37

QUANTITATIVE APPROACHES TO EMPIRICAL LEGAL RESEARCH

LEE EPSTEIN AND
ANDREW D. MARTIN[1]

[1] For research support, we thank the National Science Foundation, Northwestern University School of Law, and the Center for Empirical Research in the Law at Washington University. For their very helpful comments, we thank the editors of this volume. We adapt some of the material in this Chapter from Epstein and King (2001); Epstein and Martin (2005); Epstein, Martin and Boyd (2007); Epstein, Martin and Schneider (2006); and Epstein and Martin's annual workshop, *Conducting Empirical Legal Research*.

The title of this Chapter seems too wordy. Why call it doing "empirical *legal* research," and not simply doing "empirical research"? After all, regardless of whether empirical researchers are addressing a legal question or any other, they follow the same rules—the rules that enable them to draw inferences from the data they have collected (Epstein and King, 2002; King, Keohane, and Verba, 1994). What's more, because empirical research in law has methodological concerns that overlap with those in Biology, Chemistry, Economics, Medicine and Public Health, Political Science, Psychology, and Sociology, empirical legal researchers can adopt methods from these other disciplines to suit their own purposes.

On the other hand, in virtually every discipline that has developed a serious empirical research program—law not excepted—scholars discover methodological problems that are unique to the special concerns in that area. Each new data source 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. As of this writing, there is no "legalmetrics" but that should happen soon enough (though probably not before this Chapter appears in print).

In short, with a few wording substitutions here and there, much of what follows pertains to all empirical research. But *much* is not *all*. Recognizing that empirical legal work is unique in various ways, as we describe the research process we also outline some of the field's distinct challenges—most notably, how to communicate complex statistical results to a community lacking in statistical training.

We begin by describing the research process. Then, in sections II-V we flesh out the various components of the process: designing research, collecting and coding data, analyzing data, and presenting results.

I. Conducting Empirical Legal Research: An Overview

How do scholars implement quantitative empirical research? What challenges do they confront? To begin to formulate responses, consider a legal question at the center of hundreds, perhaps thousands, of lawsuits each year: Do employers pay men more than women solely because of their gender? Next consider how researchers who faced absolutely no constraints—i.e., researchers with more powers than

Batman, Superman, and Wonder Woman combined—would address this question. If we were the researchers, we would begin by creating a workplace, randomly drawing a worker from the workforce population, randomly assigning a sex (say, male) to the worker, instructing him to enter the workplace, and observing his wage.[2] Next, we would reverse time, and assign the same worker the other sex (female), send her into exactly the same workplace, and observe her wage. If we observed a difference in the wages of our two workers—such that the same employer paid the male version less than the female version—then we might conclude that, yes, gender causes pay inequities.

Unfortunately, researchers aren't superheroes; they usually don't have the power to create a workplace and assign a sex. And they certainly don't have the power to rerun history. This is known as *fundamental problem of causal inference* (Holland, 1986: 947). It simply means that researchers can only observe the factual (e.g., a female worker's salary, if in fact the worker was a female) and not the counterfactual (e.g., a male worker's salary, if in fact the worker was female).[3]

This is a problem without a solution but scholars have developed various fixes. The gold standard along these lines is a proper experiment—that is, an experiment in which the researcher randomly selects subjects from the population of interest and then randomly assigns the subjects to treatment and control conditions (see Ho et al., 2007). Very few experiments in empirical legal studies actually meet the first condition (random selection from the population) but some scholars have tried to meet the second. Jeffrey J. Rachlinkski and his colleagues (2006), for example, recruited 113 bankruptcy court judges to participate in an experiment designed to detect whether the race of a party affected the judges' decisions.[4] They asked the judges to read the same case materials but unbeknownst to the judges, the researchers randomly assigned them to a control or treatment group. Those judges in the control group were led to believe that the debtor was white; those in the treatment group were led to believe that the debtor was black. (It turned out that race did not affect the judges' decisions.)

This is a reasonable approach to the fundamental problem of causal inference. But, sadly, it is infeasible for many empirical legal projects—including studies of pay equity (no experiment can assign a sex to workers). It is not even feasible for most analyses of judicial behavior (the Rachlinkski et al. study is a notable exception). To provide but one example, suppose we wanted to investigate the extent to which female judges affect the decisions of their male colleagues. No US Court of Appeals would allow us to manipulate the composition of panels so that we could identify

[2] Though it should be obvious, for this hypothetical we are assuming that the employer is assigning wages intentionally, not randomly.

[3] For a more formal accounting of this type of analysis, many scholars have adopted a potential outcomes framework—posited by Neyman (1935) and Rubin (1973, 1974), thoroughly reviewed in Holland (1986), and recently applied in the social sciences by Imai (2005), Epstein, et al. (2005), and Boyd, et al. (2008).

[4] Their research tested for other biases as well, including anchoring and framing.

a possible gender effect. We could say the same of the other institutions of government. Can you imagine the President of the United States agreeing to nominate two judicial candidates identical in all respects except that one is highly qualified and the other highly unqualified just to enable us to learn whether qualifications affect the confirmation votes of US senators? We can't.

The upshot is that most empirical legal researchers simply do not have the luxury of analyzing data they developed in an experiment (i.e., experimental data). Instead, they must make use of data the world—not they—generated (i.e., observational data): salaries paid to workers by real companies; the decisions of judges in concrete cases; the votes cast by senators over the President's nominee to the federal courts. And this, of course, substantially complicates the task empirical legal researchers confront. While experimental data—generated by random assignment to treatment and control groups—effectively minimize the confounding effects of other variables, the same cannot be said of observational data. For those data, researchers must invoke statistical techniques (discussed below) to accomplish the same thing.

Because observational datasets are so much more common in quantitative empirical legal research, in what follows we focus on strategies for working with them. It is important to keep in mind, however, that other than issues of data generation and control (statistical versus experimental), experimental and observational studies are not altogether different for our purposes. Either way, scholars tend to execute them in four steps: they design their projects, collect and code data, conduct analyses, and present results.[5]

Research design largely (though not exclusively) involves the process of moving from the conceptual to the concrete. To return to our example of pay equity, suppose the researcher hypothesizes that once she takes into account the experience of the workers, males earn no more than females. However plausible this hypothesis, the researcher confronts a non-trivial problem in assessing it: how to operationally define the concept of "experience." Is it years from degree? Years in the workforce? Months in the same job? More generally, before researchers can answer empirical legal questions—actually before they can even collect the first

[5] These are indeed the key components, and in the Sections to follow we describe them in order, from designing research to conducting analyses. Nonetheless, empirical legal scholars rarely regard their research as following 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. On the other hand, being flexible does not mean that researchers do or should do ad or post hoc adjustment of theories to fit idiosyncrasies. Adjustments made to harmonize theory with data, of course, do not constitute any confirmation of the theory at all. While it is fine to use data to create theory, investigators know they must consult a brand new data set, or completely different and previously unanticipated testable consequences of the theory in the same data set, before concluding that data confirm their theory. For more on the idea of research as a "dynamic process conforming to fixed standards," see Epstein and King (2001).

piece of data—they must devise ways to clarify concepts such as experience so that they can observe them. All of this and more appear on that first (metaphorical) slide.

Data collection and coding entails translating information in a way that researchers can make use of it. For a study of pay equity, the researcher may have piles of pay stubs and employee records. Unless the researcher can transform the piles into data she can analyze the study cannot proceed.

Data analysis typically consists of two activities. First, researchers often summarize the data they have collected. If, for example, we collect information on a sample of 50 workers' salaries in a firm with 500 workers, it may be interesting to know the average salary for the men in our sample and the average salary for the women. Second, analysts use data to make inferences—to use facts they know (about the salaries, gender, experience, and so on of the 50 workers in their sample) to learn about facts they do not know (the salaries, gender, experience, and so on of the 500 workers). To perform inference in quantitative studies, researchers employ various statistical methods. Worth noting, though is that use of statistics presupposes that the study is well designed and the data are of a sufficiently high quality. If either the design is poor or the data inadequate, researchers will be unable to reach inferences of high quality. In other words, without a proper research design no statistical method can provide reliable answers; not even the best statistician cannot make lemonade without lemons.

Finally, once empirical legal analysts have drawn inferences from their data, they must be able to communicate their results to a community that may have little (or no) knowledge of even simple statistics. Doing so effectively blends both art and science, and requires careful consideration of both the project and the intended audience.

These are the contours of the research process. Let us now flesh them out to the extent possible given space constraints.

II. Designing Research

It should go without saying that before researchers can design their project, they must have one. To "have a project" usually means that the analyst has a *question* she wishes to answer and has *theorized* about possible responses.

Research questions in empirical legal studies come from everywhere and anywhere. Perhaps scholars see a gap in the existing literature or perhaps they think the literature is incomplete or even wrong. Sometimes questions come

from current events—whether a new law is having the desired (or any) effect or whether a court decision is efficacious—and sometimes they come from history. A perusal of any socio-legal journal would provide evidence of these and other motivations.

The variation is not unexpected. Empirical legal scholars are a diverse lot, with equally diverse interests. What their questions have in common, though, may be just as important: virtually all are quite conceptual. Consider a variation on the question we asked at the onset:

Do males and females who have the same level of experience earn the same amount of money?

However important this question, it is not one that even the best empirical legal project can ever address. Rather, the question the study will actually answer comes closer to this:

Do males and females who have been in the workforce for the same number of years net the same salary per month?

Note that the first form of the question contains several concepts—"earn" and "experience"—which researchers cannot directly observe. Only by clarifying these concepts, as the second form does, can the researcher empirically answer the question. Because this is more or less true of every empirical project, a major research challenge is to tighten the fit between the question asked and the question actually answered. If it is too loose the researcher cannot, at the end of the day, claim to have answered the question she initially posed.[6]

Once analysts have settled on a research question, they usually begin *theorizing* about possible answers they can use to develop *observable implications* (sometimes called hypotheses or expectations).[7] A theory is simply a reasonable and precise answer to the research question. An observable implication is a claim about what we would expect to observe in the real world if our theory is right—typically, a claim that specifies a relationship between (or among) a dependent variable (what we are trying to explain) and an independent variable(s) (what our theory suggests explains the dependent variable) (Epstein and King, 2002: 61–2).

Theorizing is a big topic, one to which we can hardly do justice in this short Chapter. So two observations will have to suffice. First, theorizing in empirical legal scholarship comes in many different forms: in some projects theories are quite big

[6] How to ensure a good fit? We turn to this question when we tackle the subject of measurement.

[7] Some might argue that these steps are unnecessary in research motivated purely by policy concerns. Not so. Because the statistical methods we describe momentarily are designed to test hypotheses, the researcher should, well, develop some hypotheses to test.

and grand, seeking to provide insight into a wide range of phenomena (e.g., rational choice theory in law and economics); others are simple, small, or tailored to fit particular situations. For the purposes of conducting an empirical study, this distinction may not be very important.

What is important—and this takes us to the second key point—is that the researcher extract observable implications from the theory. The reason is simple. Just as analysts almost never actually answer the question they pose, they almost never directly test their theory. Rather, they only indirectly assess it by evaluating the observable implications that follow from it.

To see the point, return to our question about pay equity between males and females, and consider the following theories and their observable implications.

Difference Theory
Owing to discriminatory judgments about worth, employers pay females less than comparable males.

Please confirm we have retain the underline.

Observable Implication
All else being equal (e.g., experience), if my theory is correct, we should observe females earning less than males.

Efficiency Theory
Because labor markets are efficient, any observed differences between male and female workers are a product of experience, quality, productivity, and so on.

Observable Implication
All else being equal (e.g., experience), if my theory is correct, we should observe females and males earning the same

Note that in neither instance—no matter how good their design, their data, and their methods—will the researchers be able to conclude that their theory is right or wrong (that discriminatory judgments lead to pay inequity or that efficient markets lead to pay equity). All they will be able to say is whether their data are *consistent* with the observable implications following from their theory.

And even saying that involves hard work. The problem, yet again, is that observable implications are *conceptual* claims about the relationship between (or among) variables. To evaluate these, researchers must delineate how they actually can observe them in the real world. They must, in short, move from the abstract to the concrete—a task that forms the core of research design and that Figure 1 depicts.
Note that in the clarification process the researcher translates abstract notions, such as "experience" and "earnings," into the far more concrete "years in the workforce" and "gross annual income." Unlike the abstractions, researchers can observe and measure "years in the workforce" and so on.

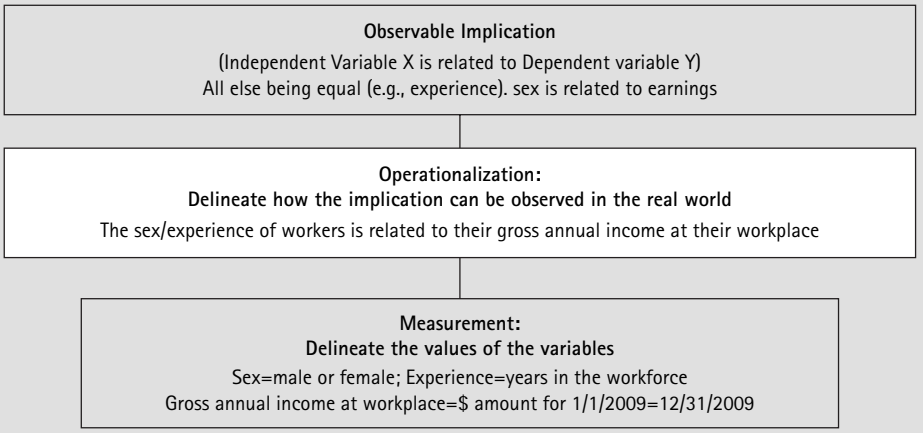


Figure 1. The process of clarifying observable implications so that researchers can evaluate them.

Note that in the clarification process the researcher translates abstract notions, such as "experience" and "earnings," into the far more concrete "years in the workforce" and "gross annual income." Unlike the abstractions, researchers can observe and measure "years in the workforce" and so on.

But how do analysts evaluate their choices and procedures? Why "years in the workforce" and not "years from degree," "months in the same position," or any of the other many plausible measures of experience? Typically, researchers look to the reliability and validity of their measures. Reliability is the extent to which it is possible to replicate a measure, reproducing the same value (regardless of whether it is the right one) on the same standard for the same subject at the same time. Measures of high reliability are preferable to those with lower levels of reliability. Validity is the extent to which a reliable measure reflects the underlying concept being measured. Along these lines, we might consider whether the measure is facially valid, that is, whether it comports with prior evidence or existing knowledge, among other criteria.

There is another test to which many researchers put their measures: robustness checks. Suppose we settled on "years in the workforce" as our measure of experience but believed that "months in the same position" was plausible as well. In our statistical work, we might try both hoping to obtain consistent results regardless of the particular measure. This procedure does not tell us whether "years in the workforce" is a better measure than "years in the same position" but it does help to anticipate a question put to many empirical legal scholars: "What if you had used measure Y instead of measure X? Would your results have been the same?"

III. Collecting Data and Coding Variables

Once researchers have designed their project—that is, they have filled out the first slide—they typically turn to *collecting* and *coding* their data—the makings of the second slide. By this point, it should go without saying, though we shall say it anyway, that we can hardly scratch the surface of either; both deserve Chapters of their own.

What we can do instead is offer some brief counsel, beginning with data collection—actually, with a crucial step before data collection: determining whether the data the researcher needs already exist in the form she needs it. For decades now, empirical legal scholars have been amassing datasets—some for particular projects and others, the so-called "multi-user" datasets, designed for application to a wide range of problems. Either way, it is entirely possible (even probable in some areas of empirical legal studies) that researchers can locate suitable data without having to invest in costly from-scratch data-collection efforts.

A few examples suffice to make the point. If analysts are interested in cases decided by the US Supreme Court, they should proceed directly to the US Supreme Court Database (<http://supremecourtdatabase.org>). This remarkable resource houses scores of variables on Supreme Court cases decided since 1953, including the legal provisions under analysis, the identity of the majority opinion writer, and the votes of the justices. A similar dataset, the US Courts of Appeals Database, exists for cases decided by the US circuit courts (at: http://www.cas.sc.edu/poli/juri/). For the researcher interested in public opinion, the General Social Survey and the American National Election Study (both available via an intuitive interface at: <http://sda.berkeley.edu/archive.htm>) are natural places to look for relevant data. For other types of projects, we recommend visiting the websites of the Inter-University Consortium for Political and Social Research (<http://www.icpsr.umich.edu/>) and the IQSS Dataverse Network <http://dvn.iq.harvard.edu/dvn/>), both of which serve as repositories for (or have links to) existing datasets. Federal and state governments and agencies too retain enormous amounts of information of interest to empirical legal scholars, including data on population demographics, economic indicators, and court caseloads. Last but not least, experience has taught us that a well-formulated Internet search can unearth datasets that scholars maintain on their own websites.

If the data simply do not exist in an analyzable form, empirical legal researchers can and do make use of a wide variety of data-generation mechanisms. They amass numerical data from structured interviews or surveys, from field research, from public sources, from private papers, and on and on. Each has its strengths and weak-

nesses (as do archived datasets) and it is the researchers' job to learn, understand, and convey them.

Still, within all this variation, two principles governing the data-collection process apply to most empirical legal research projects. One is simple enough: As a general rule, researchers should collect as much data as resources and time allow because basing inferences on more data rather than less is almost always preferable. To see the point, think about a study designed to study gender pay equity in academia. The more professors included in the study, the more certain the conclusions the analyst can reach. As a practical matter, however, diminishing returns kick in and settling on a sample size (as opposed to including all professors) is good enough. For example, one can estimate a proportion with ±2% margin of error with a random sample of approximately 2400 observations; the number increases dramatically to 9,600 for ±1%. This is why most public opinion surveys query, at most, a couple thousand respondents. As discussed in more detail below, this "margin of error" is sometimes referred to as the "sampling error" or the "confidence interval" (e.g., "CI ±3%" in examples below).

Second, if researchers cannot collect data on all members of the population of interest (e.g., all professors)—and they rarely can—they must invoke selection mechanisms that avoid selection bias (mechanisms that don't bias their sample for or against their theory). For large-*n* studies (where *n*=number of participants) only *random probability sampling* meets this criterion.[8] A random probability sample involves identifying the population of interest (all professors) and selecting a subset (the sample) according to known probabilistic rules. To perform these tasks, the researcher must assign each member of the population a selection probability and select each person into the observed sample according to these probabilities. (Collecting all the observations is a special case of random selection with a selection probability of 1.0 for every element in the population.)[9]

Researchers can implement random sampling in various ways depending on the nature of the problem. For a study of pay equity in the academy, for example, we could draw an equal probability sample—a sample in which all professors have an equal chance of being selected. If, on the other hand, we wanted to include all racial and ethnic groups in our study and worried that our sample, by chance, might not include, say, any American Indians, stratified random sampling may be a better strategy. The idea is to draw separate equal-probability-of-selection random samples within each category of a variable (here, race/ethnicity).

[8] For advice on small-*n* studies, see Epstein and King (2002: 112–13); King, Keohane, and Verba (1994: 124–8).

[9] Dealing with data collected on a population raises some foundational statistical issues. One approach is to argue that an observed population is a "sample" from possible histories, and as such, traditional inferential statistics can be used. Another option is to simply summarize the data and not report measures of uncertainty. The ideal approach, from our perspective, is to adopt a Bayesian approach and treat the parameters as random variables, not the data.

Whatever the procedure (so long as it involves random selection for large-*n* samples!), the legal researcher will typically end up with piles or computer files of questionnaires, field notes, court cases, and so on. *Coding variables* is the process of translating the relevant properties or attributes of the world (*i.e.,* variables) housed in the piles and files into a form that the researcher can then analyze systematically (presumably after they have chosen appropriate measures to tap the underlying variables of interest).

Coding is a near-universal task in empirical legal studies. No matter whether their data are quantitative or qualitative, from where their data come, or how they plan to analyze the information they have collected, researchers seeking to make claims or inferences based on observations of the real world must code their data. And yet, despite the common and fundamental role it plays in research, coding typically receives only the briefest mention in most volumes on empirical research; it has received almost no attention in empirical legal studies.

Why this is the case is a question on which we can only speculate, but an obvious response centers on the seemingly idiosyncratic nature of the undertaking. For some projects researchers may be best off coding inductively, that is, collecting their data, drawing a representative sample, examining the data in the sample, and then developing their coding scheme. For others, investigators proceed in a deductive manner, that is, they develop their schemes first and then collect/code their data. For still a third set, a combination of inductive and deductive coding may be most appropriate.[10]

Nonetheless, we believe it is possible to offer three generalizations about the process of coding variables. First, regardless of the type of data they collect, the variables they intend to code, or even of whether they plan to code inductively or deductively, at some point empirical legal researchers require a coding schema, that is, a detailing of each variable of interest, along with the values of each variable. For example, in a study of the effect of female judges on the votes of their male colleagues, the variable Vote of the Judge would obviously figure prominently; for this variable we might code three values: the judge voted to "affirm," to "reverse," or "other." With this sort of information in hand, investigators can prepare codebooks—or guides they employ to code their data and that others can use to replicate, reproduce, update, or build on the variables the resulting database contains and any analyses generated from it.

Second, depending on the type of data and variables, developing schema and creating codebooks are not always easy or straightforward tasks. To see this, reconsider the seemingly simple example of the variable Vote of the Judge. We just listed three

[10] Some writers associate inductive coding with research that primarily relies on qualitative data and deductive coding, with quantitative research. Given the [typically] dynamic nature of the processes of collecting data and coding, however, these associations do not always or perhaps even usually hold. Indeed, it is probably the case that most researchers, regardless of whether their data are qualitative or quantitative, invoke some combination of deductive and inductive coding.

values (affirm, reverse, and other) but what of a vote "affirming in part and reversing in part"? Should we code this as "other," even if the judge gave the plaintiff some relief? For that matter, what should we make of the "other" category? Depending on the subjects under analysis, it may be appropriate (meaning that it would be an option exercised infrequently) or not. But our more general point should not be missed: Accounting for the values of the variables of interest, even of seemingly simple ones, may be tricky.[11]

To be sure, following best practices can help; for example, ensuring that the values of the variables are exhaustive, creating more (rather than fewer) values, establishing that the values of the variables are mutually exclusive, and more generally, pretesting the schema (for more details see Epstein and Martin, 2005). But there is one assumption that all the rules and guidelines make—and this brings us to our third point: Researchers must have a strong sense of their project, particularly about the piece of the legal world they are studying and how that piece generated the data they will be coding, as well as the observable implications of the theory that they will be assessing (see, e.g., Babbie, 2007: 384; Frankfort-Nachmias and Nachmias, 2007). Even adhering to simple rules will be difficult, if not impossible, if the researcher lacks a deep understanding of the objects of her study and an underlying theory about whatever feature(s) of their behavior for she wishes to account.

IV. Analyzing Data

If research design is the first overhead slide and collecting and coding data, the second, then data analysis enables researchers to compare their overlap. When the overlap between the observable implications and data is substantial, analysts may conclude that the real world confirms their hunches; if the overlap is negligible, they may go back to the drawing board or even abandon the project altogether.

How do empirical legal scholars perform this task? The answer depends in no small part on their goals. If the goal is to summarize the data they have collected (say, the salaries of all male and female professors at their school), then some simple measures of central tendency (e.g., means, medians) and dispersion (e.g., standard deviations, ranges) might suffice. These will give researchers a feel for the distributions

[11] More generally, the relative ease (or difficulty) of the coding task varies according to the types of data with which the researcher is working, the level of detail for which the coding scheme calls, and the amount of pretesting the analyst has conducted.

of their variables that, depending on the number of cases, they could not possibly develop from looking at a column of data.

For the vast majority of empirical legal projects, however, making inferences—using facts we know to learn about facts we do not know—is the goal. Rarely do we care much about the, say, 50 individuals or 100 cases in our sample. Rather, we care about what those 50 individuals or 100 cases can tell us about all the employees of the corporation or all the cases. In quantitative research, inferences come in two flavors: *descriptive* and *causal*. Descriptive claims themselves can take several forms but some seem quite a kin to data summaries. Suppose, for example, that we collected data on 100 court cases involving employment discrimination and learned that, on average, appellate court panels held for the plaintiff in 40% of the cases. In and of itself this figure of 40% (a summary of the data), probably isn't all that interesting to our readers or us. What we want to learn about is the fraction of *all* employment discrimination cases in which all courts held for the plaintiff. That is, we want to use what we know (the 100 cases we have collected) to learn about what we do not know (the cases we haven't collected). This is the task of drawing a descriptive inference. We do not perform it by summarizing facts; we make it by using facts we know—the small part of the world we have studied—to learn about facts we do not observe (the rest of the world). Researchers call the "small part" a sample and the "world" a population. (An important part of performing descriptive inference is *quantifying* the uncertainty we have about that inference. We discuss this in greater detail below.) It is important to keep in mind that when dealing with data coming from a non-probability sampling neither descriptive nor causal inferences can be drawn.

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 (for example, the fraction of cases decided in favor of the plaintiff) takes on when a "treatment" is applied (for example, a female judge serves on the panel) and the average value the dependent variable takes on when a "control" is applied (for example, if no female judge sits on the panel). The *causal effect*—the goal of the process of causal inference—is this difference, the amount the fraction of decisions in favor of the plaintiff increases or decreases when we move from all-male panels to panels with a female.

How do quantitative empirical researchers go about making descriptive or causal claims? Assuming they have appropriately designed their projects and appropriately amassed and coded their data, they make use of *statistical inference*, which entails examining a small piece of the world (the sample) to learn about the entire world (the population), along with evaluating the quality of the inference they reach. Conceptually, statistical inference is not all that hard to understand; actually we confront such inferences almost every day. When we open a newspaper, we might find the results of a survey showing that 70% (± 5% margin of error) of American voters have confidence in the US president. Or when we read about a scientific study indicating that a daily dose of aspirin helps 60% (95% CI ± 3%) of

Americans with heart disease. (95% CI and ± X% are explained below.) In neither of these instances, of course, did *all* Americans participate. The pollsters did not survey every voter, and the scientists did not study every person with heart problems. They rather made an inference (in these examples, a descriptive inference) about all voters and all those stricken with heart disease by drawing a sample of voters and of ill people.

But how do the researchers go about making the statistical inference (for example, 70% of all American voters have confidence in the president) and assess its quality (that is, indicate how *un*certain they are about the 70% figure, as indicated by the ± 5%)? It is one thing to say that 70% of the voters in the sample have confidence in the president (this is summarizing or describing the data); but it is quite another to say that 70% of *all* voters have confidence (this is the descriptive inference).

To support the first claim, all analysts need do is tally (i.e., *summarize*) the responses to their survey. To support (and evaluate) the second, they must (1) draw a random probability sample of the population of interest and (2) determine how certain (or uncertain) they are that the value they observe from their sample of voters (70%), called the *sample statistic,* reflects the population of voters, the *population parameter.*

We already have discussed (1)—drawing a random sample—so we only need reiterate here that this step is crucial. If a sample is biased (for instance, if Democrats had a better chance of being in the pollsters' sample than Republicans), researchers cannot draw accurate conclusions.

Assuming researchers draw a random probability sample, they can move to (2) and make a (descriptive) inference about how well their sample reflects the population. Or, to put it another way, they can convey their *degree of uncertainty* about the sample statistic. Surveys reported in the press, for example, typically convey this degree of uncertainty as "the margin of error," which is usually a 95% confidence interval (or 95% CI). When pollsters report the results of a survey—that 70% of the respondents have confidence in the president with a ±5 margin of error—they are supplying the level of uncertainty they have about the sample statistic of 70%. That is, the true fraction of voters who have confidence in the president will be captured in the stated confidence interval in 95 out of 100 applications of the same sampling procedure. The fact that the data come from a random sample is what makes it possible to use the rules of probability to compute these margins of error.

Note that this information does not say exactly where, or whether, the population (parameter) lies within this range. (In fact, the parameter either falls within the interval or not; only an all-knowing researcher would ever know.) What is critical, however, is that if the researcher continues to draw samples from a population of voters, the mean of the samples of voters will eventually equal the mean of the population, and if the researcher creates a specialized bar graph called a histogram showing the distribution of the individual sample means, the resulting shape would resemble a

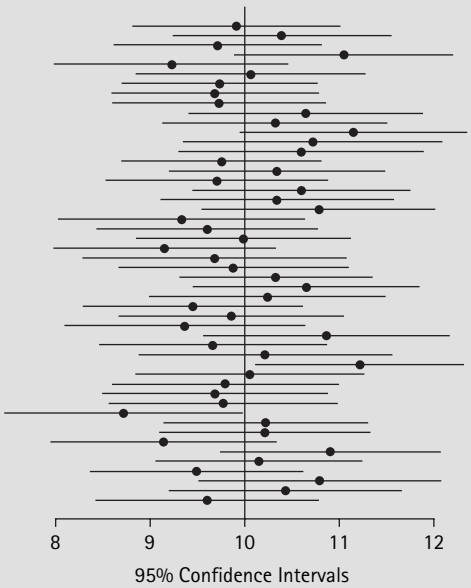


Figure 2. Confidence intervals for a known population mean ten for fifty random samples from a population.

normal distribution. This is what enables researchers to make an inference—here, in the form of a sample statistic and a margin of error—about how all voters (the population) feel about the president by observing a single sample statistic. For the sake of illustration, consider Figure 2. Here we show the confidence intervals computed from 50 random samples from a population where the known parameter of interest is ten. The 95% confidence intervals are constructed to contain the true parameter 95% of the time. Here in all but two samples the horizontal confidence intervals contain the known parameter value. Of course, in any application we do not know the parameter value (if we did we would not need to perform inference!), but we use confidence intervals that over repeated samples will return the right answer a high percentage of the time.

This pertains to descriptive claims but it is important to draw a statistical inference when performing causal inference as well. Suppose that the average monthly income for the male professors in our sample of employees was $4,200, while for the females it was $3,900, yielding a difference of $300 in this sample. There are two possible explanations for the $300 difference (assuming *all else is constant*, a phrase we explain below). It might be the case that it is due solely to the particular sample we randomly drew; in other samples from the population the difference might only be $10, or women might make, on average, $250 more than men. It is also possible that in the population, men actually earn more than women.

The process researchers use to make this determination is called *hypothesis testing*. A hypothesis test tells us whether differences across groups are simply an artifact of sampling (the first possible explanation), or whether meaningful differences exist in the population (the second possibility). In the latter case we would say the difference is *statistically significant*. All statistical significance means is that sampling alone cannot explain the observed difference, and as such, it is likely that differences exist in the population. One would conclude a relationship is statistically insignificant when the difference in the sample can be explained by sampling alone.

In addition to statistical significance, it is important to consider the *substantive significance* of any finding. A $1,000 per month difference in salary is certainly large; an $8 per month difference is not. Both could be statistically significant, but only the first would be substantively significant. Accordingly, it is crucial for empirical researchers to compute and report the size of the differences—in addition to reporting the results of hypothesis tests—so that the reader can ascertain whether the findings are substantively important. In the following section we recommend using graphics to report these differences.

But before turning to data displays, one final topic deserves some attention: the assumption of "all else being constant" or "all things being equal." This assumption takes us back to a point we made at the onset; namely, when working with data generated by the world, most of the time "all else is constant" or "all things being equal" is untenable. It is quite possible, for example, that male professors in our sample do not have the same experience as females. Thus, just naively comparing the average salaries across the two groups would not provide a reliable causal inference.

Today, there are two approaches commonly used for making causal inferences from observational data. One type of analysis is *multiple regression* analysis, and related regression models (such as logistic regression). Regression models work by allowing the researcher to hold all other measured variables constant while assessing the relationship of interest. In this example, we could see whether the difference in salaries persisted by controlling for experience. Regression models have been used for decades and are the most common tool in empirical legal research. For many types of research they work quite well, but they do require some strong assumptions about the relationship between the key causal variable and the outcome variable of interest (see Imai, 2005).

Another set of methods called *matching methods* is becoming more popular in applied statistics. These cutting-edge tools are making their way into empirical legal studies (Epstein, et al. 2005; Greiner 2008), and for many reasons we predict that their use will increase in the coming decades. The idea, to return to the example of pay equity, is to match most-similar male and female professors, and then compute differences between the matched observations. Once researchers have made the matches, these methods allow them to treat observational data as if it were experimental.

Regardless of whether one uses regression analysis or matching to control for alternative explanations, a causal inference is just a statistical inference about a difference. At bottom what researchers want to know is whether observed differences in a sample represent the same differences in a population.

V. The Last Step: Presenting the Results of Empirical Legal Research[12]

Just as scholars have been improving methods for causal inference, they have been working on approaches to convey the results of their studies. These developments should be of particular interest to quantitative empirical legal scholars who often must communicate their findings to judges, lawyers, and policy-makers—in other words, to audiences who have little or no training in statistics. Too often, though, analysts fail to take advantage of the new developments thus missing an opportunity to speak accessibly to their community.

To see the problem, consider an example adapted from a study that seeks to explain the votes cast by US senators on Supreme Court nominees (Epstein et al., 2006).[13] Briefly, the authors operate under the assumption that electorally minded senators vote on the basis of their constituents' "principal concerns in the nomination process" (Cameron, Cover, and Segal, 1990: 528). These concerns primarily (though not exclusively) center on whether a candidate for the Court is (1) qualified for office and (2) ideologically proximate to the senator (i.e., to his or her constituents). Consequently, the two key causal variables in their statistical model are (1) the degree to which a senator perceives the candidate as *qualified* for office and (2) the *ideological distance* between the senator and the candidate, such that the more qualified the nominee and the closer the nominee is to the senator on the ideological spectrum, the more likely the senator is to cast a yea vote. Also following from the extant literature, the researchers control for two other possible determinants of senators' votes: whether the President was "strong" in the sense that his party controlled the Senate and he was not in his fourth year of office; and whether a senator is of the same political party as the President.

To assess the extent to which these variables help account for senators' votes, the researchers employed logistic regression, a common tool in legal scholarship when

[12] We draw material in this section from Epstein, Martin & Boyd (2007); Epstein, Martin & Schneider (2006); Gelman, et al. (2002); King, et al. (2000).

[13] Since publication of their study, Epstein, et al. have updated their dataset (available at: <http://epstein.law.northwestern.edu/research/Bork.html>). We rely on the updated data.

Table 1. The "Ugly" Table. Logistic regression analysis of the effects on individual senators' votes on 41 Supreme Court nominees (Black through Alito). Cell entries are logit coefficients and robust standard errors. *p <.01.

Variable	Coefficient	Standard Error
Lack of Qualifications	−4.11*	0.22
Ideological Distance	−3.92*	0.23
Strong President	1.01*	0.13
Same Party	1.45*	0.15
Constant	3.32*	0.15
N	3809	
Log-likelihood	−916.91	
$X^2_{(4)}$	632.68	

the dependent variable is binary. Table 1 displays the results, and they seem to lend support to the researchers' hypothesis. For example, the * on the coefficient for lack of qualifications variable tells us that a statistically significant relationship exists between qualifications and voting: the lower a nominee's qualifications, the higher the likelihood that a senator will vote against the nominee.

On the other hand, tables of this sort (which run rampant in empirical legal scholarship) are not just ugly and off-putting to most readers; they communicate virtually no information of value either to the audience or even to the researchers themselves. Most lawyers, judges, and even law professors do not understand terms such as "statistical significance," much less "logit coefficient."

How might empirical legal scholars improve their data presentations? Adhering to three general principles would be a good start. First, we recommend that analysts communicate substance, and not only statistics. Reconsider this statement:

> In looking at Table 1, we see that the coefficient on the variable lack of qualifications of -4.11 is "statistically significant."

This is not wrong but the emphasis on the coefficient is more than off-putting; it fails to convey useful information. In fact, all we learn from the -4.11 coefficient on lack of qualifications is that, controlling for all other factors, as we move from the most qualified to the most unqualified nominee we move down 4.11 on a logit scale. To make matters worse, because the logit scale is non-linear, moving down 4.11 units will result in different probabilities of a yea vote depending on where we start on the scale.

Because few of their readers would understand what any of this means, it is no wonder many empirical legal scholars simply say "the coefficient on lack of qualifications is statistically significant at the.01 level." But this too isn't an informative statement to many readers; it isn't even informative to readers with statistical

training (a very small fraction of the legal community). It tells us is that qualified candidates are more likely to receive a yea vote than unqualified candidates but not how much more likely. 0.2 times more likely? 2 times? Or perhaps even 4 times? We probably wouldn't be very impressed, for example, if all else being equal, the predicted probability of senator voting for a very qualified candidate was 0.11 and for a very unqualified candidate was 0.14. Certainly, a quantity such as a predicted probability is what matter most to readers of empirical legal scholarship. But it is not one that they can learn from a tabular display of logit coefficients.

This is why we recommend supplying readers with a quantity of interest; that is, replace "In looking at Table 1, we see that the coefficient on the variable lack of qualifications of -4.11 is statistically significant" with:

> Other things being equal,[14] when a nominee is perceived as highly unqualified the likelihood of a senator casting a yea vote is only about **0.24**. That probability increases to **0.92** when the nominee is highly qualified.

Statements of this sort are easy to understand even by the most statistically challenged members of the legal community.

Second, we suggest that when they perform inference, researchers convey their uncertainty. To see the point, think about the statement above—that the likelihood of a senator casting a yea vote is only about 0.24 when the candidate is unqualified. This figure of 0.24 represents the researchers' "best guess" about the likelihood of a senator voting yea based on qualifications. But we know that error or uncertainty exists around that best guess. It is simply a fact of statistical analysis that we can never be certain about our guesses because they themselves are based on estimates.

Most quantitative empirical legal scholars appreciate this fact and supply the error surrounding their *estimated coefficients*. Statements such as this are not uncommon:

> In looking at Table 1, the coefficient on the variable lack of qualifications (-4.11 with a standard error of 0.22) is statistically significant at the.01 level.

True, this conveys uncertainty in the form of a standard error around the estimate but of what value is it? None, it turns out, because all the error value supplies is an estimate of the standard deviation of the estimated coefficient—which, standing alone, is of interest to no one, readers and scientists alike.[15]

One possible fix is for empirical legal scholars to follow other disciplines and report far-more-meaningful 95% (or even 99%) confidence intervals rather than (or

[14] We use the term "other things being equal" to signify that all variables in the model (other than the variable interest, here qualifications) are fixed at particular values. In this example, we set ideological distance at its mean and strong president and same party at 0.

[15] Its value, rather, lies in computing confidence intervals.

in addition to) standard errors. In the case of lack of qualifications, the values of that interval are a lower bound of -4.54 and an upper bound of -3.69.

This interval comes closer than the standard error to conveying useful information: the researchers' best guess about the coefficient on lack of qualifications is -4.11 but they are "95% certain" that it is in the range of -4.54 to -3.69. Because 0 is not in this range (the confidence interval), the researchers and their readers can safely reject the null hypothesis of no relationship between the nominees' qualifications and senators' votes.

But even denoting the confidence interval around a coefficient would not be making the most of the model's results. When researchers say they are "95% certain" that the true logit coefficient lies between -4.54 to -3.69, they lose half their audience. What we recommend instead is combining the lesson here of relating uncertainty with the first principle of conveying substantive information:

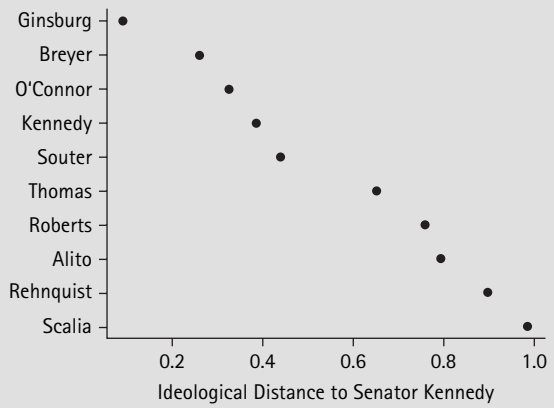
> Other things being equal, when a nominee is perceived as highly unqualified the likelihood of a senator casting a yea vote is only about 0.24 (±0.05). That predicted probability increases to 0.93, (±0.02) when the nominee is highly qualified.

Now readers need no specialized knowledge about standard errors or even confidence intervals to understand the results of the study—including uncertainty about the results. They can easily see that the researchers' best guess about the predicted probability of yea vote for a highly unqualified candidate is 0.24, though it could be as low as 0.19 or as high 0.29. Such accessible communication creates a win-win for empirical legal researchers and their audience: both are now in a far better position to evaluate the study's conclusions.

Our final recommendation is that analysts graph their data and results. With this, we are trying to convey two ideas. One is just a general point: if the goal is to give readers a feel for patterns or trends in the data, graphs are superior to tables—even for small amounts of data. Figure 3 provides an example from the project on Supreme Court nominees.

To be sure, if we looked at the table long enough some of the patterns we observe in the figure would emerge but it takes a lot more cognitive work on the part of the reader. Plus, it is unlikely that readers of empirical legal studies need such specific, precise information as in the table. So in most instances graphic displays can convey the right information without losing much.

The second idea, more relevant to the communication of results (rather than data, as in Figure 3), is that figures enable analysts to combine the first two principles we set out above (substance and uncertainty) across *many* values. Think about it this way: while substantive claims of the form "When a nominee is perceived as highly unqualified the likelihood of a senator casting a yea vote is only about 0.24 (±0.05)" may be informative, they exclude a lot of information—the values in between "highly unqualified" and "highly qualified." To provide these quantities, we could generate



Ginsburg	.092
Breyer	.261
O'Connor	.327
Kennedy	.387
Souter	.440
Thomas	.651
Roberts	.759
Alito	.794
Rehnquist	.897
Scalia	.985

Figure 3. Tables versus Figures.

Both the table and the figure provide information on the ideological distance between Senator Edward Kennedy (D-Mass.) and ten recent Supreme Court nominees. Juxtaposed against the table, the dot plot provides a more visually and cognitively appealing solution to the problem of providing the reader with information about variables of interest.

a long series of statements such as

- Other things being equal, when a nominee is perceived as highly unqualified the likelihood of a senator casting a yea vote is 0.24 (±0.05).
- Other things being equal, when a nominee is perceived as about average on the qualifications scale, the likelihood of a senator casting a yea vote is 0.83 (±0.03).
- Other things being equal, when a nominee is perceived as highly qualified the likelihood of a senator casting a yea vote is 0.93 (±0.02).

But graphing the results is a far more parsimonious, pleasing, and, for the readers of empirical legal work, cognitively less demanding approach. Underscoring these points is Figure 4. Here the reader gets a real sense of the (1) results and (2) uncertainty across the values of qualifications without having to sift through a long series of claims.

Even better, and usually necessary in multivariate analysis, is to bring in other variables of interest, as Figure 5 does. Here, we've juxtaposed qualifications against another variable: ideology, when senators and nominees are ideologically very close and when they are very distant. Specifically, in the two panels we show the probability of a senator casting a yea vote across the range of lack of qualifications and when we set ideological distance at its minimum and maximum levels. In both panels we depict our uncertainty, in the form of 95% confidence intervals, with vertical lines.

This display, we believe, is a good example of what we mean by parsimony. It conveys a great deal of information—actually it encodes 66 pieces of information—quite

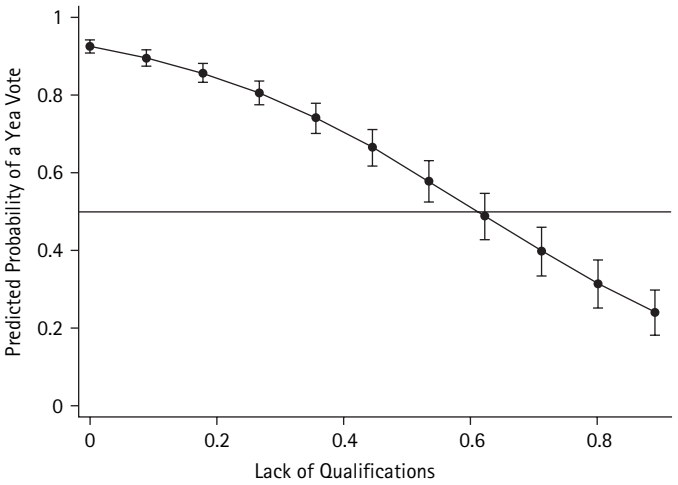


Figure 4. The effect of qualifications on Senate votes over Supreme Court nominees, from Black (1937) through Alito (2006).

The figure shows the predicted probability of a senator casting a yea vote over the range of lack of qualifications (0 is the most qualified), when we set ideological distance at its mean and the other variables in the statistical model at 0. The small vertical bars are 95% confidence intervals. Created using S-Post.

efficiently or at least more efficiently than the 66 sentences it would have taken to describe each and every result depicted in the two panels and certainly more accessibly than a table of logit coefficients.

Little more than a decade ago, implementing a graph of the sort depicted in Figure 4 would have been quite the chore: estimating the confidence intervals, in particular, was not possible for most empirical legal scholars. But now, because contemporary software packages use simulations (repeated sampling of the model parameters from their sampling distribution) to produce estimates of quantities of interest (e.g., predicted probabilities), generating assessments of error (e.g., confidence intervals) is quite easy.[16] Moreover, using the software requires no additional assumptions beyond those the researcher already has made to perform statistical inference.

Once researchers have prepared their results for presentation (and, ultimately, publication), their work would seem to be done. And, for the most part it is. But we in the empirical legal community should demand that they take one final step: archive their data and documentation. So doing ensures that empirical legal scholars adhere 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 (King, 1995). This rule does not actually require anyone to replicate the results of an article or book; it only requires that researchers provide information—in

[16] King et al.'s (2000) Clarify is an example. It uses the Monte Carlo algorithm for the simulations, and can be implemented via the Clarify plug-in for Stata.

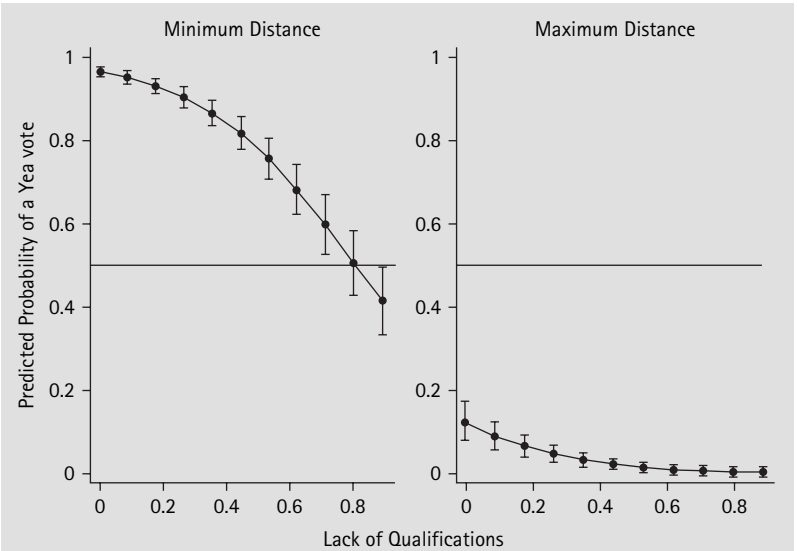


Figure 5. The effect of qualifications on Senate votes over Supreme Court nominees, from Black (1937) through Alito (2006), when the ideological distance between the Senator and nominee is very close (minimum) and very distant (maximum) and all other variables in the statistical model are set at 0.

the article or book or in some other publicly available or accessible form—sufficient to replicate the results in principle.

Why is such documentation a requisite step in conducting empirical research (regardless of whether the work is qualitative or quantitative in nature)? Epstein and King (2002) supply two answers. The first centers 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. The second reason is straightforward enough: As this Chapter has (hopefully!) made clear, the analyst's procedures may, and in most instances do, influence the outcomes they report. Readers deserve an opportunity to evaluate the researchers' choices, not to mention their data.

* * *

Designing research, collecting and coding data, analyzing data, and presenting results represent the four chief tasks of quantitative empirical legal scholarship, and

we have tried to explain some of the basics. But readers should keep in mind that mastering the four requires far more than we can possibly convey here; it requires training. That is why Ph.D. programs in the social sciences offer (at the least) a one-semester course on each.

Reading some of the books and articles we cite below would be a good start for legal scholars wishing to learn more—but only a start. To develop a full appreciation for the research process, we strongly recommend that readers contact their local social science departments.

References

Babbie, E. (2007). *The Practice of Social Research*, 11th edn. Belmont, CA: Thomson.

Cameron, C.M., Cover, A.D. and Segal, J.A. (1990). 'Senate Voting on Supreme Court Nominees: A Neo-Institutional Model', *American Political Science Review* 85: 525–34.

Epstein, L. and Martin, A.D. (2005). 'Coding Variables', in K. Kempf-Leonard (ed.), *The Handbook of Social Measurement*. Academic Press.

Epstein, L. and King, G. (2002). 'The Rules of Inference', *University of Chicago Law Review* 69: 191–209.

Epstein, L., Lindstädt, R., Segal, J.A. and Westerland, C. (2006). 'The Changing Dynamics of Senate Voting on Supreme Court Nominees', *Journal of Politics* 68: 296–307.

Epstein, L., Martin, A.D. and Boyd, C. (2007). 'On the Effective Communication of the Results of Empirical Studies, Part II', *Vanderbilt Law Review* 60: 798–846.

Epstein, L., Martin, A.D. and Schneider, M. (2006). 'On the Effective Communication of the Results of Empirical Studies, Part I', *Vanderbilt Law Review* 59: 1811–71.

Epstein, L., Ho, D.E., King, G. and Segal, J.A. (2005.) 'The Supreme Court During Crisis', *NYU Law Review* 80: 1–116.

Frankfort-Nachmias, C. and Nachmias, D. (2007). *Research Methods in the Social Sciences.* Worth.

Gelman, A., et al. (2002), 'Let's Practice What We Preach: Turning Tables into Graphs', *The American Statistician*, 56: 121.

Greiner, D. James (2008). 'Causal Inference in Civil Rights Litigation', *Harvard Law Review* 122: 533.

Ho, D.E., et al. (2007). 'Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference', *Political Analysis* 15: 199.

Holland, P.W. (1986). 'Statistics and Causal Inference', *Journal of American Statistical Association* 81: 945–70.

Imai, K. (2005). 'Do Get-Out-The-Vote Calls Reduce Turnout: The Importance of Statistical Methods for Field Experiments', *American Political Science Review* 99: 283–300.

King, G., et al. (2000). 'Making the Most of Statistical Analyses', *American Journal of Political Science* 44: 50.

King, G. (1995). 'Replication, Replication, Replication,' *PS: Political Science and Politics* 28: 443–99.

King, G., Keohane, R.O. and Verba, S. (1994). *Designing Social Inquiry,* Princeton University Press.

Rachlinksi, J. A., Guthrie, C. and Wistrich, A. J. (2006). 'Inside the Bankruptcy Judge's Mind', *Boston University Law Review* 86: 1227–65.

Rubin, D.B. (1973). 'Matching to Remove Bias in Observational Studies', *Biometrics,* 29: 159–83.

Rubin, D.B. (1974). 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies', *Journal of Educational Psychology* 6: 688–701.