Yale Law School
Yale Law School Legal Scholarship Repository

Faculty Scholarship Series | Yale Law School Faculty Scholarship

1-1-2010

The Costs of Judging Judges by the Numbers

Kate Stith
Yale Law School

Marin K. Levy
Duke Law School

Jose A. Cabranes
United States Court of Appeals for the Second Circuit

Follow this and additional works at: http://digitalcommons.law.yale.edu/fss_papers

Part of the Law Commons

Recommended Citation
Stith, Kate; Levy, Marin K.; and Cabranes, Jose A., "The Costs of Judging Judges by the Numbers" (2010). *Faculty Scholarship Series.* Paper 1289.
http://digitalcommons.law.yale.edu/fss_papers/1289

This Article is brought to you for free and open access by the Yale Law School Faculty Scholarship at Yale Law School Legal Scholarship Repository. It has been accepted for inclusion in Faculty Scholarship Series by an authorized administrator of Yale Law School Legal Scholarship Repository. For more information, please contact julian.aiken@yale.edu.

YALE LAW & POLICY REVIEW

The Costs of Judging Judges by the Numbers

Marin K. Levy, Kate Stith,** and José A. Cabranes****

INTRODUCTION

What is to be gained by using empirical evidence to rank or "judge" judges? Such empirical studies claim two major benefits. First, because the criteria are ostensibly apolitical, the resulting rankings should identify the "best" judges across the political spectrum and thereby improve, for instance, the Supreme Court nomination process.[1] Second, because the criteria are "objective"[2] rather than "subjective,"[3] the possibility of unconscious bias is reduced greatly and these studies may at last yield answers to two difficult questions: (1) whether female judges are better than male judges;[4] and (2) whether appointed judges are better than their elected counterparts.[5] These claims in turn rest on two

* Lecturing Fellow, Duke University School of Law.

** Lafayette S. Foster Professor of Law, Yale Law School.

*** Judge, United States Court of Appeals for the Second Circuit.

The authors thank participants in the lively and enlightening workshop on quantitative analysis of judging held at Duke Law School in September 2009.

1. Stephen J. Choi & G. Mitu Gulati, *Choosing the Next Supreme Court Justice: An Empirical Ranking of Judge Performance*, 78 S. CAL. L. REV. 23, 25-32 (2004) [hereinafter Choi & Gulati, *Choosing*]; Stephen Choi & Mitu Gulati, *A Tournament of Judges?*, 92 CAL. L. REV. 299, 300-04 (2004) [hereinafter Choi & Gulati, *Tournament*].

2. *See* Choi & Gulati, *Choosing*, *supra* note 1, at 29.

3. *See id.* at 36.

4. *See* Stephen J. Choi et al., *Judging Women* 1 (Univ. of Chi. Law & Econ., Olin Working Paper No. 483, 2010), *available at* http://ssrn.com/abstract=1479724.

5. *See* Stephen J. Choi, G. Mitu Gulati & Eric A. Posner, *Professionals or Politicians: The Uncertain Empirical Case for an Elected Rather than Appointed Judiciary* 2 (Univ. of Chi. Law & Econ., Olin Working Paper No. 357, 2007), *available at* http://ssrn.com/abstract=1008989 ("To test the conventional wisdom that appointed judges are better than elected judges, we use a tripartite definition of judicial quality—productivity, opinion-quality, and independence.").

HeinOnline -- 28 Yale L. & Pol'y Rev. 313 2009-2010

assumptions: that the legal empiricists have selected the right qualities to measure and that their methodology for doing so is accurate.

In the years since the first empirical rankings were published, both judges and scholars have cast grave doubt on the accuracy of these assumptions,[6] and we seek to contribute to this discussion.[7] We believe that there is now a general consensus that (1) the judicial virtues the legal empiricists set out to measure probably have little bearing on what actually makes for a good judge; and (2) even if they did, the empiricists' chosen variables have not measured those virtues accurately. As matters now stand, the benefits professed by the empiricists have not been achieved by their models.[8]

The failure of empirical rankings to deliver the benefits promised by their proponents is enough to raise serious doubts about whether the enterprise is worthwhile. But, even more importantly, these rankings stand to create—and indeed, already may have created—several significant harms. By generating and then publicizing unreliable claims about the relative quality of judges, these studies mislead both decision-makers and the public, degrade discussion of judging, and could, if taken seriously, perniciously alter the behavior of judges themselves. Far from improving the process for selecting judges and justices, reliance on such empirical rankings would distract all involved from serious inquiry and debate about the nature of judging, the process by which judges should be selected, and contested visions of justice itself.

In this Essay, we focus on the presently dominant models of judging judges advanced by Professors Stephen Choi, Mitu Gulati, and Eric Posner, among

6. *See, e.g.*, Scott Baker, Adam Feibelman & William P. Marshall, *The Continuing Search for a Meaningful Model of Judicial Rankings and Why It (Unfortunately) Matters*, 58 DUKE L.J. 1645 (2009); Jay S. Bybee & Thomas J. Miles, *Judging the Tournament*, 32 FLA. ST. U. L. REV. 1055 (2005); 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 (2005); Patrick S. Shin, *Judging Merit*, 78 S. CAL. L. REV. 137 (2004); Laura Denvir Stith, *Just Because You Can Measure Something, Does It Really Count?*, 58 DUKE L.J. 1743 (2009).

7. The authors include a judge who was one of the seventy-four federal appellate judges that Professors Choi and Gulati ranked in their article, Choi & Gulati, *Choosing*, *supra* note 1. This judge learned in conversation at a Duke Law School workshop in September 2009, that he was listed near the bottom quartile on the "independence" ranking, *see id.* at 114 tbl.H, a fact that he finds amusing and not meaningful.

8. We do appreciate, however, that other common measures of judicial quality are also wanting—in particular the ubiquitous reliance on a judge's "reputation" on the bench and within the bar. To the extent that Choi and his coauthors are motivated by the desire for more reliable gauges, we are sympathetic to this motivation.

HeinOnline -- 28 Yale L. & Pol'y Rev. 314 2009-2010

others ("Choi et al.").[9] We explain in Part I why these models are deeply flawed—revisiting arguments made by others and positing new ones—and then discuss in Part II why there is little hope for meaningful improvement. In Part III, we discuss the harms of relying on imperfect models and explain why the current enterprise of "judging judges" has significant costs.

I. Flaws in the Current Models

Choi et al. posit three "objective" measures of judicial merit: (1) productivity; (2) opinion quality (or what they sometimes call "influence"); and (3) independence. Although Choi and his coauthors have revised their models from time to time, they consistently have professed to be able to measure these qualities from datasets of appellate judicial opinions[10] by looking to, respectively: (1) how many signed opinions a judge publishes;[11] (2) how many times the judge's opinions have been cited in other jurisdictions;[12] and (3) how many times the judge has dissented in cases in which the majority opinion is written by a member of the same political party.[13] The authors then use their measures to rank judges and also to determine whether female judges are "better" than male judges, and whether appointed judges are "better" than

9. The first few articles to advance a model to "rank" judges were authored by Stephen Choi and Mitu Gulati. Eric Posner and others subsequently collaborated with them.

10. The empiricists do not apply their model to trial judges, who make innumerable decisions from the bench as well as in written opinions. Some commentators have suggested that Professor Choi and his colleagues would fare better by studying trial judges instead of appellate judges. *See* Ahmed E. Taha, *Information and the Selection of Judges: A Comment on "A Tournament of Judges,"* 32 Fla. St. U. L. Rev. 1401, 1412 (2005) ("Because federal district judges generally make decisions alone, their opinions are less likely to be affected by the preferences of their colleagues. In addition, they have almost complete discretion over choices such as whether to publish an opinion in a particular case. Thus quantitative measures of judicial performance will be measures only of a particular judge's behavior rather than also of the influence of other judges on a panel." (footnotes omitted)). Professor Mitu Gulati and Duke Law School Dean David Levi (who is a former United States District Judge), have argued that because "the literature on judicial measurement has largely ignored the question of how best to measure trial court performance," it would be beneficial for empiricists to study trial judges. David F. Levi & Mitu Gulati, *Judging Measures*, 77 UMKC L. Rev. 381, 403-04 (2008). Gulati and Levi note, however, that empiricists would need to develop different measures to do so because, for example, publication rates are less meaningful for trial judges than for appellate judges. *Id.* at 404-07.

11. Choi & Gulati, *Choosing*, *supra* note 1, at 42-43.

12. *Id.* at 49-50.

13. *Id.* at 58-59.

HeinOnline -- 28 Yale L. & Pol'y Rev. 315 2009-2010

elected judges. As noted above, several commentators have identified serious flaws in this general model.[14]

First, the judicial virtues selected by Choi et al. do not adequately capture the qualities good judges possess and the activities in which they engage. Professors Scott Baker, Adam Feibelman, and William P. Marshall point out that "because of data limitations, [Choi et al.'s] study ignores aspects of judicial behavior that are arguably more important than the ones proxied, such as integrity, fairness, open-mindedness, thoroughness, and temperament."[15] As Chief Justice Laura Denvir Stith of the Supreme Court of Missouri has explained: "No matter how much one breaks down the criteria used as measures, those criteria still miss the essence of judging."[16] Moreover, a model with only the three aforementioned virtues fails to account for the other important work that appellate judges do in addition to opinion writing, "including judicial administration and lower court supervision, serving on government commissions, teaching at law schools, writing scholarly articles, and training other judges,"[17] and, we would add, expending significant time editing or commenting on other opinions that will be published under the name of a colleague. All of this work is both critical for the functioning of courts and a significant part of what "good" judges do.

Second, even if we could agree that Choi et al. had selected judicial qualities that are possessed by the "best" judges, they have failed to measure those qualities accurately. Beginning with productivity, it seems obvious beyond peradventure that a judge may be productive by doing things beyond publishing her own signed opinions. In many courts, judges issue far more unpublished opinions than published ones. Furthermore, the measure fails to account for another common and important form of judicial production: the per curiam opinion (which is published without the identification of the author). Although practices with respect to per curiam opinions differ among state courts and among federal circuit courts, in many jurisdictions they constitute a significant portion of all published opinions. And although per curiam opinions tend, on average, to be shorter than their signed counterparts, they still consume much of the time and energy that signed published opinions require. Because the datasets used by Choi et al. do not include per curiam opinions, and because the writing of these opinions is not distributed evenly within a given court, the statistical analysis undertaken by Choi et al. is incomplete and quite possibly misleading. Indeed, judges in the state dataset

14. *See* sources cited *supra* note 6.

15. Baker et al., *supra* note 6, at 1657.

16. Stith, *supra* note 6, at 1744.

17. Baker et al., *supra* note 6, at 1657. Judges also serve on rules committees and engage in activities of the bar more generally. *Id.* at 1659.

HeinOnline -- 28 Yale L. & Pol'y Rev. 316 2009-2010

published on average only about twenty-two signed opinions per year;[18] issuing even a few per curiam opinions more than the average should significantly affect the judge's "productivity" ranking.

The point is not simply that per curiam opinions (along with unpublished opinions) are a major category of written judicial production that Choi et al. have failed to consider in their methodology. The more critical point is that their methodology will *never* be able to account for these opinions. Because, as their designation suggests, per curiam opinions are the opinions "by the court," they are not signed by the principal drafter (and, similarly, unpublished opinions are unsigned). In short, there is no way to know which judge should receive credit for the effort of writing these opinions. This important category of data is simply missing, and it cannot be discovered by Choi et al. or anyone else.

"Opinion quality," too, is an ambiguous and potentially broad-ranging concept. Yet, all that Choi et al. measure is the number of citations that a judge's opinions receive in published opinions from other jurisdictions. Citation counts need not, of course, reflect quality. Even as a rough measure of one type of judicial *influence*, the chosen proxy is wanting. There are many factors that contribute to opinion citation counts that have nothing to do with the insight or persuasiveness of the opinion and nothing to do with the ongoing influence of the opinion writer.[19] If a novel issue arises in one state court that has been addressed in only a few other states, the opinions from those other jurisdictions are likely to be cited whether or not they are persuasive or otherwise of high quality. There is a simple, temporal component to citations as well. Unless there is a seminal case for a particular proposition, many judges prefer to cite recent cases articulating the governing standard or principle. Accordingly, there may frequently be situations in which Judge *A* publishes an opinion enunciating a particular rule or standard, and a few years later, Judge *B* follows and restates this very same principle of law. Simply by virtue of being more recent, the later opinion may garner more out-of-state citations going forward.[20]

Moreover, using citation counts as a measure of opinion quality appears to introduce a bias in favor of those decisions (and those courts) that reach out to resolve issues left to legislative determination by courts in other jurisdictions.[21]

---

18. Stephen J. Choi, Mitu Gulati & Eric A. Posner, *Judicial Evaluations and Information Forcing: Ranking State High Courts and Their Judges*, 58 DUKE L.J. 1313, 1320 (2009).

19. *See, e.g.*, Frank B. Cross & Stefanie Lindquist, *Judging the Judges*, 58 DUKE L.J. 1383, 1391-93 (2009) (discussing criticisms of citation analysis).

20. *See* Lawrence B. Solum, *A Tournament of Virtue*, 32 FLA. ST. U. L. REV. 1365, 1392 (2005) ("Subsequent opinions will be more likely to be cited if they are the most recent opinion stating the proposition . . . .").

21. Relatedly, as Chief Justice Stith noted: "[T]he number of out-of-state citations largely measures a characteristic other than the intrinsic worth of the opinions; it measures which state courts take the lead in addressing new and developing areas

HeinOnline -- 28 Yale L. & Pol'y Rev. 317 2009-2010

Once one court identifies or articulates a new right or rule of law, litigants in other states or circuits will urge their courts to quite literally follow suit—resulting perhaps in opinions that cite the novel judicial proposition even if that proposition is thoroughly rejected. This situation reflects neither judicial influence nor opinion quality, much less "good" judicial behavior that should be rewarded in judicial rankings.

Finally, there are serious flaws in the way that Choi et al. measure "independence." Although we readily agree that independence—in the sense of not being influenced by certain personal, political, or other inappropriate considerations—is a valuable attribute in a judge, that quality should not be confused with what the model actually measures, which is a disposition to dissent. Dissenting frequently, regardless of the political affiliation of other panel members, might well be due to characteristics associated with "bad" judging.[22] As Professor Baker and his coauthors have noted, "independence defined in this way may actually measure a lack of quality in judging, including deficiencies in collegiality and leadership, and a propensity for judicial activism outside the ideological mainstream."[23]

We believe this problem deserves even greater emphasis. There is an obvious characteristic of good judges that is in tension with the measure of independence chosen by Choi et al.: the ability to build a coalition or achieve consensus. Nowhere does their analysis credit the judge who is skilled in convincing other judges to join his or her opinions. Perhaps such a judge writes particularly sound opinions and therefore has an easier time convincing other judges to sign on with her. Or perhaps the judge is simply adept at finding compromises or alternative routes to reaching a decision that command greater

---

of the law." Stith, *supra* note 6, at 1749. Even if one believes that addressing developing issues is a component of judicial *influence*, it is not, for the reasons that Chief Justice Stith and we have noted, necessarily related to the *quality* of the opinion in which the new issues are addressed.

22. *See* Selya, *supra* note 6, at 1290 ("[T]here plainly is a point at which dissenters cross the line from enriching thought into either intellectual preening or obstructionist polemicism.").

23. Baker et al., *supra* note 6, at 1661; *cf.* José A. Cabranes, Some Brief Reflections on Judging in the Federal Courts 9-10 (Nov. 10, 2006) (unpublished manuscript on file with the Yale Law & Policy Review) ("[J]udicial activism *at the trial level* is not necessarily an ideological question or a political question or a jurisprudential question—it is, rather, a *psychological* question; whether a judge is 'active' or 'passive' is often simply a reflection of the judge's psychological disposition rather than his political worldview. The most important 'ism' in a trial judge [isn't liberalism or conservatism but] . . . *metabolism*. . . . '[J]udicial activism' at the *trial* level often may mean little more than that a particular judge is assertive and demanding—even overbearing—in the assertion of his will and his authority. In some such circumstances, the assertion of judicial power may as readily be a function of an individual jurist's disposition to command and to take charge . . . .").

HeinOnline -- 28 Yale L. & Pol'y Rev. 318 2009-2010

agreement. Yet, such a judge will dissent less frequently, regardless of the political affiliation of the other panel members.[24]

Here again, the fundamental flaw with the approach of Choi et al. is that there is no clear way to solve the measurement problem. To credit dissents as evidence of independence necessarily discredits coalition-building as relevant to judicial quality. Without additional information, there is no way of knowing whether judges who frequently dissent are independent, are poor at persuasion and coalition building, or simply have idiosyncratic views.

II. Why the Current Models Cannot Be Improved

In response to these objections, the empiricists' natural reaction might well be to continue adjusting their current proxy variables or adding new variables in subsequent studies. They may hope that by refining their model, they can eventually fix whatever deficiencies currently exist and improve the enterprise of judging judges. We hope that this is not the direction taken.

The overarching problem is not that the methodology is flawed (though it is), but that the enterprise itself is misguided. Choi et al. cannot simply tweak their model to incorporate many of the concerns we and others have noted, because the data inevitably conflate countervailing qualities. Given the complexity of judging, it is not surprising that a single data point (e.g., a published opinion, a citation, or a dissent) may be both consistent and inconsistent with admirable characteristics. Most importantly, the very meaning of "good" judging is contested in a heterogeneous, pluralistic society. For some opinion elites (including, apparently, Choi et al.),[25] the Supreme Court nomination process would be improved if it were less partisan and if it were less dependent on a nominee's views on politically charged issues. For others, this development would be worrisome, denying what is ultimately the political character of both presidential nomination and senatorial consent.[26] Similarly,

24. In any event, myriad factors not subject to quantitative analysis may prompt a judge to avoid writing a dissent even when the judge has doubts about the soundness of a majority opinion. Such factors may include a desire to avoid solidifying the law in favor of the majority's position resulting from a dissenting opinion that loudly proclaims *what the law is not*; that is, a judge may prefer to wait for another case in which she might prevail.

25. *See* Choi & Gulati, *Choosing*, *supra* note 1, at 82 ("[T]he goal for our tournament was to provide an improvement over the current system. The harder question is whether our measures could play a role in a more bipartisan, but subjective, selection process of the type that some would claim we have had in the past. Our view is that they should; objective measures will serve as a check on the inevitable biases that any system of subjective analysis will possess (and vice versa).").

26. *See* Laurence H. Tribe, God Save This Honorable Court: How the Choice of Supreme Court Justices Shapes Our History 106-10 (1985) (arguing that the President and the Senate should consider a nominee's political leanings to create the best-functioning Court); David S. Law, *Appointing Federal Judges: The President, the Senate, and the Prisoner's Dilemma*, 26 Cardozo L. Rev. 479, 500

HeinOnline -- 28 Yale L. & Pol'y Rev. 319 2009-2010

analyses purporting to ascertain whether female judges are better than male judges and whether appointed judges are better than elected ones founder not just on data and methodological limitations, but also (and unavoidably) on disputed understandings of what it means to be "better."

III. The Harms of Relying on Imperfect Models

While the project of Choi et al. is unsustainable on its own terms, its false promise—that the quality of a judge (indeed, of judging) can be captured by a composite ranking—threatens to do real damage to the judicial system and public understanding and perception of judges. First, far from improving the Supreme Court nomination process, increased reliance on anything like these rankings could arbitrarily derail or encourage particular nominations. Consider the appellate judge who writes more than her share of per curiam or unpublished opinions, who hesitates to propound novel legal theories, and who seldom dissents because she knows how to build a majority. This judge will rank toward the bottom of all three measures employed by Choi et al.: signed opinions, citations by other courts, and dissents from others appointed by presidents of her party. Is there now a presumption—one that is irrebuttable as a practical matter[27]—that she is unqualified? If the available empirical models are treated as actually indicating something important, will we lose such judges, either because the media will be unduly influenced by this seemingly "expert" and "objective" ranking or because political actors will make the calculation, in part on the basis of media reports, simply not to nominate them?

The delusive allure of the empirical judicial ranking model already has distorted critical debates about the judiciary. Recently, several articles in the popular media have reported the findings of Choi et al. as definitive conclusions. In May 2009, the *New York Times* published an online piece entitled, "Assessing Sotomayor's Influence," which purported to reveal how influential Justice Sonia Sotomayor had been as a judge based on how many

(2005) ("It is, arguably, both necessary and appropriate to examine closely the ideological leanings of those seeking lifetime appointment to high federal office.") (footnote omitted); Press Release, Senator Charles E. Schumer, 'Modesty and Stability' on the Bench: How John Roberts Can Convince Democrats To Vote for Him (July 27, 2005), *available at* http://schumer.senate.gov/new_website/record.cfm?id=260552 ("I think we've won the argument that judicial philosophy and ideology are important and proper considerations in confirming a nominee.... Senators from across the political spectrum—from Brownback to Specter to Kennedy—have emphasized the importance of knowing a candidate's views before voting to confirm.").

27. *See* Baker et al., *supra* note 6, at 1666 ("The question is whether the information presented in response to the rankings will actually be able to rebut the presumptions they create. We do not believe they will."); Bybee & Miles, *supra* note 6, at 1073 ("The ranking of judges creates a new set of data easily grasped by the public. A President who nominates number forty-two on the list will have to explain himself.").

HeinOnline -- 28 Yale L. & Pol'y Rev. 320 2009-2010

citations her opinions had received.[28] The piece did not mention any criticisms of the use of citation counts as a measure of either quality or influence.

Indeed, it is instructive to examine how Judge Sotomayor and Judge Samuel Alito, both nominated to the Supreme Court after the first article by Professors Choi and Gulati was published, have fared in the wake of this scholarship. Judge Alito, sitting on the Third Circuit from January 1, 1998 through December 31, 2001 (the time period for the study), ranked sixteenth out of the seventy-four judges. Considering the company, perhaps that is "not too shabby," though it does not sound as good when stated slightly differently—that he was in the seventy-eighth percentile of federal appellate judges.[29] In any event, the media did not trumpet his overall ranking, but rather, that he was "among the most independent judges."[30] Like all numbers, the Choi et al. rankings can be spun.

In fact, the numbers have been spun not only by the media but also by one of Professor Choi's coauthors. Judge Sotomayor was not included in the Choi and Gulati ranking of federal judges because she was not appointed to the court of appeals until mid-1998.[31] Shortly after Judge Sotomayor was nominated to the Supreme Court in 2009, Professor Eric Posner published an online analysis that sought to extrapolate from Judge Sotomayor's 1999-2001 data where she would have placed in the original ranking. Posner concluded:

> The bottom line is that Judge Sotomayor is about average, or maybe a bit below average, for a federal appellate judge. These results are far from conclusive, but one might think that put the burden on Judge Sotomayor's defenders to come forward with stronger reasons for her nomination than they have so far.[32]

---

28. *Assessing Sotomayor's Influence*, N.Y. TIMES.COM, May 28, 2009, http://www.nytimes.com/interactive/2009/05/28/us/politics/0529-judge-graphic.html. This multimedia presentation was linked to a print article, Jo Becker & Adam Liptak, *Assertive Style Raises Questions on Demeanor*, N.Y. TIMES, May 29, 2009, at A14.

29. *See* Choi & Gulati, *Choosing*, *supra* note 1, at 113 tbl.H.

30. *See, e.g.*, *Study Ranks Alito Among Most Independent Judges*, USA TODAY.COM, Nov. 22, 2005, http://www.usatoday.com/news/washington/judicial/2005-11-22-alitoresearchers_x.htm.

31. *See Judge Sonia Sotomayor: What the Data Show*, Posting of Eric Posner to The Volokh Conspiracy, http://volokh.com/2009/05/13/judge-sonia-sotomayor-what-the-data-show/ (May 13, 2009, 11:40 EDT) ("Unfortunately, Choi and Gulati excluded Judge Sotomayor from the data set because she was appointed in 1998 and thus does not have complete data for that year.").

32. *Id.*

HeinOnline -- 28 Yale L. & Pol'y Rev. 321 2009-2010

Not surprisingly, the defenders did come forth. Two weeks later, Professor Posner announced a different conclusion: "Sotomayor looks good."[33] Working with Professors Gulati and Choi, he examined Judge Sotomayor's opinions between 2004 and 2006, and compared her not to all other federal judges but to a subgroup consisting of "other court of appeals judges who were rumored to be on President Obama's shortlist, and with a kind of control group consisting of court of appeals judges rumored to be on President Bush's shortlist in 2005."[34] Compared to this group, she ranked high both in total number of citation counts and in citations per opinion (the latter being a measure discussed, but not employed, in the rankings previously published by Choi and Gulati).[35] Posner referred to the total citation-count data as "a measure of general influence."[36] He noted that the use of citation counts had been criticized but addressed that concern only with the non sequitur: "Still, the numbers seem pretty robust, and the comparison here is with well-respected judges, not with ordinary judges."[37] Of course, if the measure does not make sense, it is not clear how the numbers can be "robust." Moreover, the second blog posting seemed to undercut the core of the larger empirical project, for it both downplayed the other two measures—productivity and independence[38]—and reverted to subjective general reputation ("not ordinary") as a measure of quality.

Even though commentators have powerfully criticized their model, the authors cannot seem to resist trumpeting its claims—without acknowledging the model's severe limitations. In October 2009, Professor Choi and his coauthors published an article in *Slate*, entitled "Do Women Make Better Judges? Asked and Answered—With Data,"[39] which the *New York Times* picked

33. *Judge Sotomayor: More Data, and a New Conclusion*, Posting of Eric Posner to The Volokh Conspiracy, http://volokh.com/posts/1243482653.shtml (May 27, 2009, 23:50 EDT).

34. *Id.*

35. *See* Choi & Gulati, *Choosing*, *supra* note 1, at 54.

36. Posner, *supra* note 33.

37. *Id.*

38. Judge Sotomayor did not fare as well as most of the subgroup on productivity, namely the number of signed opinions. Professor Posner quite plausibly suggested that the Second Circuit may have a norm of publishing fewer signed opinions than other circuits—an influence for which Professor Choi and his colleagues usually control in their empirical analysis, *see* Choi & Gulati, *Choosing*, *supra* note 1, at 45-46, but which apparently was not controlled for here. Posner pronounced Judge Sotomayor "about average" with respect to the number of dissenting opinions per signed opinions (a measure not included in the published ranking). Owing to a "lack of time," however, he did not undertake to measure her "independence" ranking. Posner, *supra* note 33.

39. Stephen Choi et al., *Do Women Make Better Judges? Asked and Answered—with Data*, SLATE, Oct. 2, 2009, http://www.slate.com/id/2231166/.

HeinOnline -- 28 Yale L. & Pol'y Rev. 322 2009-2010

up in a blog posting, "Women Judges Less Credentialed, but Equally Good."[40] Thus, both the media and the model's authors proactively publicize the "results" of what the authors know to be a limited or flawed model. This activity looks more like false advertising than scholarship.

Conclusion

Studying judges—both what they do and what it means for them to do their job well—is important. The common, informal practice of judging judges on the basis of general "reputation" within the bar or among other judges may be inadequate, fraught with misunderstanding, and subject to manipulation by judges or by those who comment on the work of judges, among others. But a more rigorous study of judges must be approached with an appreciation that the job of judging is complicated and multifaceted; judging judges should not be reduced to the counting of discrete data points that yield ordinal rankings. Given the limitations of available data, the model developed by Choi et al. cannot accurately measure even what it purports to measure: productivity, opinion quality, and independence. Given the nature of judging, we doubt that any such data-driven model ever can describe what constitutes "good" or "bad" judging, at least not to the satisfaction of those who take the time to study the models.

Whatever "good judging" may be, it is not to be found in how individual judges score on a law-professor-generated rating system based on crude data and a shallow understanding of the art and science of judging. The biggest cost of judging judges by these numbers would be if political decision-makers (not to mention judges themselves) alter their behavior in response to the existence of the Choi et al. rating system. In that event, the entire ranking project not only will have failed to identify good judging, but also will be counterproductive to that end.

40. Posting of Catherine Rampell to Economix, http://economix.blogs.nytimes.com/2009/10/07/women-judges-less-credentialed-but-equally-good/ (Oct. 7, 2009, 19:45 EDT).

HeinOnline -- 28 Yale L. & Pol'y Rev. 323 2009-2010

HeinOnline -- 28 Yale L. & Pol'y Rev. 324 2009-2010