

EMPIRICAL MEASURES
OF JUDICIAL PERFORMANCE:
AN INTRODUCTION TO THE SYMPOSIUM

STEVEN G. GEY* AND JIM ROSSI**

I. CONCEPTUAL CRITIQUES OF THE JUDICIAL TOURNAMENT: CAN WE MEASURE JUDGES, AND IF SO, HOW?.....	1004
II. NEW EMPIRICAL WORK ON JUDICIAL PERFORMANCE: HOW DO WE OPERATIONALIZE AND MEASURE IT?.....	1007
III. DO MEASURES OF JUDICIAL PERFORMANCE CREATE POSITIVE BEHAVIORAL AND INSTITUTIONAL INCENTIVES?.....	1011
IV. CONCLUSION	1014

In the highly charged political process of vetting, presenting, and approving federal judicial nominees, it is commonplace for Presidents, Senators, and interest groups to make claims about a nominee's merit or lack thereof. Both supporters and opponents of nominees often phrase their positions in objective terms of merit.

Independent groups such as the American Bar Association (ABA) also appeal to seemingly objective terms of merit in evaluating judicial candidates. The ABA's ratings of judicial appointees are specifically phrased in terms of the nominee's professional qualifications, and in recent years the objectivity of these ratings has frequently been the subject of controversy. In the recent dispute over the nomination of California Supreme Court Justice Janice Rogers Brown to the U.S. Court of Appeals for the Ninth Circuit, for example, no one on the fifteen-member ABA evaluation committee rated Brown "well qualified," a majority rated her merely "qualified," while some members of the committee rated her "not qualified." This resulted in the ABA committee releasing an oxymoronic rating of "qualified/not qualified."¹ Of course, the rhetoric over judicial qualifications is most impassioned in the debate over selecting U.S. Supreme Court Justices. After two members of the ABA evaluation committee rated Clarence Thomas "not qualified" to be appointed to the Supreme Court, President Bush proposed ending the organization's fifty-year role in the judicial appointments process,² thus eliminating even the

* David and Deborah Fonvielle and Donald and Janet Hinkle Professor, Florida State University College of Law.

** Harry M. Walborsky Professor and Associate Dean for Research, Florida State University College of Law.

1. Neil A. Lewis, *Battle Lines Already Forming Against a Bush Court Selection*, N.Y. TIMES, Oct. 18, 2003, at A8.

2. Neil A. Lewis & David Johnston, *Bush Would Sever Law Group's Role in Screening Judges*, N.Y. TIMES, Mar. 17, 2001, at A1.

pretense that an objective assessment of judicial quality was possible.

Claims concerning judicial nominees' professional qualifications, however, are rarely gauged against empirical evidence which might support or contradict them. In recent years, extensive data on the judiciary and individual judges has become widely available to political scientists, economists, and legal scholars, presenting a fertile opportunity for the empirical study of courts and judges. Today, the outcomes of the judicial process—judicial opinions—are widely available in searchable form through electronic databases, including Westlaw and LEXIS. More extensive common datasets for many court systems are also available to researchers.³

If you build a squash court, it does not take long for a player to find it. Indeed, given the rich availability of data, in recent years, empirical studies of judicial institutions⁴ and judicial behavior⁵ have proliferated. So have studies of judges.⁶ One of the more provocative recent studies of judicial behavior, by Stephen Choi and Mitu Gulati, builds on the rich availability of data on the performance of individual judges.⁷ Their judicial tournament, which purports to introduce calm objectivity into the study of judicial performance and judicial selection, has generated a furor of sorts. Choi and Gulati argue that

3. In 2000, *Judicature*, the official journal of the American Judicature Society, devoted several articles to the topic of the use of data on courts to understand the judicial process. The issue surveyed the emergence of a broad range of common datasets, covering a diverse range of courts including the U.S. Supreme Court, federal appellate courts, and many state courts. See Symposium, *Social Science, the Courts, and the Law*, 83 JUDICATURE 217 (2000).

4. F. Andrew Hanssen, *The Effect of Judicial Institutions on Uncertainty and the Rate of Litigation: The Election Versus Appointment of State Judges*, 28 J. LEGAL STUD. 205 (1999); Alexander Tabarrok & Eric Helland, *Court Politics: The Political Economy of Tort Awards*, 42 J.L. & ECON. 157 (1999).

5. See Orley Ashenfelter et al., *Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes*, 24 J. LEGAL STUD. 257 (1995); Frank B. Cross, *Decision-making in the U.S. Circuit Courts of Appeals*, 91 CAL. L. REV. 1457 (2003); Theodore Eisenberg & Sheri Lynn Johnson, *The Effects of Intent: Do We Know How Legal Standards Work?*, 76 CORNELL L. REV. 1151 (1991); Tracey E. George, *Developing a Positive Theory of Decisionmaking on U.S. Courts of Appeals*, 58 OHIO ST. L.J. 1635 (1998); Deborah Jones Merritt & James J. Brudney, *Stalking Secret Law: What Predicts Publication in the United States Courts of Appeals*, 54 VAND. L. REV. 71 (2001); Richard L. Revesz, *Congressional Influence on Judicial Behavior? An Empirical Examination of Challenges to Agency Action in the D.C. Circuit*, 76 N.Y.U. L. REV. 1100 (2001); Richard L. Revesz, *Environmental Regulation, Ideology, and the D.C. Circuit*, 83 VA. L. REV. 1717 (1997); Gregory C. Sisk et al., *Charting the Influences on the Judicial Mind: An Empirical Study of Judicial Reasoning*, 73 N.Y.U. L. REV. 1377 (1998).

6. See Lee Epstein et al., *The Norm of Prior Judicial Experience and Its Consequences for Career Diversity on the U.S. Supreme Court*, 91 CAL. L. REV. 903 (2003); William M. Landes et al., *Judicial Influence: A Citation Analysis of Federal Courts of Appeals Judges*, 27 J. LEGAL STUD. 271 (1998).

7. Stephen Choi & Mitu Gulati, *A Tournament of Judges?*, 92 CAL. L. REV. 299 (2004) (raising the idea of a judicial tournament).

the selection of Justices for the U.S. Supreme Court ought to be based on a tournament, in which judges who possess the most merit, as measured empirically, would be selected over their lower-ranked peers. If their tournament were seriously adopted, promotions to the U.S. Supreme Court would be based on quantitative measures of, rather than qualitative claims to, merit.

When Choi and Gulati first circulated their proposal, many legal scholars (including us) believed it a clever experiment of Swiftian proportions. Choi and Gulati are more than witty provocateurs. They are also hard-working, smart, and careful scholars. Since proposing the tournament, Choi and Gulati have collected and analyzed empirical data on federal appellate court judges, ranking every sitting appellate court judge over their test period.⁸ The federal judiciary is not yet setting salaries based on their tournament ranking, but their effort to operationalize and measure the performance of judges is serious business. Not only would we choose the next Supreme Court Justice from among their tournament's high performers if Choi and Gulati's understanding of merit were the standard, but they have done the heavy lifting for the next President to propose a nomination to the Supreme Court, actually conducting their tournament by ranking sitting appellate court judges based on measures of their productivity, influence, and independence.

Inspired by the burgeoning empirical literature on the judiciary, as well as Choi and Gulati's proposal that empirical study (including a tournament) be used in judicial selection, the editors of the *Florida State University Law Review* have given Choi and Gulati a little company on the court by soliciting some essays addressing the topic of empirical measures of judicial performance from leading scholars as well as some federal judges who, willingly or not, played in their tournament. The essays in this Symposium address empirical measures of judicial performance from a variety of methodological perspectives, but they can roughly be organized around three basic themes. First, many of the essays critique the empirical enterprise itself and especially the tournament strategy for evaluating judges, although these essays also raise important issues for future empirical study of judges. Second, many of the essays in the Symposium propose new ways of operationalizing the empirical study of judicial performance or present fresh empirical evidence about judges and courts. Third, some of the essays focus on the behavioral and institutional implications of empirical studies on judges and courts.

8. Stephen J. Choi & G. Mitu Gulati, *Choosing the Next Supreme Court Justice: An Empirical Ranking of Judge Performance*, 78 S. CAL. L. REV. 23 (2004) (applying the tournament idea to data on federal appellate court judges).

I. CONCEPTUAL CRITIQUES OF THE JUDICIAL TOURNAMENT: CAN WE MEASURE JUDGES, AND IF SO, HOW?

It is certainly not new to claim that quality in judging is incapable of empirical measurement. Judge Harry Edwards of the U.S. Court of Appeals for the D.C. Circuit is perhaps the strongest critic of efforts to quantitatively measure the activities of judges and courts. For example, in a much-cited article on attempts to use quantitative techniques to study the judiciary, he observes that quantitative studies of judicial decisions “must be viewed with great caution.”⁹ He further states, “Regression analysis does not do well in capturing the nuances of human personalities and relationships, so empirical studies on judicial decision making that rely solely on this tool are inherently flawed.”¹⁰

Many of the essays in this Symposium echo Judge Edwards’ skepticism toward quantitative studies of the judiciary. The essays add several fresh perspectives and concerns to the mix of critics of quantitative measures of judicial performance. Some of the essays raise a concern with quantitative measures of what is fundamentally qualitative, as did Judge Edwards. For instance, one concern that reverberates throughout the various criticisms of Choi and Gulati’s tournament is that it interferes with judicial independence. Other concerns relate to the imperfections of various quantitative measures of judicial performance, suggesting that we look to different criteria than Choi and Gulati’s. In addition, some reject the extension of ranking or quantification of any sort to individual judges and judicial institutions.

Brannon Denning expresses some skepticism about an empirical judicial appointments tournament by challenging some of the assumptions underlying Choi and Gulati’s proposal.¹¹ In particular, Denning questions the assumption that politics has overwhelmed the judicial selection process. Denning first argues that Choi and Gulati have not carefully defined the “politics” with which they are concerned. Denning argues that when Choi and Gulati refer to “politics,” what they really mean is “ideology.” Denning then critiques the claim that ideology should not enter the judicial appointments process. Denning suggests that the problem (if it is a problem) with the introduction of ideology into the process lies not with the President and individual Senators, but with the constituents they represent. If

9. Harry T. Edwards, *The Effects of Collegiality on Judicial Decision Making*, 151 U. PA. L. REV. 1639, 1656 (2003).

10. *Id.*; see also Harry T. Edwards, *Collegiality and Decision Making on the D.C. Circuit*, 84 VA. L. REV. 1335 (1998).

11. Brannon P. Denning, *Empirical Measures of Judicial Performance: Thoughts on Choi and Gulati’s Tournament of Judges*, 32 FLA. ST. U. L. REV. 1123 (2005).

members of the public view judging as at least in part an ideological process, Denning asks, then why should an appointee's ideology be excluded from the range of concerns discussed during the confirmation process? After posing this question, Denning moves on to praise Choi and Gulati for providing one possible model of a more sophisticated framework for critiquing nominees during the judicial selection process. Denning concludes that something like the Choi and Gulati system could improve both the process itself and the public's understanding of that process.

Steven Goldberg's essay, *Federal Judges and the Heisman Trophy*,¹² challenges Choi and Gulati's basic premise that lower federal courts are the most logical places to identify candidates with the proper qualifications to be great Supreme Court Justices. Goldberg points out that most of the individuals who have become Supreme Court Justices did not serve on lower federal courts. They came, instead, from the state courts, the executive or legislative branches of government, or private practice. Even more striking, Goldberg surveys eleven lists of highly successful Justices and finds that the Justices who served on lower federal courts fared very badly on those lists. Only one of the twenty-three Justices who appear on two or more of those eleven lists had previously served on a lower federal court. Goldberg concludes by drawing an analogy between Choi and Gulati's tournament of judges and other flawed efforts to predict greatness in other fields. In particular, Goldberg notes the failure of most Heisman Trophy winners to succeed in professional football. Goldberg suggests that success at the lower levels of the federal courts may require different talents than success at the highest judicial level, just as success in college football may require different talents than success at the professional level. Goldberg leaves open the possibility that the talents required to succeed at the highest judicial level can be empirically assessed but concludes that Choi and Gulati's tournament is unlikely to predict greatness in Supreme Court Justices.

John Orth's essay looks to history as a reminder that judges have been judged as long as we have had courts.¹³ Of course, elite and lay public opinion frequently is directed toward judicial decisions as well as individual judges. Many states directly elect judges, a relic of the Jacksonian Era. Litigants, too, frequently cast judgment on the judges before whom they have appeared. In addition, the political process may serve to critique judges through the process of impeachment and other removals. Despite these many institutionalized mechanisms for

12. Steven Goldberg, *Federal Judges and the Heisman Trophy*, 32 FLA. ST. U. L. REV. 1237 (2005).

13. John V. Orth, *Who Judges the Judges?*, 32 FLA. ST. U. L. REV. 1245 (2005).

evaluating judges, Orth is reluctant to embrace efforts to measure judicial performance empirically as a basis for judicial selection and promotions. To the extent that most empirical metrics focus on the outcome of courts—the *decision* rather than the *opinion*—Orth argues that empirical studies fail to capture important aspects of the judicial process and function. Moreover, Orth argues that many empirical measures of influence, such as citation studies, rely primarily on judicial peers to judge performance, thus threatening the constitutional balance of powers by promoting consensus candidates, at the cost of political judgment by other branches.

Lawrence Solum, who was one of the first in print to criticize Choi and Gulati's tournament idea,¹⁴ poses a serious conceptual challenge to any empirical tournament of judges.¹⁵ As he suggests, by beginning with the available data rather than a concept of judicial excellence, Choi and Gulati beg a fundamental question: What is judicial excellence? Solum begins his project with virtue rather than *measures* of virtue. Solum identifies the “thin” judicial virtues—those on which there is widespread agreement—as extending much broader than Choi and Gulati's measures. These virtues include incorruptibility and judicial sobriety, civic courage, judicial temperament and impartiality, diligence and carefulness, judicial intelligence and learnedness, and craft and skill. In addition, Solum makes a plea for recognition of “thick” judicial virtues: *nomimos* (roughly, the virtue of justice) and *phronimos* (roughly, judicial wisdom). If, as Solum suggests, these virtues in the practice of judging are what constitute excellence, then, clearly, it cannot be measured empirically. Solum further observes that to the extent Choi and Gulati measure something less than excellence, he sees many opportunities for gaming this already imperfect measure. Solum argues that the tournament fails to produce accurate and meaningful statistics, let alone a useful and legitimate process of judicial selection based on excellence. Nevertheless, he sees the judicial tournament as a useful thought experiment to the extent that it asks valuable questions about both the content of judicial excellence and the selection of judges who possess it.

David Vladeck brings a litigator's perspective to bear to the critique of an empirical tournament as a measurement of judicial performance.¹⁶ Vladeck shares Choi and Gulati's distaste for a political judicial appointments process and litmus tests as a basis for judicial selection, suggesting that litigators look for competent judges who value collegial-

14. See Lawrence B. Solum, *A Tournament for Judges. Mad? Brilliant? Clever?*, LEGAL THEORY BLOG (Apr. 17, 2004), at http://lsolum.blogspot.com/archives/2003_04_01_lsolum_archive.html#200162580.

15. Lawrence B. Solum, *A Tournament of Virtue*, 32 FLA. ST. U. L. REV. 1365 (2005).

16. David C. Vladeck, *Keeping Score: The Utility of Empirical Measurements in Judicial Selection*, 32 FLA. ST. U. L. REV. 1415 (2005).

ity and efficiency in dispute resolution. Most litigators would rather be spared the “inept, plodding, or even mediocre”¹⁷ judges, even if those judges are seen as the champions of important political values. Vladeck is wary of efforts to attribute objectivity and quantitative measurement to the judicial attributes that litigators value. For example, he observes that focus on the product of the judicial decision fails to capture much of the collaborative process of judging. Yet he does see some modest promise to the enterprise of empirical measurement of judges as delimiting a baseline for determining judicial competence in judicial selection. Vladeck would broaden Choi and Gulati’s factors to include many questions relating to the experience of a potential nominee in practice, government, and the community. He would see the absence of empirical performance on the relevant criteria as a danger signal, but superior performance on one or more criterion should not necessarily be a determinative factor in influencing judicial selection.

II. NEW EMPIRICAL WORK ON JUDICIAL PERFORMANCE: HOW DO WE OPERATIONALIZE AND MEASURE IT?

Choi and Gulati’s study poses many new questions as to how we should best measure variables empirically. Several of the essays in this Symposium refine the operationalization of the criteria Choi and Gulati put forth. Others attempt new measurements of judicial performance or put Choi and Gulati’s criteria to a reality test against the data. These essays add important knowledge to the body of empirical scholarship on courts and judges.

James Brudney starts his essay with an assessment of how effectively Choi and Gulati’s proposed tournament would have predicted the careers of two Supreme Court nominees from an earlier era.¹⁸ Brudney applies the Choi and Gulati analysis to the appointments of Chief Justice Burger and Justice Blackmun. Brudney concludes that the Choi and Gulati analysis would have produced inaccurate predictions about their Supreme Court careers. Under the Choi and Gulati analysis, Chief Justice Burger should have done very well on the Supreme Court. According to the Choi and Gulati criteria, Burger scored favorably on productivity and independence measures while serving as a court of appeals judge, while Justice Blackmun scored markedly lower in both areas. In contrast to this prediction, Burger is widely viewed as a poor Chief Justice, while Blackmun is equally widely viewed as relatively distinguished. Brudney concludes that Choi and Gulati’s criteria fail to measure intangible personal factors that contribute to greatness in Supreme Court Justices—such as the

17. *Id.* at 1416.

18. James J. Brudney, *Foreseeing Greatness? Measurable Performance Criteria and the Selection of Supreme Court Justices*, 32 FLA. ST. U. L. REV. 1015 (2005).

arrogance and aloofness that characterized Chief Justice Burger's tenure on the Court and the ability to change and grow that characterized Justice Blackmun's Supreme Court career. Brudney also draws the broader conclusion that Choi and Gulati's system unwisely attempts to remove policy and ideological considerations from the judicial selection process. Brudney notes that many observers (and members of the general public) view these factors as highly significant, especially in light of the Supreme Court's important role in shaping the basic collective values of the American political system.

Stephen Choi and Mitu Gulati continue their effort to devise objective measures of quality and judicial merit in their contribution, *Which Judges Write Their Opinions (And Should We Care)?*.¹⁹ They use techniques from computational linguistics and other methods to explore both the desirability and feasibility of determining whether individual judges write their own judicial opinions. In the first part of their essay, Choi and Gulati argue that knowing the authorship of judicial opinions is highly relevant when deciding whether to elevate a judge to a higher court. They also argue that this information will be useful in assessing the ongoing performance of sitting judges and in making determinations about the allocation of judicial resources. Assuming this information is useful, Choi and Gulati attempt to devise a method to assess judicial authorship. Unfortunately, the methodology chosen by Choi and Gulati failed to identify judges who, by reputation, are known to write their own opinions. Choi and Gulati speculate on the failure of this methodology and suggest possibilities for future research using a more finely honed version of the methodology, which would control for opinions regarding different subject matter and would focus on factors such as citation practices; opinion, paragraph, and sentence length; and other, frequency-related measures.

Sharing Brudney's views, Lee Epstein, Jeffrey Segal, Nancy Staudt, and René Lindstädt are skeptical about the broad assumptions about the judicial nomination process that motivate Choi and Gulati's proposed tournament of judges.²⁰ Specifically, these four authors dispute the conclusion that policy determinations and ideological concerns have come to dominate the process of judicial nomination and confirmation. They base this claim on their comprehensive study of all votes cast by individual Senators on Supreme Court nominees since Earl Warren in 1953. Their study indicates that although ideology plays a significant role in determining whether a

19. Stephen J. Choi & G. Mitu Gulati, *Which Judges Write Their Opinions (And Should We Care)?*, 32 FLA. ST. U. L. REV. 1077 (2005).

20. Lee Epstein, Jeffrey A. Segal, Nancy Staudt & René Lindstädt, *The Role of Qualifications in the Confirmation of Nominees to the U.S. Supreme Court*, 32 FLA. ST. U. L. REV. 1145 (2005).

Senator will vote for or against a nominee, the qualifications of the nominee will play virtually the same role. It is the interplay of these two factors—ideology and qualifications—that provides the best predictor of a Senator’s vote. While Senators will almost certainly vote for a nominee whose ideology is similar and who is highly qualified and will almost certainly vote against a nominee who is ideologically dissimilar and unqualified, there are certain conditions in which Senators will often vote for an ideologically dissimilar nominee who is highly qualified. The authors conclude that despite common complaints about the nomination system, the qualifications of a nominee continue to play a significant role in the confirmation process. The authors argue that this conclusion should give pause to those—including Choi and Gulati—who propose to fundamentally alter the current judicial selection process.

Daniel Farber starts his essay, *Supreme Court Selection and Measures of Past Judicial Performance*,²¹ by praising Choi and Gulati’s contribution to the judicial selection literature. Farber specifically compliments them for showing how objective measures of past judicial performance could improve the judicial selection process. He then suggests, however, that Choi and Gulati’s proposals are far too ambitious based on the evidence they have mustered. Farber argues that we should be reluctant to adopt their proposal because it is impossible to measure past judicial performance as precisely as they claim and because professional merit should not be the only factor in the judicial selection process. With regard to the first point, Farber notes that some of the criteria used by Choi and Gulati to measure productivity and independence are imperfect because high scores on these criteria may be a function of a judge’s personal characteristics (such as self-centeredness) that would not well serve a Supreme Court Justice. Other measures, such as the use of citation counts as a proxy for influence, are imperfect because of outlier effects and feedback loops that generate citations primarily due to a judge’s personal prominence, which may or may not be directly correlated with objective merit. In the end, Farber engages Choi and Gulati on their own terms more than most of the commentators—to the extent their tournament provides at least a rough measure of judicial merit—but he also argues that merit alone should not be (and probably will never be) the deciding factor in judicial selection.

Michael Gerhardt’s essay surveys the problems with defining three critical factors that are often central to the judicial selection

21. Daniel A. Farber, *Supreme Court Selection and Measures of Past Judicial Performance*, 32 FLA. ST. U. L. REV. 1175 (2005).

process: merit, the “mainstream,” and ideology.²² In the first part of his essay, Gerhardt documents the different ways in which merit is defined by different constituencies and discusses the assumptions about the judicial role that underlie these different definitions. The second part of Gerhardt’s essay investigates the effort to define the “mainstream” in battles over judicial appointments. Although the fight over the “mainstream” is common to almost all judicial appointments battles, it is unclear how this middle ground should be defined. Gerhardt provides several options and considers the various perspectives from which the middle ground should be viewed. The third part of his essay assesses the ways in which the effects of ideology can be assessed in judicial appointments battles. Despite the difficulties presented by these three aspects of the judicial appointments process, Gerhardt argues that an objective empirical study of merit is both possible and desirable. In the absence of some mechanism to assess judicial appointments by reference to a coherent and comprehensive definition of merit, Gerhardt concludes, judges will be viewed as little more than “policymakers who just happen to wear robes.”²³

Michael Solimine’s contribution, *Judicial Stratification and the Reputations of the United States Courts of Appeals*,²⁴ advances the empirical project by applying the tournament concept to performance across, rather than within, appellate courts. Like many of the other contributions, Solimine disentangles the concepts of reputation, prestige, and influence in hopes of identifying a dependable measure of quality. Insofar as the subject of empirical inquiry is federal appellate court judges, he suggests that an effort to separate empirical study of the performance of judges from the performance of the courts they compose is useful. After surveying historical efforts to measure the reputations of courts of appeals, in which the D.C. Circuit and Second Circuit seem to consistently enjoy the highest reputation and most influence, Solimine considers how studies of the citation analysis of particular appellate court judges relate to the reputation of the circuit on which they sit. Some of the measures of influence, Solimine observes, place the Seventh Circuit above the D.C. Circuit and Second Circuit. Solimine’s analysis of the rise and fall of reputation sheds light on the nature and importance of reputation of appellate courts generally. For instance, he observes that during the latter part of the twentieth century, reputation among circuits became homogenized and the importance of reputation fell among litigants.

22. Michael J. Gerhardt, *Judicial Selection by the Numbers*, 32 FLA. ST. U. L. REV. 1197 (2005).

23. *Id.* at 1235.

24. Michael E. Solimine, *Judicial Stratification and the Reputations of the United States Courts of Appeals*, 32 FLA. ST. U. L. REV. 1331 (2005).

Solimine's analysis also illuminates disconnects in reputation between appellate circuits.

III. DO MEASURES OF JUDICIAL PERFORMANCE CREATE POSITIVE BEHAVIORAL AND INSTITUTIONAL INCENTIVES?

Regardless of whether and how we measure judicial performance, everyone agrees that the empirical enterprise has important implications for behavioral incentives and institutions. If we measure—and especially if we rank or otherwise evaluate—actors in an institutional setting based on their performance on a set of criteria, these actors will adjust their conduct in response to the positive and negative incentives created by this information. This seems uncontroversial in concept. But any evaluation of judicial incentives requires some specification of the motivations of judges. In a famous article, Judge Richard Posner suggested that judges maximize “the same thing everybody else does,”²⁵ but among (most) lawyers and judges, it remains controversial to claim that judges are motivated by anything other than doing justice. For political scientists, economists, and some legal scholars, discussion of incentives affecting judicial behavior is much less maligned. Several of the essays make important behavioral and institutional insights which will have implications for further empirical and conceptual research on judges and courts.

Judge Jay Bybee and Thomas Miles join other commentators in questioning Choi and Gulati's central argument that an empirical tournament can settle disputes over the qualifications of candidates for judicial appointments.²⁶ While acknowledging that empirical assessments are an important factor in the appointments process, Bybee and Miles argue that other, less quantifiable subjective measures of quality are even more significant. For example, they suggest that an analysis of a candidate's position on “hot-button” issues such as affirmative action and abortion can often reveal more about a candidate's qualifications than the empirical measures suggested by Choi and Gulati. Also, Bybee and Miles observe that the tournament would create strong incentives for ambitious judges to engage in undesirable behavior simply to increase their judicial score. Ambitious judges will be encouraged to “judge to the tournament,”²⁷ which will have the effect of undermining the spirit of judicial independence and pursuit of judicial quality that the tournament is supposed to encourage.

25. Richard A. Posner, *What Do Judges and Justices Maximize? (The Same Thing Everybody Else Does)*, 3 SUP. CT. ECON. REV. 1 (1993).

26. Hon. Jay S. Bybee & Thomas J. Miles, *Judging the Tournament*, 32 FLA. ST. U. L. REV. 1055 (2005).

27. *Id.* at 1068-73.

Judge Richard Posner's contribution to the issue, *Judicial Behavior and Performance: An Economic Approach*,²⁸ should be taken very seriously by both judges and scholars—and not only because Posner is the likely victor in Choi and Gulati's tournament of appellate judges. As is characteristic of Posner's work, the contribution is brimming with valuable insights about the implications of empirical measures of individual performance of judges. Posner challenges empirical and behavioral scholars of the judiciary to start with a more serious institutional framework for understanding decisions rather than simply with data. As Posner's approach would suggest, "judicial behavior is likely to differ across national legal systems and indeed within a nation's legal systems to the extent that components of the system . . . differ in the incentives and constraints that they impose on judges."²⁹ He extends the analysis not only to appellate court judges—for which he, like Brudney, would like to see greater historical research of how a judge's performance on the U.S. court of appeals maps onto performance as a Justice of the U.S. Supreme Court—but also to the different institutional settings of U.S. district court judges, state court judges, and arbitrators. Posner illuminates how each of the institutional settings presents unique behavioral incentives for individual judges. His analysis of behavioral incentives in the institutional context raises many additional questions and hypotheses for empirical research not only of the behavior of judges but also of other actors within the judicial system.

Judge Bruce Selya evaluates the dual objective of the tournament: merit-based evaluation and increased incentive to perform.³⁰ The proliferation of multiple objective metrics may undermine the goal of political transparency: he suggests that "any ranking system will face constant criticism that it is a proxy for either political affiliation or ideological leanings rather than for merit."³¹ As to Choi and Gulati's criteria, Judge Selya takes issue with several of them—particularly their incentive effects. Commenting on the manipulability of citation measures, he suggests that "[a]ny judge worth his salt will tell you that there are ways to write opinions that make citation more likely."³² He also characterizes the measurement of judicial independence by use of dissents as useless as well as perverse in its incentives. Selya points out that, over time, "judicial rankings will say less about actual merit and more about agility—the ability to game

28. Richard A. Posner, *Judicial Behavior and Performance: An Economic Approach*, 32 FLA. ST. U. L. REV. 1259 (2005).

29. *Id.* at 1259.

30. Hon. 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).

31. *Id.* at 1286.

32. *Id.* at 1290.

the system.”³³ The inevitable result of these incentives, Judge Selya warns, is to undermine the very goals of the tournament.

Russell Smyth’s essay illustrates the promise of taking seriously the institutional context of judging.³⁴ Smyth rejects the suggestion of Judge Harry Edwards³⁵ that the collegiality of judging cannot be measured, suggesting that “there is much evidence of the success of regression analysis in capturing these nuances of human behavior.”³⁶ Smyth examines how the idea of the judicial tournament would translate in the setting of judicial selection in Australia. He carefully builds from an evaluation of judicial incentives, pointing out both behavioral and quantitative problems with many of the Choi and Gulati criteria along the way. Smyth argues that a tournament transplanted to the Antipodes would best focus on minimum qualifications for judges, rather than on the highfliers. His essay illustrates how one need not reject empirical study of judges in order to accept the criticisms of the judicial tournament many others make.³⁷

Ahmed Taha’s essay provides a limited defense of an empirical judicial tournament.³⁸ Taha focuses on the effects of rankings on the judicial nomination and confirmation process as well as the effects of rankings on judges’ behaviors. He argues that the case for a tournament of judges as a judicial selection device is stronger for the positions to which Choi and Gulati apply the tournament—U.S. courts of appeals—than for positions on the U.S. Supreme Court. As he argues, basing selection of U.S. Supreme Court Justices on a tournament may undermine Choi and Gulati’s goals by allowing “a President to nominate more politically extreme candidates who happen to have high merit rankings.”³⁹ By contrast, “a ranking of federal district [court] judges would avoid many of the problems that might be created by ranking appellate judges.”⁴⁰ Further, based on an assessment of judicial behavior, he suggests that, if subjected to the tournament, federal district judges are more likely to respond to the positive incentives it creates.

33. *Id.* at 1294-95.

34. Russell Smyth, *Do Judges Behave as Homo Economicus, and if So, Can We Measure Their Performance? An Antipodean Perspective on a Tournament of Judges*, 32 FLA. ST. U. L. REV. 1299 (2005).

35. Edwards, *supra* note 9, at 1640-43, 1656.

36. Smyth, *supra* note 34, at 1319-20.

37. *See supra* Part I.

38. Ahmed E. Taha, *Information and the Selection of Judges: A Comment on “A Tournament of Judges,”* 32 FLA. ST. U. L. REV. 1401 (2005).

39. *Id.* at 1406.

40. *Id.* at 1403.

IV. CONCLUSION

There is an elegance about the empirical tournament as a mechanism for making the intangible knowable and defusing political rhetoric in judicial selection. Of course, it does not succeed in these respects. As the essays in this Symposium suggest, the project of empirical evaluation of judicial performance must also address normative questions of quality, and it raises important issues about measurement, behaviors, and incentives. A focus on empirical issues relating to the activity of judging and its outputs not only feeds the prurience of the legal community for gossip about what lies beneath the robes of judges, but it can help to shed light on fundamental issues of judicial behavior, important to students and scholars in law, political science, and economics. It also challenges scholars to disentangle important questions relating to the quality of judging and judicial institutions. For example, to what extent can judicial performance be reduced to a judge's reputation? Or does quality in judging represent something more? To what extent, if any, does judicial quality in one institutional setting, such as in the context of the U.S. circuit courts of appeals, necessarily translate into quality in another judicial setting, such as in the context of the U.S. Supreme Court? Does a Supreme Court nomination system that keeps Judge Posner from becoming Justice Posner really fail to recognize or reward merit?

Choi and Gulati's empirical tournament presents a much-welcomed challenge: How might we introduce greater objectivity into discussions of merit in judicial selection? The scholarly ruminations in the pages of this Symposium certainly cannot answer all of the questions empirical studies of judging and courts present. Much as the initial furor the tournament provoked, some of the commentators in this Symposium refuse Choi and Gulati's challenge on normative terms. Nevertheless, as the Symposium contributions clearly indicate, an empirical tournament tells us much about several enterprises: normative theories of judging, quality, judicial independence, and judicial selection; empirical discussions of measurement of judges and courts; and behavioral and incentive-based evaluations of institutions. As the essays in the Symposium indicate, the empirical tournament has inspired some important advances in the discourse about measurement of performance in the context of the judiciary and its relevance to the selection of judges and the judicial process. That is a discourse that will be certain to continue as long as we have data, judges, and courts.