[345] and, from some perspectives, is a waste of effort, but from the perspective of empirical research, it has two important advantages. First, it connects the extant scholarship to existing literatures. This is one of the few ways that legal academics fulfill our admonitions about the importance of developing a community of scholars. Second, elaborate footnotes enable readers to locate any text cited in an article and learn about the content of that text. And if the text is unpublished, scholars are able to obtain it from the author or the law reviews themselves, which ask authors to provide unpublished materials for storage in their archives. We certainly cannot say the same of the norms in almost any other academic discipline.

Given the importance (and value) of this norm of documentation to the legal community, it is surprising that violations are rampant when it comes to non-textual sources of information—most relevant here, quantitative or qualitative data analyzed in empirical research. So, for example, while the law reviews regularly obtain unpublished material from authors, they do not typically store qualitative or quantitative data or documentation necessary to replicate the studies that they publish.[346] Along the same lines (and with few notable exceptions), we found it impossible to obtain data used in law review articles from public sources, from the law reviews, or from the authors of the articles directly. Even those "notable exceptions" came with strings attached or other complications. In one instance, the author was willing to provide his data but only if we were willing to sign a legal document placing near-Draconian limits on our use; in another, the data came in a form that made analysis near impossible.[347]

---

[345] See, for example, Abner Mikva, "Goodbye to Footnotes," 56 U Colo L Rev 647 (1985); Arthur J. Goldberg, "The Rise and Fall (We Hope) of Footnotes," 69 American Bar Association Journal 255 (1983); Rodell, Goodbye to Law Reviews, supra note 45 at 41; David Mellinkoff, Legal Writing: Sense and Nonsense (1982), at 92; Cramton, "The Most Remarkable Institution: The American Law Review, supra note 326, at 5; Lasson, Scholarship Amok: Excesses in the Pursuit of Truth and Tenure, supra note 42, at 939.

[346] A related (and troubling) development is that neither the editors of many law reviews nor the producers of the studies they publish have apparently taken steps to force various companies (*e.g.,* LEXIS-NEXIS) that produce electronic versions of their journals to include tables and figures. Where these elements are supposed to appear in the text, the reader is directed to see the print version. Surely editors and faculty would (and should) raise the roof if LEXIS-NEXIS treated footnotes in the same way. Fortunately, at least the tables are available somewhere; such is not the case with most original quantitative or qualitative data.

[347] What this experience shores up is that authors—whether law professors or natural, physical or social scientists—are often of little value in helping others to replicate research that did not originally meet the replication standard. In fact, many efforts to replicate work that does not meet the standard fail even with the author's help. See William G. Dewald, Jerry G. Thursby, &Richard G. Anderson, "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," 76 Am Econ Rev 587 (1986). This is not surprising since scholars are not professional archivists and frequently do not retain all the information necessary for replication. Of course, even if authors were routinely helpful, never lost anything, and lived forever, relying on them rather than on published material on the public record means that scholarship, interpretation, and all the facts are not shared. As such, the benefits of a scientific community are lost.

The upshot of these practices is that the very basis of the most important documentary evidence in empirical law review articles is forever lost. This monumental waste of the resources should not stand. How can the scholarly community evaluate such work? How can future scholars build on it? For that matter, how can even the original author conduct follow-up research? How can the scholarly community correct mistakes, improve its methods, or benefit from the most important advantages of having a scholarly community in the first place?

We recommend that law reviews, at a minimum, require documentation of empirical data with as much specificity as they do for textual documentation. And, just as for textual documentation, this should be a prerequisite for publication. This means simply making it possible for any reader to traverse the chain of empirical evidence amassed to support the conclusions published. Citing public-use data sets is one way to comply with this rule, but in virtually all situations the only way to ensure full compliance is to require researchers to deposit their original data, and all information necessary to replicate their results, in some *public archive*. This may sound like an unusual idea, but scholars in every field who have tried to replicate another's empirical work know how hard that task is to accomplish without the original data. Even those using public data sets would normally need to deposit at least the full details of their calculations—how they moved from the publicly available data to their numerical results—and exact information about which version of the public data set they analyzed. By the same token, researchers conducting surveys should deposit the individual-level responses to their questions (removing only information necessary to protect the identity of the respondents), and any calculations performed (*e.g.,* how missing data were handled). Investigators coding cases would deposit their data sets, complete coding rules, and the precise connections between their numerical data and the original cases from which they were coded. Scholars studying speeches of legislators would deposit the texts of the speeches if they were not easily retrieved from other sources, or detailed citations to all speeches consulted.

Many public archives exist, including Publication Related Archive of the ICPSR,[348] the International Studies Association data archive,[349] Qualidata: The ESRC Qualitative Data Archival Resource Centre,[350] the Database Registry of the Economic History Association,[351] and Statlib, a repository of the statistics community.[352] A healthy procedure would be for law reviews individually or collectively to establish their own data archives, so that they can keep empirical evidence and satisfy the norms of the legal profession. Examples of archives associated with individual journals are those created by the *Journal of Applied Econometrics*[353] or *Political Analysis*.[354] The Virtual Data Center project provides easy public domain tools that journals and others can use to set up their own archives, as well as the exact standards for citing empirical data.[355]

This recommendation centers on the law reviews; another pertains to legal scholars themselves: Those who comply with this rule ought receive credit for it. Legal academics should list the data sets they have made publicly available on their vitae, just as they now list published articles. And hiring and tenure and promotion committees, and other sanctioning bodies need to recognize the contribution that publicly available data make to the scholarly community.

We realize that following this recommendation, the others we have offered, and, of course, the rules of inference to which we devoted most of this Article will confront law schools and their

---

348 http://www.icpsr.umich.edu/index.html
349 http://csf.colorado.edu/isa/data/data-archive.html
350 http://www.essex.ac.uk/qualidata/
351 http://eh.net/ehresources/
352 http://lib.stat.cmu.edu/
353 http://qed.econ.queensu.ca/jae/
354 http://web.polmeth.ufl.edu/
355 http://TheData.org

faculty with a host of challenges. Meeting them, we suspect, will not be altogether difficult; after all, the interest in empirical research and the norms supporting documentation are in place. It is now a matter of making productive use of them by heading in the direction we have commended—a direction that we cannot help but believe will lead to substantial improvements in legal scholarship, as well as in public policy.