EVALUATING JUDGES AND JUDICIAL INSTITUTIONS: REORIENTING THE PERSPECTIVE

Mitu Gulati
David E. Klein
David F. Levi

In September 2009, we hosted an unusual workshop at Duke Law School. The workshop focused on the empirical evaluation of judges, judging, and judicial institutions. Most work in this area has been driven by the agendas and constraints of empirical researchers, and empiricists from multiple disciplines—including history, sociology, anthropology, political science, and law and economics—participated in the workshop. But they were joined by judges and legal theorists, who were invited to take the lead in selecting the specific issues to be discussed at the workshop. The reason for the workshop's unusual makeup and structure was our conviction that the empirical analysis of judging can be dramatically strengthened through the active participation of judges and theorists. In this Essay, we explain why we think conversations among these three groups are important. Then, drawing on the workshop experience, we describe where and how we believe that cooperation could do the most to advance the empirical study of the judiciary, with special attention to issues of evaluation.

Before beginning, we should note that we paint with a broad brush here and likely fail to give credit where it is due. This Essay should be read as a comment on general tendencies rather than on individual studies or people. To the extent that it can be understood as reflecting on individuals, we are not ourselves exempt from the criticisms.

I. GOALS OF THE WORKSHOP

The empirical analysis of judicial behavior is one of the fastest growing areas of scholarship in the legal academy. The three of us bring different perspectives to this literature. Two of us, a legal scholar and a political scientist, have been involved in producing portions of that empirical literature. The third, a former federal prosecutor and United States District Judge and currently a law school dean, has been sometimes a critic but also a proponent. Disagreements among us are intense, with each at times finding the others' perspectives on courts and judges perplexing and frustrating, if not utterly misguided. Yet our debates have resulted in agreement on three important points: the emergence of this literature in legal academia is something to be celebrated, its potential has not yet been realized, and its potential would be realized more quickly if judges and legal theorists played a larger role in producing it.

One reason to celebrate the growth of this literature is the increased interaction between legal scholars and social scientists. Despite much disagreement between social scientists and legal academics on the how and why of studying judges, a number of scholars from each side have begun talking and working together, realizing that they can gain both new insights from each other and bigger audiences for their work. Because of the research experience and methodological expertise that social scientists bring to this

partnership, the resulting body of work is likely to be more rigorous and reliable than if it were produced by legal scholars alone.

But this collaboration also brings dangers—in particular, that methodological considerations will dominate theory rather than serve it, resulting in research that is hyper-technical and theoretically narrow or even irrelevant. If this happens, the research will be of little utility or interest to those who should care most about it, including the primary subjects of the literature: judges and those who depend most upon our judicial institutions. Further, neither judges nor scholars with training in other disciplines will be able to engage and be involved in the research project if it takes such a technical turn.

To our eyes, there are already disturbing signs of a trend in this direction. Specifically, in its themes and methodological approaches, the emerging empirical research in the legal academy tends to resemble the work that social scientists were already doing. Part of the cause, we think, is that legal academics and judges have been too reticent about the strengths they bring to empirical research and therefore have not pushed as hard as they might for work to move in new directions. Or perhaps it is that social scientists have been too aggressive in pushing their own perspectives, sometimes in a framework that is seen by judges as attacking them or mocking their aspirations. Either way, we do not think this trend best serves any of the groups involved.

There is a different direction available, one in which the collaboration between legal scholars and social scientists expands to incorporate more perspectives, particularly those of the primary research subjects, and becomes more accessible, interesting, and relevant. Perhaps another way to think of this approach, congenial to law professors, is to think of the judges and the judiciaries as if they were clients and not subjects.

A skeptic might correctly point out that our goals here cut against the dominant paradigm in empirical research across a range of disciplines, in which social scientists study phenomena from an outsider's perspective. They observe and measure and theorize about their research subjects, but do not necessarily feel any need to interact with them; and certainly not as collaborators. We are overstating, of course. Our colleagues in anthropology and sociology, especially, incorporate the subject-perspective into their research. But their work has not figured prominently in the current enthusiasm in American law schools for empirical research on courts. We hope that in the future judges' perspectives will play an increasing role in the research on courts.

There is a different reason for our push toward increased collaboration between researchers and their subjects and that has to do with our goals. These goals are at least partially normative. We hope that the research can have payoffs in the near future in terms of yielding insights into how to improve the functioning of the judicial system. The three of us believe that the quality of the judicial system is important as a social and as an economic matter, and that aspects of the system can be measured and studied to help determine whether it can be improved and how. By contrast, there are others who are skeptical of the measurement project itself—arguing that no measurement is better than

partial measurement.[1] The threshold questions, then, are what should be measured and how.

II. WHAT TO MEASURE AND WHY

The questions for our workshop—what does it mean to judge well, how well do judges perform, and how can judicial institutions be arranged to promote the best possible performance?—are examples of topics that could benefit from greater intellectual cross-fertilization. Like all public officials entrusted with substantial power, judges should be subjected to critical appraisal: holding them accountable for their performance, identifying judges worthy of promotion, helping to decide who is fit to be a judge in the first place, or reforming judicial institutions to promote better judging. Judges themselves, in our experience, are interested in the question of what makes a good judge and, in many cases, would welcome research that attempted to tackle that question, particularly when the outcome of that research might be concrete suggestions for better judicial techniques or institutional arrangements. We are hardly making radical statements here; evaluative statements about judges and judging are far from rare. Indeed, we have colleagues who, although hostile toward any attempt to quantify aspects of judicial behavior, are comfortable evaluating the quality of this or that judge based on a selection of noteworthy opinions.

The challenge we confront, for which we would welcome help from judges and theorists, is in identifying evaluative standards that are widely held, firmly grounded in theory, and amenable to rigorous empirical assessment. If we were to ask observers of courts about judicial performance, we might well reveal some consensus about how well judges do in general and even about which judges stand out as particularly strong or weak. But if we were to press our respondents to explain the grounds for their judgments, we suspect that the answers would differ, with many struggling to give an explanation or even define their terms.

If we are correct that there is room for improvement in the critical evaluation of judging, the main reason is not that judges and scholars have been uninterested in the topic. It is not uncommon for judges to share their thoughts about proper judging in print,[2] and one of us can attest that many judges who do not write on the topic still contemplate it privately and with colleagues. There also have been a handful of judges who have engaged the academic debates primarily to criticize academic attempts to

---

[1] *See, e.g.*, Marin K. Levy, Kate Stith & José A. Cabranes, *The Costs of Judging Judges by the Numbers*, LEGAL WORKSHOP, (DUKE L.J., Feb. 25, 2010); William P. Marshall, *Be Careful What You Wish for: The Problems with Using Empirical Rankings to Select Supreme Court Justices*, 78 S. CAL. L. REV. 119, 134 (2004) ("[P]lacing too much emphasis on quantifiable measures alone may . . . inhibit the selection of those with the qualities most needed for a successful Supreme Court tenure.").

[2] *E.g.*, Armistead M. Dobie, *A Judge Judges Judges*, 1951 WASH. U. L.Q. 471, 474–84; Ruggero J. Aldisert et al., *What Makes a Good Appellate Judge? Four Views*, JUDGES' J., Summer 1983, at 14, 14, 16–17; Joseph P. Nadeau, *What It Means to Be a Judge*, JUDGES' J., Summer 2000, at 34, 34–35.

measure judicial behavior through empirical study.[3] Despite the apparent hostility of some judges, our impression from both reading and observation at our workshop is that the tone of their critiques is driven less by hostility to the idea that judicial behavior can be evaluated (and that there are better and worse performing judges and courts) than by the perception that academics are aiming wide of the mark in terms of conducting the type of research that might help improve the judicial system.[4]

On the academic side, there is some work directly on the question of how to evaluate judging—including Solum's (2003) theoretical exploration[5] and Cann's (2007) empirical analyses.[6] Empirical studies of judges and courts have become more common, and many of these studies implicitly adopt some view of judging. Concerns about the quality of judging are an important motivator of recent research into heuristics and biases in judging.[7] And even if they often go unexpressed, normative considerations about the legitimacy of judges' behavior underlie the question that has garnered more attention

---

[3] *E.g.*, Harry T. Edwards, Essay, *Collegiality and Decision Making on the D.C. Circuit*, 84 VA. L. REV. 1335, 1364–70 (1998); Bruce M. Selya, *Pulling from the Ranks? Remarks on the Proposed Use of an Objective Judicial Ranking System to Guide the Supreme Court Appointment Process*, 32 FLA. ST. U. L. REV. 1281, 1281–83 (2005); Laura Denvir Stith, Response, *Just Because You Can Measure Something, Does It Really Count*?, 58 DUKE L.J. 1743, 1743–45 (2009).

[4] Further, it seems that judges perceive a tone of disrespect in some of the academic work that seeks to rank judges on simple measures and reveal the secret "political" agendas of judges. David F. Levi & Mitu Gulati, *"Only Connect": Toward a Unified Measurement Project*, 58 DUKE L.J. 1181, 1183 (2009) ("Judges . . . resent what they see as the obsession of some empiricists with proving that judges determine case outcomes based on their judicial philosophies, which the political scientists insist on calling 'political bias.'"); Ernest A. Young & Erin C. Blondel, Response, *Does the Supreme Court Follow the Economic Returns? A Response to a Macrotheory of the Court*, 58 DUKE L.J. 1759, 1782 (2009) ("[M]any empiricists . . . seem to default to less plausible explanations for judicial behavior—for example, that judges are voting their political viewpoints or trying to affect the economy. These conclusions seem . . . inaccurate—even offensive—to judges.").

[5] Lawrence Solum, *Virtue Jurisprudence: A Virtue-Centered Theory of Judging*, 34 METAPHILOSOPHY 178, 198–99 (2003) ("[J]udicial virtues include . . . temperance, courage, good temper, intelligence, wisdom, and justice. . . . Judges ought to be selected on the basis of their possession of . . . the judicial virtues." (footnote omitted)).

[6] Damon Cann, *Beyond Accountability and Independence: Judicial Selection and State Court Performance*, 90 JUDICATURE 226, 229 (2007) (basing an empirical study of merit selection efficacy on a survey of 2,428 state court judges who chose "'making impartial decisions,' 'ensuring fairness under law,' 'defending constitutional rights and freedoms,' and 'providing equal justice for rich and poor'" as the "most important" judicial duties).

[7] *E.g.*, Chris Guthrie, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Blinking on the Bench: How Judges Decide Cases*, 93 CORNELL L. REV. 1, 5 (2007).

from students of judicial behavior than any other: the extent to which judges' personal policy preferences or moral views trump impartial interpretations of legal materials in determining their decisions. Outside of empirical studies, one may see the same implicit evaluation issues in certain theoretical work, such as in the literature on constitutional interpretation.

Lack of attention, then, is not a major obstacle to progress in the study of judicial performance. In our view, a far more important obstacle is the dearth of intellectual engagement among judges, theorists, and empiricists. The result is empirical work that is often too far removed from the core concerns of theorists and judges to reward their attention and theoretical work that is typically too abstract to lend itself to empirical testing.

Research into ideological voting illustrates this problem. Empirical scholars have amassed mountains of evidence suggesting that ideology plays an important role in judicial decisions, especially at the United States Supreme Court. But this evidence seems to have had only a limited impact on the way most theorists and judges think. Empiricists are often frustrated by what seems like a stubborn refusal to confront the implications of their findings, but there may be more to the reactions than obstinacy. For example, it may be that the distinction between the "legal" and "attitudinal" models does not capture all, or even a large part, of what is important for the legitimacy of judicial decisions.

If we are right to claim that there is a problem, what can be done about it? At a general level, the crucial step is for judges, theorists, and empiricists to engage in structured conversations. Workshops like ours can help foster such conversations, and we hope to hold more of them. Larger conferences, such as those sponsored by the Society for Empirical Legal Studies, may also serve this purpose. In the end, however, there is no substitute for reading each other's writings. There have been signs of cross-disciplinary awareness in recent years. For instance, Judge Posner, who has always engaged the social science literature, is especially attentive to it in his latest book on judging;[8] and two recent books by theorists include extensive discussions of social science research.[9] Yet these authors are in a small minority. On the other side, many empiricists care about theoretical issues; in fact, as noted, the much-maligned attitudinal versus legal model debate is, at bottom, about the legitimacy of judges' behavior and self-presentation. Still, caring about theoretical issues is not quite the same as reading theorists and judges closely and designing studies specifically to test their ideas or address their concerns.

Of course, writers cannot place all the blame on readers. Empiricists might boost readership among judges and theorists by: a) explaining their methods and results in ways that are clear and unintimidating even to those without much training in empirical research or statistics; b) avoiding resting their analyses on assumptions that strike others

---

[8] RICHARD A. POSNER, HOW JUDGES THINK (2008).

[9] MICHAEL J. GERHARDT, THE POWER OF PRECEDENT (2008); BRIAN Z. TAMANAHA, BEYOND THE FORMALIST-REALIST DIVIDE: THE ROLE OF POLITICS IN JUDGING (2008).

as too unrealistic to take seriously; c) making the theoretical and practical implications of their research more explicit; and d) increasing their understanding of the law or legal framework so as to avoid making inaccurate statements or assertions. Perhaps most beneficial would be a greater focus in the first place on questions that judges and theorists could be expected to care about. For instance, in choosing criteria for evaluating judges or judicial institutions, they could pay close attention to the normative arguments of theorists and the practicalities of real life judging, the latter with an eye toward what we can reasonably demand of human judges or what they can reasonably hope to achieve.

Like empiricists, theorists and judges are more likely to attract readers outside their own circles to the extent they refrain from insularity, eschew jargon, and avoid assumptions of knowledge or beliefs not shared by those outside the circle—admittedly, easier said than done. Most importantly, in thinking about their own work, whether academic or on the bench, theorists and judges could try to be more aware of when that work raises questions about the empirical world or rests on assumptions about the empirical world that are questionable. Identifying such questions could make it more obvious to empiricists why they should read what theorists and judges write.

The benefits of having theorists and judges suggest topics for empirical analysis would not end there. Many empiricists would likely find studying the suggested topics intellectually rewarding. Their work would, in turn, be read by more judges and theorists. The result, we suspect, would be a virtuous circle, with ever-increasing engagement among the different groups.

III. THE NEXT STAGE

Our workshop experiences and impressions from reading tell us that suggesting questions for empirical analysis does not come easily to judges and theorists, perhaps because of reticence, skepticism, or certain habits of thinking. And empiricists are not especially inclined to listen to either theorists or research subjects about what they should be studying and how. That said, despite some apparent distrust or misunderstanding at the initial stages, there was ultimately a high level of intellectual engagement at the workshop. Whatever the causes for the initial difficulties in getting the conversations going, we hope theorists and judges will push to play a larger role in setting the empirical research agenda, whether through calls for action or, if they wish to be more directly involved, through active collaboration with empirical researchers.

We end with four sets of more specific suggestions (or pleas) to different combinations of key players. The first, to academics—both theoretical and empirical—is to consider spreading their attention more evenly across a broad range of courts and judicial behaviors. The law touches people’s lives far more often and directly through state trial courts than through federal appellate courts. And then there are the local courts tackling small claims, traffic violations, and family matters; the administrative law tribunals; the international law courts; and similar court systems. All of these settings potentially provide rich sources of insight into the workings of legal institutions. Some of

these settings have been examined by researchers, but these examinations are relatively rare and are frequently ignored in mainstream discussions of judges and courts.

As important as decisions on the merits of cases are, it is just as important for us to understand how judges gather information, evaluate evidence, interpret precedents, rule on motions, choose language for their opinions, and so on. Further, whether in the criminal or civil justice systems, most parties' experience of the courts is not the traditional trial or a series of opinions culminating in the Supreme Court of the jurisdiction. Rather, it is a settlement system, through plea bargain or negotiation. But these truths, although reflected in many individual studies, are not well reflected in the literature as a whole, especially in political science and legal theory. Of course, we are not advocating that scholars stop paying attention to the U.S. Supreme Court and federal appellate decisions. Those who wish to reach judges and produce research with wider application to the world outside of academia, however, might achieve more success by focusing more on the issues of most concern to the typical judge and the typical litigant on a typical day.

The second suggestion, to empiricists alone, is to consider embracing greater methodological flexibility. Both theorists and judges at our workshop seemed impatient with what they viewed as empirical researchers' insistence on quantification, usually in the context of large-sample studies. Their criticism is overstated, given the large number of empirical researchers who employ qualitative techniques. Nevertheless, it has some validity both for the literature as a whole and for the emerging branch of that literature in legal journals. By no means do we think it would be appropriate for empiricists to weaken their standards in a way that would allow conclusions to be drawn from data that do not adequately support them. But, as long as they explicitly recognize limitations in their data, it seems to us that it may be worthwhile to sacrifice some reliability[10] and precision if it allows them to get at things that really matter.

Our third plea is to theorists and judges. They were no more shy about expressing criticisms of empirical work at our workshop than they have been in print. But their criticisms are seldom as constructive as they might be. It is of some help to an empirical researcher to hear why a particular method of measuring a key concept is flawed; it is far better to receive suggestions for improving the method. Is there any way of assessing the

---

[10] To illustrate, Professors Gulati and Klein have collaborated on research employing types of citation counts to measure aspects of judicial reputation and performance. *E.g.*, Stephen J. Choi, Mitu Gulati & Eric A. Posner, *Judicial Evaluations and Information Forcing: Ranking State High Courts and Judges*, 58 DUKE L.J. 1313 (2009); David E. Klein & Darby Morrisroe, *Prestige and Influence on the U.S. Courts of Appeals*, 29 J. LEGAL STUD. 271 (1998). Because the computation of these measures does not require independent judgment, they are highly reliable. On the other hand, although we believe that the measures are also valid, we readily concede that they only partially capture the phenomena of interest and could usefully be supplemented by measures that approach the phenomena from other angles, even if dependent on greater coder judgment and so more susceptible to reliability problems.

concept, even if imperfect, that would yield useful information? If not, is there a similar concept that could be empirically observed, allowing at least some light to be shed on the question? The key here, we think, is patience—for theorists and judges to recognize that a methodological difficulty is not necessarily an impossibility and, instead of dismissing the problem, to contribute their insights in an attempt to solve it.

Finally, a request of judges. One of the most important things they could do to promote empirical scholarship that is significant and that matters to them is to actively embrace the spirit of scholarly inquiry. No one much enjoys being the focus of critical scrutiny, especially when being evaluated by measures that seem crude. (Consider how academics regard student evaluations of their teaching or their dean's annual determination of whether they have "contributed" or not). But to the extent judges can overcome discomfort or resentment, cooperate with researchers' efforts to study them, and suggest ways for researchers to improve their studies, they can significantly contribute to the research enterprise. And there is no reason why judges should only be subjects of research. They can also engage in research informally or formally, whether keeping their eyes open for how things are done in other courts and comparing those methods with their own, engaging in experimentation to test the effectiveness of different practices or institutions, or even conducting full-scale studies and publishing the results. We recognize that in the current political environment there are groups and persons who seek to damage the judiciary in general and individual judges in particular. From our point of view, this is lamentable. But these malevolent forces and special interests will gather and publicize their own flawed data and empirical studies. We ask the judges to consider that more and better empirical study of judging and judicial institutions has the potential to lead to a stronger judiciary and to better judging. It is also an antidote to slanted and partisan attacks disguised as objective studies.

At the end of the day, we realize we are asking for a lot and that others might not be as optimistic regarding the value of collaboration among judges, theorists, and empiricists. What we saw at the workshop itself was a great deal of openness and willingness to engage. Given what we saw, we are certainly willing to do whatever we can to keep the conversations going.

ACKNOWLEDGEMENTS

Mitu Gulati is a Professor of Law at the Duke University School of Law.
David E. Klein is an Associate Professor of Politics at the University of Virginia.
David F. Levi is the Dean and a Professor of Law at the Duke University School of Law. Previously, Dean Levi was the Chief Judge of the United States District Court for the Eastern District of California.
The workshop was made possible by a grant from the National Science Foundation.