

WHICH JUDGES WRITE THEIR OPINIONS (AND SHOULD WE CARE)?

STEPHEN J. CHOI* AND G. MITU GULATI**

PROLOGUE: LAWYERS AND THE ATTRIBUTION OF AUTHORSHIP	1077
I. INTRODUCTION	1078
II. THE UPSIDES AND DOWNSIDES TO DETERMINING JUDICIAL AUTHORSHIP	1081
A. <i>Upsides</i>	1083
1. <i>Promotions and the Quality of Judicial Output</i>	1083
2. <i>Fine-Tuning Incentives</i>	1086
3. <i>Allocation of Resources</i>	1088
4. <i>Research on Judicial Behavior</i>	1089
5. <i>Information for Law Clerks</i>	1091
6. <i>Informational Benefits to Opinion Users</i>	1092
B. <i>Downsides</i>	1094
1. <i>Danger of Unjustified Inferences</i>	1094
2. <i>Imperfect Measurement</i>	1095
III. TESTING AUTHORSHIP	1096
A. <i>Tracking Judicial Fingerprints</i>	1099
B. <i>Ranking Judges</i>	1105
C. <i>White-Box Tests</i>	1111
1. <i>Citation Practices</i>	1111
2. <i>Subject Matter-Neutral Language Patterns</i>	1116
3. <i>Revisiting the Black-Box Tests</i>	1120
IV. CONCLUSION: OTHER POSSIBLE APPLICATIONS OF AUTHORSHIP TECHNOLOGY	1121
A. <i>Securities Fraud Complaints</i>	1121
B. <i>Boilerplate Contract Evolution</i>	1122

PROLOGUE: LAWYERS AND THE ATTRIBUTION OF AUTHORSHIP

If one of our students were to pass off someone else’s work as his or her own and we were to discover that misrepresentation, the penalties would be harsh. But as lawyers, we hold ourselves to less rigorous standards for plagiarism than those to which we hold our students. Where research assistants make significant contributions to

* Professor of Law, New York University School of Law.

** Visiting Professor of Law, University of Virginia School of Law; Professor of Law, Georgetown University Law Center. We are indebted to David Keeney, Robert Mezey, and Craig Hoffman for conversations about linguistic theory, authorship styles, and how one might track and measure style differentials in judicial opinions as opposed to fiction or poetry. We are grateful to the computer scientists and programmers at InApp in India, including Satish Babu, M.C. Jayakrishnan, and R.V. Suchitra, for their innovations in designing programs to run standard authorship tests on judicial opinions. For conversations about the ideas here, we also thank Gail Agrawal, Scott Baker, Michael BeVier, Devon Carbado, Adrienne Davis, Susan French, Dennis Hutchinson, Edeanna Johnson, Kimberly Krawiec, Richard Posner, Un Kyung Park, Tom Rowe, Michael Solimine, David Vladeck, and P. Vijaykumar. Our thanks also go to Associate Dean Jim Rossi and the Law Review editors at the Florida State University College of Law for organizing this Symposium. The library staffs at the University of Virginia School of Law and the Georgetown Law Center provided us with invaluable research assistance. Chris Knott and Kent Olson were especially generous with their assistance. Finally, we owe a debt to Lee Epstein for giving us the initial impetus for the project.

their professors' papers and sometimes even draft large sections of those papers, they are rarely given anything more than an acknowledgement in a footnote.¹ Law firm partners, we suspect, think nothing of asking junior associates to draft entire articles or book chapters and then send them out under their name (again, often without acknowledgement of the flunky's authorship). And then there is the matter of judges. Law clerks are said to draft the vast majority of opinions for judges. Yet, if one were to ask most lawyers and judges whether authorship credit should be given to the individual clerks, they would in all likelihood think the question ridiculous. Why do we, as lawyers, consider proper attribution of authorship so important for our students and so unimportant for ourselves?²

I. INTRODUCTION

Federal judges enjoy a large amount of discretion in how they go about writing opinions. Once a judge is assigned an opinion, the judge may choose to write the opinion alone, doing both the research and writing without any assistance. Judges may also turn to their clerks to help research relevant law or to draft parts of the opinion. Some judges may allow clerks to draft entire opinions, while they themselves contribute only a light editorial overview. Commentators have discussed the practice of delegating significant portions of the opinion-writing task to clerks, and more than a few have criticized it.³ But no one seems to know exactly how much delegation goes on;

1. Law professors, their misuse of sources, and failures to properly attribute credit have been in the news lately. See Daniel J. Hemel & Lauren A.E. Schuker, *Prof Admits to Misusing Source*, HARV. CRIMSON ONLINE, Sept. 27, 2004, at <http://www.thecrimson.com/article.aspx?ref=503493> (quoting Harvard Law Professor Alan Dershowitz as making the justificatory point that although the rules governing plagiarism by undergraduates are clear (and harsh), the norms of attribution and citation in the legal profession—where judges frequently rely on lawyers' briefs and clerks' memoranda in drafting opinions—are less clear); Sara Rimer, *When Plagiarism's Shadow Falls on Admired Scholars*, N.Y. TIMES, Nov. 24, 2004, at B9 (discussing criticisms of Harvard Law Professors Laurence Tribe and Charles Ogletree Jr.).

2. See Marilyn V. Yarbrough, *Do as I Say, Not as I Do: Mixed Messages for Law Students*, 100 DICK. L. REV. 677 (1996) (discussing the mixed messages about plagiarism that law professors and legal professionals send to law students).

3. For examples of works discussing the practice of delegating to law clerks, see FRANK M. COFFIN, *ON APPEAL: COURTS, LAWYERING, AND JUDGING* (1994); JONATHAN MATTHEW COHEN, *INSIDE APPELLATE COURTS* 9-11 (2002); WILLIAM DOMNARSKI, *IN THE OPINION OF THE COURT* 42-45 (1996); RICHARD A. POSNER, *THE FEDERAL COURTS: CHALLENGE AND REFORM* 139-59 (1996); David Crump, *Law Clerks: Their Roles and Relationships with Their Judges*, 69 JUDICATURE 236, 238 (1986); Sally J. Kenney, *Puppeteers or Agents? What Lazarus's Closed Chambers Adds to Our Understanding of Law Clerks at the U.S. Supreme Court*, 25 LAW & SOC. INQUIRY 185, 200-06 (2000); J. Daniel Mahoney, *Law Clerks: For Better or for Worse?*, 54 BROOK. L. REV. 321, 332-34, 338-44 (1988); David McGowan, *Judicial Writing and the Ethics of the Judicial Office*, 14 GEO. J. LEGAL ETHICS 509, 555-67 (2001); Nadine J. Wichern, Comment, *A Court of Clerks, Not of Men: Serving Justice in the Media Age*, 49 DEPAUL L. REV. 621 (1999); Marvin E. Frankel, *A Matter of Opinions*, N.Y. TIMES, May 15, 1994, at E15; Alain L. Sanders, *Putting a Thumbprint on History*, TIME, Aug. 6, 1990, at 75. Going back further in time, see THE FORGOTTEN

commentators rely on anecdotes, rumors, and, at best, informal surveys.⁴

In prior work, we asked why, in selecting Supreme Court Justices, there was little attempt to use the available data on the relative performances of federal circuit court judges (the primary pool from which Supreme Court Justices are chosen).⁵ In this Essay, we ask why few have attempted to provide a systematic analysis of authorship patterns among federal circuit court judges. We use generic techniques from computational linguistics, as well as several methods tailored for the judicial setting, to explore both the desirability and feasibility of determining the authorship of judicial opinions.

Information about the production process (as opposed to information about the end product alone) can be useful in decisions regarding judicial promotion. Whether a particular judge expends effort in authoring her own opinions may have bearing on how suited the judge is for elevation to a higher court, including the Supreme Court. Knowing whether individual judges use their scarce time to manage their clerks' writing or to engage in the writing task themselves may help determine whether there is a need for more judges and resources for the judiciary. If we observe that a spike in the volume of cases coincides with judges suddenly placing greater reliance on their clerks to write opinions, we might want to support an expansion of the federal judiciary. Authorship information may also help identify judges who are no longer able or willing to perform their tasks adequately. Once identified, peer pressure and other forms of public opprobrium may lead the judges to either increase their activity level or

MEMOIR OF JOHN KNOX: A YEAR IN THE LIFE OF A SUPREME COURT CLERK IN FDR'S WASHINGTON (Dennis J. Hutchinson & David J. Garrow eds., 2002); Chester A. Newland, *Personal Assistants to Supreme Court Justices: The Law Clerks*, 40 OR. L. REV. 299, 312-16 (1961); see also HENRY J. ABRAHAM, *THE JUDICIAL PROCESS* 238 (3d ed. 1975) (quoting Dean Acheson on his clerkship experience with Justice Brandeis); DREW PEARSON & ROBERT S. ALLEN, *THE NINE OLD MEN* 109 (1936) (discussing Justice Harlan Stone's use of his secretary during his drafting process). The most recent and most serious attempt to systematically survey former Supreme Court law clerks about their involvement in the opinion-writing process is contained in the forthcoming book, ARTEMUS WARD & DAVID L. WEIDEN, *SORCERER'S APPRENTICE: LAW CLERKS AT THE U.S. SUPREME COURT* (forthcoming 2005) (manuscript at ch. 5, at 36, on file with authors) ("Nineteen percent of the clerks said that their justice made revisions in most cases. Seven percent said that their justice only changed clerk-written drafts in some cases and four percent said that revisions were made in few or no cases. What is striking from these results is that thirty percent of clerks had their drafts issued without modification, as opinions by their justice at least some of the time.").

4. As the reader will later see, the criticism about the reliance on anecdote, rumor, and informal surveys is something that our project is also subject to since we needed some baseline piece of information about "true" authorship levels to test out various authorship testing methods. The goal, however, is to develop a set of methods of testing authorship such that reliance on these often nonverifiable methods can be reduced.

5. See 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).

retire. Law students applying for judicial clerkships will find it useful to know whether their potential employer does her own writing or delegates it all to her clerks. And—highly relevant from our perspective as researchers—authorship information has the potential to improve academic research on judicial behavior.

Part II of this Essay sets out a framework to explore whether it is worthwhile to investigate which judges write their opinions. We explore whether there is societal value in knowing information about the input levels that society's agents (in this case, the judges) put into the production process (the product being judicial opinions).

A variety of techniques may be employed to determine the authorship of opinions. Part III sets out tests to determine the viability of these techniques. We use the limited existing information on authorship—a handful of judges such as Richard Posner, Frank Easterbrook, and Michael Boudin participate more actively in the writing of their opinions than do most other judges⁶—to test how well the different authorship tests perform in distinguishing such judges.

Our sample pool consists of all circuit court judges who were both active and under the age of sixty-five as of May 2003 and had been on the bench for the period from January 1, 1998 to December 31, 2000.⁷ For each judge in the sample, we selected opinions for analysis at random from those generated during the three-year sample period.

Based on this sample, we report that the generic tests drawn from computational linguistics fail to distinguish Posner, Easterbrook, and Boudin as judges most likely to author their opinions. The generic tests, however, do not control for the subject matter of specific opinions. The common phrases used in opinions of a specific genre (for example, administrative law opinions) will cause the generic methodologies to treat all such opinions as more likely to be by the same author, even if different authors actually wrote the opinions. If the randomly selected opinions for one judge are all of the same subject matter but those for another judge are not, this factor alone may lead

6. Several commentators have remarked on the writing practices of Judges Posner and Easterbrook. See Robert F. Blomquist, *Dissent, Posner-Style: Judge Richard A. Posner's First Decade of Dissenting Opinions, 1981-1991—Toward an Aesthetics of Judicial Dissenting Style*, 69 MO. L. REV. 73, 74 n.12 (2004); Martha Middleton, *Shaping a Circuit in the Chicago School Image*, NAT'L L.J., July 20, 1987, at 1 (discussing Easterbrook and Posner). Informal conversations with a number of other judges and former law clerks confirmed that Judges Posner and Easterbrook author all of their own opinions. Many also mentioned Judge Boudin as someone who authored most of his own opinions. Other judges were also mentioned, albeit less frequently, as authoring substantial portions of their opinions. We use Posner, Easterbrook, and Boudin—those judges with the highest a priori likelihood of self-authorship—to calibrate the effectiveness of our authorship methodologies.

7. We generated much of this dataset in our earlier article ranking judges based on judicial performance. See Choi & Gulati, *supra* note 5. To take advantage of some of the data collected for that project, we restrict our analysis to the same pool of judges here.

the generic authorship tests to give the first judge a higher self-authorship score.

We also provide more customized tests designed to control, at least in part, for the subject matter of judicial opinions. Using these tests, we are able to distinguish our test judges—Posner, Easterbrook, and Boudin—as ranking consistently high in terms of self-authorship of judicial opinions. Part IV concludes and provides possible extensions of the authorship methodology.

II. THE UPSIDES AND DOWNSIDES TO DETERMINING JUDICIAL AUTHORSHIP

The production of judicial opinions is a joint venture between the judges and their staffs—for purposes of opinion writing, the law clerks. The outside world sees the end product in the final opinion that appears in West's case reporters. Outsiders cannot distinguish the contributions of the various participants. For each opinion, only the primary writing judge is identified (for an appellate opinion, the other judges on the panel are also identified as secondary participants, but their relative involvement is not specified).

Lore has it that many opinions are drafted primarily by the judge's law clerks and sometimes even by staff attorneys. Such delegation may help judges handle their ever-increasing caseloads.⁸ At the other extreme, there are some judges who, despite the caseloads, are reputed to take complete responsibility for opinion writing (for example, Richard Posner and Frank Easterbrook, who seem to be able to do this and more⁹). Further, occasions may arise when the other judges on the panel draft significant portions of the opinion. But the involvement of these subsidiary coauthors is never identified. In addition, norms of appropriate behavior mean that the clerks rarely reveal what occurred in their judges' chambers; the information the clerks possess about relative responsibility for the various opinions is seldom reported in any public and verifiable manner.¹⁰

8. See Thomas E. Baker, *Intramural Reforms: How the U.S. Courts of Appeals Have Helped Themselves*, 22 FLA. ST. U. L. REV. 913, 944-45 (1995) (noting that growth in caseload has led to an increase in the number of law clerks and in the amount of work delegated to them); William M. Richman & William L. Reynolds, *Elitism, Expediency, and the New Certiorari: Requiem for the Learned Hand Tradition*, 81 CORNELL L. REV. 273, 274 n.3 (1996) (citing materials that document the increase in judicial caseload).

9. See *supra* note 6.

10. When clerks do reveal the goings-on in their judges' chambers, they can sometimes receive considerable criticism. A recent example is the criticism that was leveled at Edward Lazarus' *Closed Chambers*, a memoir of his year on the Supreme Court clerking for Justice Blackmun. See Kenney, *supra* note 3 (discussing some of the criticism). Prior to that, there was criticism of the many law clerks who spoke to Bob Woodward and Scott Armstrong when they were researching their book on the workings of the Court, *The Brethren*. See George Gold, *Loose Tongues: How Stone Cautioned Clerks*, A.B.A. J., Oct. 1985, at 28; Mahoney, *supra* note 3, at 336.

The agency concept provides a useful framework for thinking about the judicial production problem. Judges are agents who perform a set of tasks, which include producing judicial opinions, for the public (the principal in the relationship). A variety of consumers—lawyers, law students, litigants, researchers, and other judges, among others—use these judicial opinions to understand the law. In order to perform their tasks, judges are authorized to employ a set of subsidiary agents (law clerks and staff attorneys), who are hired, supervised, and evaluated by the judges themselves. Society's goal as the principal is arguably to ensure that the judges produce at maximal potential, while at the same time providing for the judges' independence.

Ordinarily, principals attempt to monitor the conduct of their agents. If the agents do not work hard enough or produce high-quality products, they get fired (or, in the case of politicians, they get voted out of office). Since society wishes to ensure the independence of federal judges, however, they are provided with virtually ironclad job security and fixed salaries. These elements of job security and fixed incomes mean that the public has relatively few levers with which to incentivize the judges; for the most part, judges do their work because they want to. It may appear pointless to inquire into judges' relative levels of involvement in the production of opinions. Indeed, that view may be one reason why such information thus far has not been collected.

As illustrated by the materials mentioned above, the one context where there has been some minimal revelation of the levels of delegation to law clerks is the Supreme Court. Even here the information is highly imperfect, often a function of clerk reconstructions in contexts where the explicit goal is to praise the Justice or journalistic reports about high profile cases resting on undisclosed sources. But at least there is some information. Plus, to the extent that some former Justices have (a) opened their papers to the public and (b) kept accurate records on clerk-judge communication, researchers have a starting point to try and determine the level of clerk delegation. Among the prominent recent examples along these lines are the *Legal Affairs* article by David Garrow criticizing the amount of delegation of power by Justice Blackmun to his clerks (especially later in his career) and the *Vanity Fair* exposé, following the 2000 election, that used information from law clerks to construct a story about how the decision was made. See David Garrow, *The Brains Behind Blackmun*, LEGAL AFF., May-June 2005, available at http://www.legalaffairs.org/issues/May-June-2005/feature_garrow_mayjun05.msp; David Margolick et al., *The Path to Florida*, VANITY FAIR, Oct. 2004, at 310. For examples of former law clerks writing about their Justices, see JOHN C. JEFFRIES, JR., JUSTICE LEWIS F. POWELL, JR. (1994); J. HARVIE WILKINSON, III, SERVING JUSTICE: A SUPREME COURT CLERK'S VIEW (1974); Bruce A. Ackerman, *In Memoriam: Henry J. Friendly*, 99 HARV. L. REV. 1709 (1986); Anne M. Coughlin, *Writing for Justice Powell*, 99 COLUM. L. REV. 541 (1999); James L. Volling, *Warren E. Burger: An Independent Pragmatist Remembered*, 22 WM. MITCHELL L. REV. 39 (1996); and Kevin J. Worthen, *Shirt-Tales: Clerking for Byron White*, 1994 BYU L. REV. 349. The precise contours of what can and cannot be revealed by former clerks, however, seems to be unclear. That lack of clarity of the disclosure rules, when combined with the fondness (or fear) that most clerks have for their judges, we suspect, keeps most clerks from going on the record with information about the authorship practices in their judges' chambers.

In Part II.A, we put forward the claim that a better understanding of the level of judicial input into the opinion-writing process can potentially help the management of judicial agents in at least three circumstances: deciding on promotion when the quality of the final output is hard to evaluate, determining incentives for the judges as part of a judicial opinion production team, and assessing how best to allocate resources to the judiciary. We also identify others who may find knowledge on authorship useful: researchers studying judicial behavior, law students entering the judicial clerkship market, and those considering how much to rely on the opinions of specific judges.

Having considered the upsides, in Part II.B we note possible downsides to the project that colleagues have mentioned to us. These are the danger of improper negative inferences and the problem of imperfect measurement.

A. Upsides

1. Promotions and the Quality of Judicial Output

Information on input into the production process can assist in the evaluation of the quality of a final product where such quality is difficult to observe directly. For many products, such as legal and medical services, product quality is hard to measure. When a lawyer loses a case or a transaction fails to be completed, it is difficult to know whether the result should be attributed to the lawyer or to other factors. If a patient's health does not improve under a doctor's care, does this reflect the doctor's failure to deliver competent medical care or the patient's own initial health condition?

Where the final output is difficult to evaluate, another solution is to look to inputs.¹¹ For example, the difficulty in evaluating lawyers' product is often given as a reason for why lawyers bill by the hour.¹² If one can know the levels of effort and skill a lawyer brings to the production process, one can determine whether the bad outcome was the product of inadequate effort and skill, some other factor, or random chance. While doctors are typically not paid on an hourly basis, information on how much attention and effort a doctor vests in a particular patient is likely to be relevant to any potential medical malpractice claim against that doctor.

11. See EDWARD P. LAZEAR, *PERSONNEL ECONOMICS* 20 (1995).

12. There has been considerable debate over whether hours billed serves as a good measure of lawyer performance (or, put differently, an effective means of monitoring). See Scott Baker & Kimberly D. Krawiec, *The Economics of Limited Liability: An Empirical Study of New York Law Firms*, 2005 U. ILL. L. REV. (forthcoming) (discussing the monitoring literature in the law firm context).

In earlier work, we suggested that one way to evaluate the quality of a judicial opinion was to look at citation rates.¹³ Citation rates, however, are an imperfect measure because they are affected by a number of factors other than quality: the subject matter of the case, the reputation of the circuit, and the reputation of the judge.¹⁴ Therefore, knowing the level of the judge's participation in the production process supplements information about citation rates. A judge who rates a low self-authorship score coupled with a low citation rate may signal her relatively low level of engagement in the judicial process.

Information as to individual inputs into the production process can also be useful in instances where the employer cares about more than product quality. An employer might want to know, for example, whether a particular employee is exerting herself fully or only partially. This is information from which the employer can potentially infer a number of things, such as the employee's commitment to the job, her interest in the project, and her likely effect on her coworkers (to the extent they take their cues from her).

For some jobs, process—which might be seen to include the agent's commitment and dedication to the process—might have an importance of its own. In the judicial context, commitment and dedication to the task may suggest a concern for justice. Someone who will work day and night on her task, even if she is not the quickest at it, may be more likely to produce a just decision than someone who is highly intelligent and writes quickly but is not willing to spend more than a few hours a day on a case. The willingness to work hard, we suggest, may correlate with a judge's concern that a case be determined in a fair manner (as opposed to a slovenly willingness to decide the case more randomly). Achieving justice, vague and amorphous a concept as it may be, may be more important to society than ensuring that judges write high-quality opinions.¹⁵ The same kind of argument can be made with respect to other abstract concepts such as honesty and integrity.¹⁶

13. See Choi & Gulati, *supra* note 5, at 34, 48-61.

14. There is an extensive literature on the pitfalls and benefits of using citation rates as a measure of quality. For discussions, see Arthur Austin, *The Reliability of Citation Counts in Judgments on Promotion, Tenure, and Status*, 35 ARIZ. L. REV. 829 (1993); Choi & Gulati, *supra* note 5, at 48 n.38, 54-55; William M. Landes et al., *Judicial Influence: A Citation Analysis of Federal Courts of Appeals Judges*, 27 J. LEGAL STUD. 271 (1998); and 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).

15. See John V. Orth, *Judging the Tournament*, *Jurist* (Apr. 15, 2004), at <http://jurist.law.pitt.edu/forum/symposium-jc/choi-gulati-orth-taha.php>.

16. For example, one could question the honesty and integrity of someone who chose to exert little effort in making a decision even though the lack of effort made it more likely that the outcome was more random and only loosely related to the facts of the specific case.

Consider the problem of deciding judicial promotions. If the job at the lower level is substantially different from that at the promoted level, high-quality production at the lower-level job may not predict high-quality production after promotion. The work of trial judges, for example, is primarily in the area of trial management, with the occasional authorship of an opinion. Circuit court judges, on the other hand, focus to a greater extent on the production of judicial opinions. Where a trial judge is being considered for promotion to a circuit court position, information about dedication to the job (the inputs) might be more important than which employee produced higher-quality opinions at the lower level (the outputs).

Almost all the candidates given serious consideration for appointment to the Supreme Court tend to have prior judicial experience, most of them on the federal circuit courts of appeals.¹⁷ What, then, if we were to find out that a judge who was a candidate for promotion to the Supreme Court wrote none of her own lower court opinions, but instead delegated them all to her law clerks? And what if there was another candidate who wrote all of her own opinions? This information may not be dispositive in terms of eliminating the first candidate, but questions would be raised and inquiries would be made. If it turned out that the first judge spent a significant portion of the year skiing, sailing, or rock climbing, the public would probably conclude that this judge was not committed to the job.

On the other hand, it might turn out that this judge, instead of actually writing her own opinions, was spending her time managing her clerks and training them to write and analyze. If it also turned out that these clerk-written opinions were considered to be of high quality by experts, the final evaluation of this candidate might be positive. This would be especially so if the job at the higher level required even more management and training skills than the lower-level job. Indeed, the inability of the other judge to effectively utilize her clerks in the writing process might be seen by some as a negative factor in the promotion decision.

The information about the judge's involvement in the writing process alone is therefore unlikely to be dispositive. It may, however, lead people to look for other pieces of information so as to be able to make inferences. Judges themselves may voluntarily reveal more information about their style of decisionmaking in response to objective information on their involvement in writing opinions. Some evaluators may see direct judicial involvement in the writing of drafts as of

17. See DAVID ALISTAIR YALOF, PURSUIT OF JUSTICES: PRESIDENTIAL POLITICS AND THE SELECTION OF SUPREME COURT NOMINEES 170 (1999); 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).

paramount importance. Others may regard judges as also having a responsibility to train law clerks. The information that may emerge on judicial input into the writing may be important to both sets of evaluators, although it may lead them to different conclusions about who should be promoted.

2. *Fine-Tuning Incentives*

A much-discussed question in economics is how to devise solutions to the “team production” problem.¹⁸ The team production problem occurs when production occurs in teams and the relative levels of credit for the final product are hard to allocate.¹⁹ Where individual responsibility cannot be allocated, a free-rider problem may arise whereby team members have an incentive to shirk.²⁰ There are methods employers use to solve the team production problem, which include internal peer pressure within the team and incentives to members to monitor and manage the other team members.²¹

In the case of judicial opinion writing, the judge is the manager. The judge, who is allocated full credit for the output of her chambers, has both the ability to monitor subordinates and an incentive to ensure that they work hard. The potential problem, however, is with the judge. If a judge wishes to shirk, there is no formal mechanism to penalize her. A judge can monitor clerks and fire them or give them bad references. But there are few direct mechanisms that enable the public to monitor and penalize a federal judge.

Informal mechanisms, however, do exist. Judges care about status and prestige.²² They will care if information is released showing that some of them demonstrate no involvement in the opinion-writing process (especially if the data also shows that their colleagues are highly involved in that process). Information concerning a lack of authorship, when coupled with a low citation rate, may indicate that a judge is simply not doing a good job, thereby reducing that judge’s reputation among her peers and the general public. Judges may demonstrate greater participation in the authorship process to prevent or remedy this situation.

But what if these judges are not good writers? Might not it be better that they remove themselves from the writing portion of the pro-

18. The classic papers on the team production problem are Armen A. Alchian & Harold Demsetz, *Production, Information Costs, and Economic Organization*, 62 AM. ECON. REV. 777 (1972), and Bengt Holmstrom, *Moral Hazard in Teams*, 13 BELL J. ECON. 324 (1982).

19. Alchian & Demsetz, *supra* note 18, at 779.

20. *Id.* at 780.

21. See Margaret M. Blair & Lynn A. Stout, *A Team Production Theory of Corporate Law*, 85 VA. L. REV. 247 (1999) (describing the literature on the team production problem).

22. Smyth, *supra* note 14, at 1306-07 nn.39-42 (citing support on the point).

ject? The point is a fair one, but it leads to a more important point: the motivation of politicians. Once a judge who has no skill at writing is appointed, there may be a valid argument that the judge should not be compelled to participate in the writing process.²³ But writing skills and the willingness to write opinions should be at least two of the criteria for appointing a federal judge. And if someone who lacks those basic characteristics is appointed, the blame should fall on the politicians who made the appointment. If politicians are penalized for bad appointments on basic matters such as writing skills, they may be discouraged from using judicial appointments as political favors or from allowing their choice to be determined on the basis of litmus tests; for example, a judge's likely vote on issues such as abortion and the death penalty. Politicians may then in fact focus on more mundane, yet fundamental, matters such as writing skill.

Another indirect incentive exists. Like the rest of us, judges become ill, old, and sometimes uninterested in their jobs. Being a judge, however, brings status, power, and guaranteed income. There is the danger that a judge who is beyond the point of being able or willing to perform the job will be tempted to eschew retirement.²⁴ So long as the judge has able law clerks, casual observers on the outside will find it difficult to detect the drop in performance.²⁵ It is likely that the law clerks and the other judges will be able to observe when the judge is not able to perform adequately, but while there may be internal grumblings in the courthouse, this information will rarely find its way to the general public.²⁶ We assume that there is a strong social norm against the disclosure of such information.

If, however, information is available as to judicial input into the writing process, any significant drop in interest or capability will result in public pressure for the judge to retire. For example, it is likely that no litigant wants a case to be decided by a senile or otherwise absent judge. And while the law clerks making decisions for these

23. Alternatively, some might argue that judges who are not good writers are precisely the ones who should be incentivized to write more so that they learn to become better writers. In his diary for the online magazine, *Slate*, Judge Posner writes:

Most judges nowadays, because of heavy caseloads, delegate the writing of their judicial opinions to their clerks. It's a mistake on a number of grounds: The more you write, the faster you write; only the effort to articulate a decision exposes the weak joints in the analysis; and the judge-written opinion provides greater insight into the judge's values and reasoning process and so provides greater information—not least to the judge.

Richard Posner, *Diary: A Weeklong Electronic Journal*, SLATE (Jan. 14, 2002), at <http://slate.msn.com/?id=2060621&entry=2060676>.

24. Cf. Smyth, *supra* note 14, at 1301-03 (pointing to evidence suggesting that judicial retirement rates respond to changes in incentives, such as pension levels).

25. See RICHARD A. POSNER, *OVERCOMING LAW* 112 (1995).

26. Supreme Court Justices are perhaps an exception to the rule, as we see from the rumors about Justice Thurgood Marshall. See *infra* note 42 and accompanying text. Our primary interest in this Essay, however, is in the less visible lower court judges.

absent judges may have done well on their law school exams, the litigants might prefer to have their cases decided by judges who have at least some meaningful experience in the world rather than by students fresh out of law school.

The question that remains is as follows: Would disclosure of this information have any effect on a particular judge, a superannuated one for example? We suspect it would. While there are few external sources of pressure that can be brought to bear on a judge, status, power, and prestige are all a function of the public's perception of that person. If the public begins to lose confidence in the judge because of evidence that her ability to write has declined, the judge may consider either increasing her effort or retiring.

3. Allocation of Resources

The determination of resource allocation for the federal judiciary is made by the legislature. Given that there are no meaningful competitive forces to tell the legislature when the production process is faulty, the legislature itself has to collect the necessary information to make such evaluations. For example, it is well documented that there has been a dramatic explosion in the caseloads carried by federal judges over the past few decades.²⁷ The number of judges, however, has not kept pace with the increased workload.²⁸ This disparity has led some to call for more judgeships;²⁹ these calls have, in turn, been countered by others.³⁰

One element in the argument that more judges are needed has been that judges today, unlike judges decades ago, are forced to turn to law clerks, staff attorneys, and other shortcuts to enable them to tackle their expanded caseloads.³¹ Skeptics, however, say that the caseloads are not so large as to prevent existing judges from tackling them. Some of the questions in the debate might be answered by data

27. See COMM'N ON STRUCTURAL ALTERNATIVES FOR THE FED. COURTS OF APPEALS, FINAL REPORT 13-17 (1998); see also Jeffrey O. Cooper & Douglas A. Berman, *Passive Virtues and Casual Vices in the Federal Courts of Appeals*, 66 BROOK. L. REV. 685 (2000) (discussing the caseload explosion and the responses of the courts of appeals); Lauren K. Robel, *Caseload and Judging: Judicial Adaptations to Caseload*, 1990 BYU L. REV. 3 (describing and critiquing the methods judges have used to deal with increased caseloads).

28. COMM'N ON STRUCTURAL ALTERNATIVES FOR THE FED. COURTS OF APPEALS, *supra* note 27, at 14.

29. See, e.g., Stephen Reinhardt, *A Plea to Save the Federal Courts: Too Few Judges, Too Many Cases*, A.B.A. J., Jan. 1993, at 52; Victor Williams, *Solutions to Federal Judicial Gridlock*, 76 JUDICATURE 185 (1993).

30. See, e.g., Jon O. Newman, *1,000 Judges—The Limit for an Effective Federal Judiciary*, 76 JUDICATURE 187 (1993); J. Harvie Wilkinson III, *The Drawbacks of Growth in the Federal Judiciary*, 43 EMORY L.J. 1147, 1173-74 (1994) (expressing the concern that a significant increase in the size of the federal judiciary will hurt collegiality); Harry T. Edwards, *The Effects of Collegiality on Judicial Decision Making*, 151 U. PA. L. REV. 1639, 1675 (2003) (agreeing with Judge Wilkinson on the collegiality point).

31. See sources cited *supra* note 8.

that would indicate whether judicial involvement in the writing process has decreased in direct correlation with the increase in caseloads over the past several decades.

Court administration can also be helped in other ways. Along with the explosion in caseload, there has been an increase in the complexity and volume of many areas of federal regulation.³² In order to tackle this increased complexity and volume, judges may be focusing more attention on certain types of cases while delegating others. If certain types of cases—for example, immigration, social security, securities regulation, or habeas cases—are consistently delegated to the clerks or the staff attorneys (as measured through a low judge self-authorship score), such evidence supports the argument that these specific types of cases deserve a separate set of specialized courts with expert judges, much like the U.S. Tax Court.³³

4. *Research on Judicial Behavior*

Perhaps the most immediate application of information on authorship rates is improvement of the quality of research on judicial behavior. Authorship rates can be used as an explanatory variable in regressions that seek to explain a large number of outcome variables.

32. On the emergence of the regulatory state and the accompanying enormous expansion in federal regulation (and judicial interpretations of those regulations), see Cass R. Sunstein, *Interpreting Statutes in the Regulatory State*, 103 HARV. L. REV. 405 (1989).

33. We are not aware of any systematic research into the question of whether there is a greater use of clerks in some substantive areas of the law, although it strikes us that this is an important area of research. Cf. WARD & WEIDEN, *supra* note 3 (manuscript at ch. 5, at 38-39) (finding some evidence that Justices were more likely to make substantial revisions to draft opinions when particular issues were involved). There are statements by commentators that suggest that the level of clerk and staff attorney usage is much higher in unpublished opinions. See Patricia M. Wald, *The Rhetoric of Results and the Results of Rhetoric: Judicial Writings*, 62 U. CHI. L. REV. 1371, 1373 (1995) (noting that clerks—not judges—generally draft the opinions); Tony Mauro, *Difference of Opinion*, LEGAL TIMES, Apr. 12, 2004, at 1, 10 (quoting Judge Kozinski on the matter). If it turns out that certain categories of cases are more likely to be the subject of unpublished opinions, it should follow that clerk and staff attorney usage will be higher in this category of cases. So, for example, if ninety percent of prisoner or immigration appeals result in unpublished opinions, it may well be that the major portion of lawmaking in these areas is done by law clerks and not judges. This is an outcome that strikes us as both unsatisfactory and calling out for reforms such as specialization. For a trenchant critique of the current practices of issuing unpublished opinions, see Penelope Pether, *Inequitable Injunctions: The Scandal of Private Judging in the U.S. Courts*, 56 STAN. L. REV. 1435 (2004). On the matter of whether the federal courts should specialize more in terms of subject matter, see Jeffrey W. Stempel, *Two Cheers for Specialization*, 61 BROOK. L. REV. 67 (1995).

Along the lines mentioned above, an empirical study by one of the Symposium participants, James Brudney, finds, among other things, that Supreme Court Justices seem to apply statutory canons of construction very differently in cases that arise in the more complex and less high-profile areas of labor law than they do in what might be seen as the more sexy areas. See James J. Brudney & Corey Ditslear, *Canons of Construction and the Elusive Quest for Neutral Reasoning*, 58 VAND. L. REV. 1, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=534982.

Among these outcome variables that researchers might be interested in better explaining are voting patterns, citation rates and styles, invocation rates,³⁴ publication patterns, independence levels,³⁵ and choices about styles of argument (for example, whether one prefers the use of multifactor balancing tests). It may be that whether a judge writes a significant portion of her opinions is a positive or negative explanatory factor for one or more of these outcome variables. And that information will advance our knowledge of judicial behavior. For example, it has been suggested that there are certain types of cases (perhaps those raising important constitutional questions) where judges are more likely to be actively involved in the writing and editing of an opinion than other types of cases such as tax, securities, and bankruptcy.³⁶ If authorship rates for different opinions were known, this suggestion could be tested. And, as noted earlier, if it turned out that certain types of cases were consistently delegated to the clerks, this might be reason enough to think about specialized courts for those cases. More broadly, if higher levels of authorship predict higher citation rates (and if citation rates are considered a good proxy for opinion quality), that might suggest that judges should be given incentives to write more of their own opinions. If, instead, regressions reveal the reverse effect, judges might be encouraged to write fewer opinions. Alternatively, if citation rates were seen as an inadequate measure of judicial output quality, one could use whatever other measure (perhaps a survey of experts) and then observe whether higher-quality argumentation or explanation in judicial opinions is related to self-authorship rates.³⁷ The point is that we will not really appreciate the value of knowing authorship rates

34. Invocations involve situations where one judge invokes the name of another judge when discussing an opinion written by that other judge. Choi & Gulati, *supra* note 5, at 58. We consider invocations as a special sign of respect on the part of one judge for another. *See id.* at 58-59 (discussing the concept of invocations).

35. One way of measuring judicial independence is to examine whether a judge tends to side more systematically with judges nominated by a President of the same political party as the judge in question. *See id.* at 63. We analyze circuit court judges, active from 1998 to 2000, based on this measure of independence in Choi & Gulati, *supra* note 5, at 61-67.

36. *See* WARD & WEIDEN, *supra* note 3 (manuscript at ch. 5, at 38) (finding that while most clerks reported that the Justices did not seem to have a particular pattern of deciding whether to make edits on their drafts, there is some evidence that Justices only substantially revised their draft opinions when particular issues were involved or when the cases were landmark or important cases).

37. The claim is often made that the discipline of writing out an argument is what enables the reader to test the logic of the argument. *See* Posner, *supra* note 23. Law clerks, who have been delegated the task of writing an opinion based on an argument that the judge has suggested, are perhaps less likely to second-guess the argument than the judge herself. Alternatively, some might think that law clerks are more likely to second-guess the logic of the argument and point out flaws. Once again, we do not know the answer. Knowing the information as to which opinions were solely authored versus which ones were not, however, might help us investigate further.

until we collect the information and run empirical tests on that information. Finally, assuming the information turns out to be useful, the authorship data might also be used as an outcome variable itself, and models could be constructed and tested to determine what factors determine—or at least correlate with—authorship rates.

5. *Information for Law Clerks*

Information about judicial writing proclivities will undoubtedly prove useful to law students applying for clerkships. The nature of a clerkship experience is a direct function of what the judge does. If the judge researches and writes all of her own opinions, there is little left for the clerks to do. If the judge is off skiing or at the beach for most of the year, only calling in on occasion, there will be an immense amount for the clerks to do and correspondingly little guidance from the judge.

Clerkships provide value to potential clerks in at least two ways: training and status. Law clerks are likely to receive minimal training if their judges are at the beach and have minimal contact with them. Clerks' training is also likely to be minimal if their judges are so capable that they do all of their own research and writing.³⁸ What the clerk who is seeking training wants, then, is the judge whose approach lies between the two extremes. Other clerks, who are seeking status and the power to shape the law, might want to look for a judge who delegates the major portion of opinion-writing discretion to the clerks. Clerks who primarily value leisure might prefer judges who need no help. Effective information about judges and their proclivities will assist prospective clerks in choosing judges who best suit their abilities and desires.³⁹

38. For example, clerkships with Justice Douglas were reputed to be among the worst because, at least in part, he did all his own writing and generally kept his clerks at a distance. See WARD & WEIDEN, *supra* note 3 (manuscript at ch. 5, at 8) (reporting on interviews with former clerks for Justice Douglas).

39. Internet discussion fora, such as the Greedy Clerks list serve, have at least begun to facilitate the sharing of such information among current and former clerks. See FINDLAW'S INFIRMATION, GREEDY CLERKS BOARD, at <http://www.infirmation.com/bboard/clubs.tcl?topic=Greedy%20Clerks> (last visited Feb. 10, 2005). The question whether a particular judge writes her own opinions has been discussed in at least one thread, which reported that Judges Posner, Easterbrook, Garza, and Niemeyer are among those federal circuit judges most likely to author their own opinions. See Posting of Lurks on Sept. 20, 2004 (message no. 25762) (naming Garza), at <http://www.infirmation.com/bboard/clubs-fetch-msg.tcl>; Posting of Lurks on Sept. 20, 2004 (message no. 25768) (naming Posner), at <http://www.infirmation.com/bboard/clubs-fetch-msg.tcl>; Posting of Publius Rex on Sept. 20, 2004 (message no. 25769) (naming Niemeyer), at <http://www.infirmation.com/bboard/clubs-fetch-msg.tcl>; Posting of CTA7alum on Sept. 20, 2004 (message no. 25793) (naming Easterbrook), at <http://www.infirmation.com/bboard/clubs-fetch-msg.tcl>.

But will a more effective process for matching clerks to judges bring social benefits? In terms of efficiency, it should.⁴⁰ Clerks who desire to have a greater input in the writing process will end up with judges willing to allow their clerks to do so. Likewise, clerks more interested in learning from a judge who actively manages opinion writing will be better able to find such judges. Whether such efficiency gains are great in magnitude and result in better judicial opinions, however, is unclear. To the extent the best clerks always opt for judges with a reputation as “feeder judges” for Supreme Court clerkships, information on authorship will not change the market much (unless it indirectly affects which judges are the feeder judges).

If clerks, in fact, prefer to work for judges who tend to author their own opinions or are otherwise more active in the judicial process, an indirect incentive effect on judges is possible. Judges would be forced to get more involved or accept lower-quality clerks. Alternatively, if judges’ reputations for providing inadequate training hurt their ability to get good law clerks, they might engage in more training activities for clerks.

6. *Informational Benefits to Opinion Users*

Knowing an author’s identity serves as a useful shortcut in a variety of settings. For example, in a bookstore, one might use authors’ names to decide which books to scrutinize and which ones to ignore. Similarly, when one conducts academic research, there are some authors who can be depended on for high-quality work and are therefore likely to be read first. Shortcuts such as an author’s identity are especially useful when the reader does not have the time or ability to read all of the available material and evaluate it piece by piece.

Lawyers and judges are faced with this dilemma on a daily basis. In attempting to construct their arguments, they have to draw from a vast body of available precedent and decide which arguments to use and which ones to reject or ignore. If there are multiple opinions that relate to a particular issue and there is but a limited amount of time to allocate to the task of analysis, a choice has to be made about which of those opinions should receive more attention. One of the variables used to sort through opinions is likely to be author identity. The dictates of formal precedent aside, opinions by judicial superstars like Hand, Friendly, and Cardozo will probably receive close scrutiny.

But what if it were known that while Judge Friendly wrote all his securities law opinions, he delegated his social security opinions to

40. For research into improving the judge-clerk matching process that focuses on aspects of the market other than the one that we mention, see Christopher Avery et al., *The Market for Federal Judicial Law Clerks*, 68 U. CHI. L. REV. 793 (2001).

his law clerks? Or, what if Judge Easterbrook was known to write all his own opinions and Judge Posner was known to write none of his? Presumably, that would make a difference as well, with opinions authored solely by the judge receiving greater attention and those that were not receiving less attention.⁴¹ To push the argument further, it would help outsiders decide which opinions to concentrate on if the opinions were sorted based also on the various coauthors. So, if there were an opinion on the Second Circuit for which Judge *X* was the primary author, but to which Judge Hand had made a sufficient contribution to be named coauthor, the opinion might receive greater authority than if it were to appear under Judge *X*'s name alone (assuming, for purposes of this point, that Judge *X* did not have a high reputation). Pushing even further, law clerks could be identified as coauthors or even primary authors when their contributions so warranted. The assumption would still be that the primary decision was made by the judge. But it would help outsiders who were attempting to determine how much weight to give a particular opinion to know who actually wrote the opinion.

Greater information on authorship may also add more credibility to—or alternatively quash—rumors on the supposed influence of law clerks in the writing process. There are rumors that Justice Thurgood Marshall wrote very few of his own opinions and that he wrote fewer and fewer as he advanced in age.⁴² There are rumors that Jus-

41. Justice Brandeis famously said that the reason why the Supreme Court's reputation was so high was that everyone knew that the Justices did all of their own work (a fact that is undoubtedly not true any longer). See WILLIAM O. DOUGLAS, *THE COURT YEARS 1939-1975*, at 172-73 (1980) (describing Justice Brandeis' views on the need for judges to author their own opinions). Judge Posner, in turn, explains:

The less that lawyers and especially other judges regard judicial opinions as authentic expressions of what the judges think, the less they will rely on judicial opinions for guidance and authority. . . . The more the thinking embodied in opinions is done by law clerks rather than by judges, the less authority opinions will have.

RICHARD A. POSNER, *THE FEDERAL COURTS: CRISIS AND REFORM* 110 (1985).

42. See Paul J. Wahlbeck et al., *Ghostwriters on the Court?: A Stylistic Analysis of U.S. Supreme Court Opinion Drafts*, 30 AM. POL. RES. 166, 172 (2002) (describing the rumors regarding Justice Marshall's relative lack of involvement in the opinion-writing process). The public articulations of these rumors are contained in BOB WOODWARD & SCOTT ARMSTRONG, *THE BRETHERN: INSIDE THE SUPREME COURT* 258 (1979), and Terry Eastland, *While Justice Sleeps*, NAT'L REV., Apr. 21, 1989, at 24. See also Peter Huber, *Advice to Justice Thomas*, FORBES, Nov. 25, 1991, at 202 (finding that Justice Marshall's opinions during the 1990 term demonstrated four distinctive styles, corresponding to his four different law clerks). On the other side, those familiar with the workings of the Court and, in particular, Marshall's chambers, have disputed the view articulated above. See JUAN WILLIAMS, *THURGOOD MARSHALL: AMERICAN REVOLUTIONARY* 370 (1998) (quoting both a close friend and former law clerk of Marshall); Mark Tushnet, *Thurgood Marshall and the Brethren*, 80 GEO. L.J. 2109, 2112 (1992) (stating that while Marshall may have relied "more heavily on his law clerks for opinion writing than did the other Justices during the early years of his tenure, . . . his practices were not wildly out of line with those of the others on the Court").

tice Kennedy's opinion in *Planned Parenthood v. Casey*⁴³ was the product of undue influence from one of his liberal clerks.⁴⁴ And judges like William Douglas, Learned Hand, and Richard Posner are reputed to have written all their own opinions.⁴⁵

These rumors may unfairly hurt or benefit a judge's reputation, effects that are especially problematic when some of the rumors are driven by stereotypes based on race and gender. Law clerk denials and confirmations occur, but these lack credibility since the law clerks have an incentive to do everything to heighten the reputation of their judge, because their own status is tied to the status of their judge.⁴⁶ If, however, authorship can be determined using credible and verifiable methods, these rumors can be quashed or confirmed.

B. Downsides

When we began this project, we saw few disadvantages to collecting authorship information on judges. More information, we assumed, was a good thing. But a number of our colleagues, seeing our inquiry as problematic, disagreed. Their prime objection was that the venture was a waste of our time and resources because information about judicial authorship, even if it could be obtained, was useless—an objection that we hope Part II.A has answered. Below, we tackle two of the other objections we heard most often.

1. Danger of Unjustified Inferences

One could argue that an implicit message exists in just our attempt to test the degree of self-authorship among judges. Judges who do not author their own opinions will feel unfairly stigmatized, the argument goes, when it is not clear that they should. Judging is supposed to be about applying the law to the facts in an impartial and considerate manner. Whether the articulation of that application is

43. 505 U.S. 833 (1992).

44. See Richard Lacayo, *Inside the Court*, TIME, July 13, 1992, at 29; Edward Lazarus, *Disturbing Truths*, Jurist: Books-on-Law (July 1998), at <http://jurist.law.pitt.edu/lawbooks/revjul98.htm>. The claim that liberal law clerks have had excessive influence on the Justices was also famously made years earlier by Justice Rehnquist (prior to his joining the Court). See Wahlbeck et al., *supra* note 42, at 167. The controversy that ensued led one conservative Senator to ask for clerk confirmation by the Senate. *Id.*

45. On Learned Hand's use of his law clerks in the opinion-writing process, his former law clerk, Gerald Gunther, wrote: "No clerk for [Learned] Hand ever wrote a single word, either in producing research memoranda or in drafting opinions. Instead, the Hand-law clerk relationship was one of extraordinary intellectual intimacy: it consisted entirely of face-to-face contacts, not any written work." Gerald Gunther, *Reflections on Judicial Administration in the Second Circuit, from the Perspective of Learned Hand's Days*, 60 BROOK. L. REV. 505, 510 (1994). On Posner, see *supra* note 6. And on Douglas, see Richard A. Posner, *The Anti-Hero*, NEW REPUBLIC, Feb. 24, 2003, at 27 (reviewing BRUCE ALLEN MURPHY, *WILD BILL: THE LEGEND AND LIFE OF WILLIAM O. DOUGLAS* (2003)).

46. See *supra* note 42 (citing statements by former Marshall clerks).

self-authored or not is a consideration of minimal value, at best.⁴⁷ The identity of the author does not matter, and for us to suggest otherwise by making this inquiry is disrespectful.

We disagree. One may not want judicial or author identity to be relevant, but there is reason to think that it is. Judges have individual preferences and respond to incentives. There is debate about the degree to which judges, as a specific subgroup, are driven by these factors, as opposed to other factors such as the norms of their profession or altruism toward society.⁴⁸ Given that debate, the necessary next step should be to test the robustness of the competing models, and that requires collecting information on variables such as authorship.⁴⁹ It may be that the empirical results will pleasantly surprise our critics and authorship identity will turn out not to matter. And if that is the case, there will not be any stigma. On the other hand, if self-authorship is related to productivity, quality, and other factors of importance, then a stigma should arguably fall on those judges who do not author their own opinions.

2. *Imperfect Measurement*

A second criticism takes aim at our methodology. Our critics point out that the types of statistical tests that we describe later in the Essay can at best produce an imperfect measure of authorship rates. Given that judges themselves have perfect information about their authorship levels, why not simply ask them about their practices? The point is fair. Judges do know more about their own practices, and we should try to ask them about what they do. But doing that alone is unlikely to be sufficient for two reasons. First, judges may choose simply not to respond to a survey instrument, especially if it is from some annoying law professor. Indeed, judges seemed irate when requested to respond to a survey sent out by the Senate's Judi-

47. John Orth makes this point in his critique of our prior article where he says that any evaluation of judges should focus first on the quality of decisions (or justice) from the perspective of the individual litigants and only then, if at all, on considerations about the quality and style of the articulation. See Orth, *supra* note 15.

48. See Smyth, *supra* note 14, at 1302-09 (describing the debate).

49. There is already a literature on testing the different models of judging in law, economics, and political science, with contributions to that literature having been made by participants in this Symposium, including Brudney, Epstein, Farber, Posner, Smyth, Staudt, and Taha. See, e.g., James J. Brudney, *Foreseeing Greatness? Measurable Performance Criteria and the Selection of Supreme Court Justices*, 32 FLA. ST. U. L. REV. 1015 (2005); Lee Epstein et al., *The Role of Qualifications in the Confirmation of Nominees to the U.S. Supreme Court*, 32 FLA. ST. U. L. REV. 1145 (2005); Daniel A. Farber, *Supreme Court Selection and Measures of Past Judicial Performance*, 32 FLA. ST. U. L. REV. 1175 (2005); Richard A. Posner, *Judicial Behavior and Performance: An Economic Approach*, 32 FLA. ST. U. L. REV. 1259 (2005); Russell Smyth, *supra* note 14; Ahmed E. Taha, *Information and the Selection of Judges: A Comment on "A Tournament of Judges,"* 32 FLA. ST. U. L. REV. 1401 (2005). Our point is that information about authorship rates has the potential to improve those tests.

ciary Oversight Subcommittee.⁵⁰ Second, there is a verifiability problem with judges self-reporting their practices. What's to stop a judge from saying that she writes all her opinions (whether true or not)? Judges could ameliorate the verifiability problem somewhat by allowing their clerks free reign to disclose delegation practices. But we think this unlikely to happen.

Imperfect measures using computational techniques might be able to partially solve these problems by creating both an incentive to self-report and a verifiability mechanism. In other words, the revelation of partial information by an outside source may create an incentive for the possessor of the true information to fully reveal it, if only to correct the imperfect impression left from objective data on authorship.⁵¹ If the imperfect information is seen as having a high degree of reliability (for example, it is correct seventy percent of the time), then that will put pressure on judges who rank low on the imperfect measure, but whose true scores should be higher, to self-report their true practices. And to the extent there is a perceived verifiability problem with self-reporting, the judges may allow for verifiability through mechanisms like clerk disclosure.

III. TESTING AUTHORSHIP

So far, we have assumed that authorship identity could be determined. In theory it can be, but the task is not straightforward. Indeed, prior attempts at surveys on this very information have revealed little other than the fact that many judges do delegate some portion of the writing task to their clerks.⁵² In this Part we describe

50. See *infra* note 52 (discussing the survey attempt). The Grassley survey may not be the best example, though. The judges' annoyance at Senator Grassley's attempt to survey them may have been more a product of the somewhat undiplomatic tone and context of that particular survey attempt. See *infra* note 52.

51. For more on the argument that objective tournaments may help force participants to reveal information about themselves, see Scott Baker, Stephen Choi & Mitu Gulati, *The Rat Race as an Information Forcing Device*, 81 IND. L.J. (forthcoming 2006), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=649083.

52. There was a survey attempt directed by Senator Charles Grassley in 1996 as part of his work as chair of the Senate-Judiciary Subcommittee on Administrative Oversight and the Courts. See U.S. SENATE-JUDICIARY SUBCOMM. ON ADMIN. OVERSIGHT & THE COURTS, 104TH CONG. REPORT ON THE JANUARY 1996 JUDICIAL SURVEY (PART I, U.S. COURTS OF APPEAL) 26-27 (Comm. Print 1996) [hereinafter SURVEY REPORT]. He sent a survey to all sitting federal judges that inquired about matters such as delegation to law clerks. See *Now the Judges Face the Questions*, LEGAL TIMES, Feb. 5, 1996, at 8 (providing the full text of the Grassley questionnaire). The survey results suggested that most judges did not perceive an inappropriate level of delegation. See SURVEY REPORT, *supra*, at 26-27. (reporting that over seventy-five percent of the circuit court judges did not perceive a problem with the extent of delegation to law clerks); *Oyez Surveys*, LEGAL TIMES, Aug. 12, 1996, at 3 (reporting that Grassley's "results show that most judges believe that law clerks do not play too large a role in judicial decision making"); Deborah Pines & Bill Alden, *District, Circuit Judges Use Senate Survey to Boast, Gripe*, N.Y. L.J., Mar. 25, 1996, at 1, 4 (reporting that Judge Newman responded to the question of how much he delegated work to law

our preliminary attempts to determine authorship through other techniques.

The literature on the question of tracing authorship dates back to over a century ago. In recent years, these techniques have made it onto television talk shows and the front page of *The New York Times*.⁵³ Stories have appeared about their application in the discovery of a new Shakespeare poem and in a variety of other high-profile settings, such as unmasking the anonymous author of *Primary Colors*.⁵⁴ Law enforcement agencies and prosecutors have also used “forensic linguistics” in both the JonBenét Ramsey murder investigation and the Unabomber prosecution.⁵⁵

In attempting to answer these questions and others—such as whether Shakespeare wrote his plays or if they were really the work of Marlowe, Bacon, or some other contemporary—scholars have de-

clerks by saying that he does “not believe that law clerks exercise anything that can reasonably be called ‘power’”).

The survey caused considerable displeasure among many judges. *See, e.g.*, Bruce Brown, *Grassley Has Judges Grousing*, AM. LAW., Mar. 1996, at 16 (reporting that “Grassley’s crusade has many judges grumbling”); Pines & Alden, *supra*, at 4 (quoting Chief Justice Rehnquist, who characterized the survey as potentially “an unwarranted and ill-considered effort to micromanage the work of the federal judiciary,” and stating that the survey’s inquiry into delegation practices particularly had “raise[d] the ire of judges”). Given this displeasure, the response rate of over fifty percent was surprisingly high—approximately 600 judges returned the surveys. *See* Pines & Alden, *supra*, at 4 (reporting that over 600 out of 1148 judges had responded to the survey). Despite the high response rate, neither this nor other survey attempts have come anywhere close to estimating relative levels of clerk contributions in the different chambers in a meaningful manner. *See* Wahlbeck et al., *supra* note 42, at 168 (making this point about prior survey attempts).

53. *See* William H. Honan, *A Sleuth Gets His Suspect: Shakespeare*, N.Y. TIMES, Jan. 14, 1996, § 1, at 1 (discussing Professor Donald Foster’s efforts to trace authorship).

54. Much of the media scrutiny regarding what is sometimes referred to as “forensic linguistics” has centered around the work of Vassar College professor Don Foster. Many of Foster’s exploits, including those mentioned in the text, are detailed in DON FOSTER, AUTHOR UNKNOWN: ON THE TRAIL OF ANONYMOUS (2000). For more on Foster’s exploits, including *The New York Times* front-page story, his unmasking of Joe Klein for *New York Magazine*, and his confrontation with Dan Rather on CBS, see Jamie Allen, *On the Trail of a Literary Sleuth*, CNN.COM, Dec. 6, 2000, at <http://archives.cnn.com/2000/books/news/12/06/foster.anonymous/>. The discovery of the Shakespeare poem that brought Foster his initial fame, though, was later recanted by Foster. *See* William S. Niederkorn, *A Scholar Recants on His ‘Shakespeare’ Discovery*, N.Y. TIMES, June 20, 2002, at E1. Stronger evidence pointing to John Ford as the author of the poem, *A Funeral Elegy*, was subsequently reported by Professor Gilles Monsarrat. *Id.* Foster’s error, apparently, was to focus excessively on word usage patterns and not enough on structural features such as phrase patterns. *See id.* (reporting criticisms by Cambridge Professor Brian Vickers).

For more sophisticated linguistic treatments of the arguments over the authorship of *Funeral Elegy*, see BRIAN VICKERS, ‘COUNTERFEITING’ SHAKESPEARE: EVIDENCE, AUTHORSHIP, AND JOHN FORD’S FUNERALL ELEGY (2002); Ward Elliott & Robert Valenza, *Smoking Guns and Silver Bullets: Could John Ford Have Written the Funeral Elegy?*, 16 LITERARY & LINGUISTIC COMPUTING 205 (2001); and G.D. Monsarrat, *A Funeral Elegy: Ford, W.S., and Shakespeare*, 53 REV. ENG. STUD. 186 (2002).

55. *See* Allen, *supra* note 54 (summarizing Foster’s help to the FBI); FOSTER, *supra* note 54, at 16-17 (providing a brief discussion of the Ramsey murder investigation); *id.* at 95-142 (explaining Foster’s contributions to the Unabomber investigation).

veloped a number of techniques.⁵⁶ We make use of these techniques in discerning authorship for judicial opinions. Although some linguistics tools have found their way into FBI investigations and even into court,⁵⁷ few legal academics have made any meaningful use of them in their research on judicial authorship.⁵⁸ One reason for the lack of research is resource-related.⁵⁹ In addition, determining authorship is difficult without a set of authentic texts for each judge (for example, texts where it is known to a certainty that the judge is the author) against which to compare the judicial opinions bearing the judge's name.

There has been, however, at least one attempt to determine authorship in the judicial context. Paul Wahlbeck, James Spriggs, and Lee Sigelman used techniques from computational linguistics to determine the relative levels of delegation of the opinion-writing task to clerks by Justices Lewis Powell and Thurgood Marshall.⁶⁰ We consider whether some version of that method could be used to determine authorship patterns for a larger sample of judges.

To get a sense of the difficulty of tackling this question and as a preliminary test of the effectiveness of standard authorship methodologies in determining judicial opinion authorship, we ran our authorship tests on a sample consisting of opinions for all the active federal circuit court judges for the period from January 1, 1998 to December 31, 2000—a total of three years worth of data.

Our tentative conclusion: the ranking task can be accomplished, but it requires a significant allocation of resources—at least more than the minimal amounts that law faculty typically use on their pro-

56. Inquiries into the authorship of Shakespearean texts have been multitudinous. For a discussion of some examples, see Barron Brainerd, *The Computer in Statistical Studies of William Shakespeare*, 4 COMPUTER STUD. HUMAN. & VERBAL BEHAV. 9 (1973), and C.B. Williams, *Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon*, 62 BIOMETRIKA 207 (1975). For more on inquiries into the true authorship of the texts we attribute to Shakespeare, see Ward Elliot, *The Shakespeare Clinic*, ELLIOT ONLINE, at <http://govt.mckenna.edu/welliott/shakes.htm> (last visited Feb. 5, 2005), and THE SHAKESPEARE AUTHORSHIP PAGE, at <http://shakespeareauthorship.com> (last visited Feb. 5, 2005).

57. See FOSTER, *supra* note 54, at 95-142 (describing the use of his techniques by both the FBI and the prosecutors in the Unabomber case); see also Bryan Niblett & Jillian Boreham, *Cluster Analysis in Court*, 1976 CRIM. L. REV. 175 (describing how the "cluster analysis" technique is used to verify criminal confession statements).

58. We should note though that at least two sets of legal scholars have discussed forensic linguistics in other contexts. See LAWRENCE M. SOLAN & PETER M. TIERSMA, *SPEAKING OF CRIME: THE LANGUAGE OF CRIMINAL JUSTICE* 149-80 (2005) (discussing the application of authorship attribution research to a wide variety of criminal investigations and cases); JOHN M. CONLEY & WILLIAM M. O'BARR, *JUST WORDS: LAW, LANGUAGE, AND POWER* 161-78 (2d ed. 2005) (discussing Donald Foster's involvement in the JonBenét Ramsey murder investigation).

59. See *infra* text accompanying notes 86-88 (describing time-intensive nature of our authorship methodology).

60. See Wahlbeck et al., *supra* note 42, at 170-73.

jects. Moreover, significant “noise” may exist in comparing authorship of different judicial opinions that may make determining the true level of authorship for a particular set of opinions for any particular judge difficult. Nonetheless, the same reasons we offer in Part II.A of this Essay to support the value of determining authorship lead us to believe that more research is warranted in the area.

In Part III.A, we give a thumbnail sketch of the methodology used for doing authorship testing. In Part III.B, we then examine the question of how that methodology might be applied to the matter of ranking judges, and we provide results from generic tests of authorship drawn from computational linguistics. The generic tests perform poorly in distinguishing judges based on authorship. In Part III.C, we then report the results from better-tailored tests of authorship for the judicial context. Tests that control for the subject matter of judicial opinions perform better in assessing judicial authorship.

A. *Tracking Judicial Fingerprints*

The basic proposition here is that writers have styles of their own. Just as all of us have our own styles of walking, talking, singing, shooting a photograph or movie, throwing a baseball, and playing a guitar, we also have particular writing styles. Some of us have styles that are more distinctive than others. This is especially true when we are experts in a field and, as a result, are extensively active within the field. Michael Jordan’s style of shooting a basketball, Serena Williams’ style of serving a tennis ball, and Sachin Tendulkar’s style of hitting a cricket ball are so distinctive that even casual fans are likely to identify the players from just their playing styles. Our more literate friends can readily recognize passages from authors such as Jane Austen, Ernest Hemmingway, and F. Scott Fitzgerald.

Testing authorship by way of an author’s idiosyncratic style requires distilling the basic elements of the style that differ from that of other authors in order that these basic elements can be used as identifiers. For example, if one knows that author *A* is partial to using the word “hath” and intensely dislikes the word “have,” one might think it unlikely a document that has zero uses of “hath” and multiple uses of “have” belongs to him. The information about author *A*’s distinctive style characteristics is combined with information about the patterns in the document of unknown authorship, and Bayes’ Theorem is used to determine the probability that *A* authored the mystery text.⁶¹

61. The standard technique to make these probability calculations is to use Bayes’ Theorem. One of the problems with the use of the naïve Bayes model that has been pointed out is that the calculation assumes randomness (and, of course, words in a text are a function of each other as opposed to completely independent and random). Nevertheless, despite the

The problem with this technique is that one has to identify a large set of unique identifiers (such as uses of “hath” in place of “have”). Without such a set of identifiers, it becomes difficult to distinguish authors from one another. Relying on unique stylistic markers also requires expertise in identifying the author’s style and will work only for authors who have developed identifiable traits in the first place. This narrows considerably the usefulness of such techniques. Not all authors use unique stylistic markers in their writing (Choi and Gulati, for example, are quite mundane in their styles, although they are both partial to text parentheticals).

More broadly applicable techniques for authorship determination build on the same basic idea but look to a variety of style patterns rather than particular identifiers. Here, the premise is that authors’ writings follow patterns. These patterns may be, in part, a function of stylistic preferences, such as the use of the word “hath.” Patterns may also consist of the use of a series of more common words and the order in which such words are used. Rather than look for specific styles, authorship may be determined through an examination of the overall frequency of specific words and patterns of words throughout the entire text.

Word choice patterns (that is, diction) are also likely to be a function of bounded rationality. People engaged in the task of writing, just as with other tasks, are constrained by their cognitive capacities. There is likely to be a finite set of words that are part of a single individual’s active vocabulary. One uses these words more often than other words. Other words, though part of one’s vocabulary, are used only rarely in one’s own speech or writing yet are immediately recognizable in both the text and speech of others. And then there will be words the author simply does not know. Individual authors are likely to exhibit distinct patterns of noun and verb usage, distinct patterns

likely violation of the randomness assumption, the naïve Bayes model appears to perform well. For discussions of Bayes’ Theorem, see James Joyce, *Bayes’ Theorem*, THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY, at <http://plato.stanford.edu/entries/bayes-theorem> (last substantive content change Sept. 30, 2003). For a Bayes’ Theorem calculator, see VASSARSTATS, BAYES’ THEOREM: CONDITIONAL PROBABILITIES, at <http://faculty.vassar.edu/lowry/bayes.html> (last visited Feb. 7, 2005). On the use of the Bayes model, see FREDERICK MOSTELLER & DAVID L. WALLACE, APPLIED BAYESIAN AND CLASSICAL INFERENCE: THE CASE OF THE FEDERALIST PAPERS (2d ed. 1984). See also Sang-Bum Kim et al., *Effective Methods for Improving Naïve Bayes Text Classifiers*, PRICAI 2002: TRENDS IN ARTIFICIAL INTELLIGENCE: 7TH PACIFIC RIM INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, TOKYO, JAPAN (M. Ishizuka & A. Sattar eds., 2002); Andrew McCallum & Kamal Nigam, *A Comparison of Event Models for Naïve Bayes Text Classification*, in AM. ASS’N FOR ARTIFICIAL INTELLIGENCE, LEARNING FOR TEXT CATEGORIZATION: PAPERS FROM THE AAAI WORKSHOP, TECHNICAL REPORT WS-98-05, at 41 (1998); Yiming Yang, *An Evaluation of Statistical Approaches to Text Categorization*, 1 INFO. RETRIEVAL 69 (1999). More generally, for both a discussion and application of a variety of these methods of authorship recognition, see Michael G. Farringdon, *A Stylometric Analysis*, in MARTIN C. BATTESTIN, NEW ESSAYS BY HENRY FIELDING: HIS CONTRIBUTIONS TO THE CRAFTSMAN (1734-1739) AND OTHER EARLY JOURNALISM app. C, at 549-62 (1989).

for combining words, and distinct patterns of starting and ending sentences. All of these patterns can translate into frequency of use. Given a set of authentic texts for any author, a full set of these frequencies can be calculated. These frequencies can then be compared to the frequency patterns for the document whose authorship is unknown to determine whether the same author wrote the document in question.⁶²

Among the first serious treatments of the mathematics of word distributions, to our knowledge, was that by G. Udny Yule in 1938, in a study of *The Imitation of Christ*.⁶³ The defining studies in the area, however, are four studies published in the 1960s: Louis Milic's tests of Jonathan Swift's prose,⁶⁴ A.Q. Morton and James McLeman's study of the Pauline Epistles,⁶⁵ Alvar Ellegård's study of the *Junius Letters*,⁶⁶ and Frederick Mosteller and David Wallace's work on *The Federalist* papers.⁶⁷ The success of these projects and improvements of computer technology spurred an expansion of the literature on style statistics in the decades since. For interested scholars, extended treatments and a steady stream of sophisticated journal articles exist.⁶⁸

62. See Farrington, *supra* note 61 (describing a range of techniques used to determine authorship). For an overview of the state of authorship attribution studies and the wide range of techniques in use, see Joseph Rudman, *The State of Authorship Attribution Studies: Some Problems and Solutions*, 31 COMPUTERS & HUMAN. 351 (1998).

63. See Farrington, *supra* note 61, at 549-50; see also G. UDN YULE, THE STATISTICAL STUDY OF LITERARY VOCABULARY (1944) (investigating the measurement of word distribution). Prior to Yule, there is the suggestion that stylistic habits might be used to detect authorship in a letter from the mathematician Augustus De Morgan in 1851. See Farrington, *supra* note 61, at 549 (citing SOPHIE ELIZABETH DE MORGAN, MEMOIR OF AUGUSTUS DE MORGAN 215-16 (1882)). There is also the work on frequency distributions of words by Zipf in 1932. See GEORGE KINGSLEY ZI PF, SELECTED STUDIES OF THE PRINCIPLE OF RELATIVE FREQUENCY IN LANGUAGE (1932).

64. LOUIS TONKO MILIC, A QUANTITATIVE APPROACH TO THE STYLE OF JONATHAN SWIFT (1967).

65. A.Q. MORTON & JAMES MCLEMAN, PAUL, THE MAN AND THE MYTH: A STUDY IN THE AUTHORSHIP OF GREEK PROSE (1966).

66. ALVAR ELLEGÅRD, A STATISTICAL METHOD FOR DETERMINING AUTHORSHIP: THE JUNIUS LETTERS, 1769-1772 (1962).

67. MOSTELLER & WALLACE, *supra* note 61; FREDERICK MOSTELLER & DAVID L. WALLACE, INFERENCE AND DISPUTED AUTHORSHIP: THE FEDERALIST (1964).

68. Among the extended treatments or collections of articles are J.F. BURROWS, COMPUTATION INTO CRITICISM: A STUDY OF JANE AUSTEN'S NOVELS AND AN EXPERIMENT IN METHOD (1987); THE COMPUTER IN LITERARY AND LINGUISTIC STUDIES (Alan Jones & R.F. Churchhouse eds., 1976); PAULA R. FELDMAN & BUFORD NORMAN, THE WORDWORTHY COMPUTER: CLASSROOM AND RESEARCH APPLICATIONS IN LANGUAGE AND LITERATURE (1987); ANTHONY KENNY, THE COMPUTATION OF STYLE: AN INTRODUCTION TO STATISTICS FOR STUDENTS OF LITERATURE AND HUMANITIES (1982); LITERARY COMPUTING AND LITERARY CRITICISM: THEORETICAL AND PRACTICAL ESSAYS ON THEME AND RHETORIC (Rosanne G. Potter ed., 1989); and A.Q. MORTON, LITERARY DETECTION: HOW TO PROVE AUTHORSHIP AND FRAUD IN LITERATURE AND DOCUMENTS (1978). Current articles in the field can be found in the journals *Literary and Linguistic Computing* and *Computers and*

The question, then, is whether these techniques can be applied to judicial opinions. The answer is not obviously in the affirmative.

The fact that Jane Austen's or Ernest Hemmingway's texts have distinctive styles and are easily recognizable does not mean that the writing styles of more ordinary mortals will be recognizable. Moreover, testing judicial authorship is potentially even more problematic because of the institutionalized nature of judicial writing. To the extent that judges consciously try to follow some institutional style (and, therefore, consciously suppress their own style), the difficulties in identifying particular judicial authorship are likely further exacerbated.⁶⁹ Our project, however, is to identify the degree of judge authorship versus clerk authorship.

There is reason to believe that judicial style is likely to be different from clerk style. Judges, who tend to have been experienced lawyers or academics earlier in their careers, are likely to be more confident in their writing than their clerks, who tend to be fresh out of law schools. Judges, because of their high level of skill and confidence, may write shorter opinions with fewer citations and footnotes. Judges also are likely to attack the central issue in their cases directly. Clerks, by contrast, because of their inexperience, may tend to write lengthy opinions with numerous citations and footnotes.⁷⁰

As noted earlier, the sole application of the linguistic techniques discussed above to the question of judicial delegation to law clerks is the paper by Wahlbeck, Spriggs, and Sigelman.⁷¹ Court lore has it that Justice Thurgood Marshall delegated writing tasks heavily to his law clerks, whereas Justice Lewis Powell was more actively involved in the production of opinions.⁷² Wahlbeck and his coauthors attempted to use information available from the Justices' papers,

the Humanities. See Wahlbeck et al., *supra* note 42, at 189 n.14 (suggesting these two journals and a number of the volumes mentioned above for background in statistical stylistics).

69. For a discussion of the types of judicial opinion styles and the background literature, see Robert F. Blomquist, *Playing on Words: Judge Richard A. Posner's Appellate Opinions, 1981-82—Ruminations on Sexy Judicial Opinion Style During an Extraordinary Rookie Season*, 68 U. CIN. L. REV. 651, 656-76 (2000).

70. The point that overreliance on law clerks can significantly influence the style of opinions has been made by a number of commentators. See COHEN, *supra* note 3, at 94 (making the point that delegation of opinion writing to law clerks can affect the clarity and style of the opinion); POSNER, *supra* note 41, at 115 (suggesting that the excessive numbers of footnotes, citations, and words in opinions are all a product of heightened levels of delegation to law clerks); Wahlbeck et al., *supra* note 42, at 173 (making the point that clerks are more likely than judges to rely extensively on multifactor and balancing tests in the opinions that they draft) (citing ANTHONY T. KRONMAN, *THE LOST LAWYER* (1993), and POSNER, *supra* note 3). *But see* Samuel Estreicher, *Conserving the Federal Judiciary for a Conservative Agenda?*, 84 MICH. L. REV. 569, 574 (1986) (agreeing that law clerks may be partially responsible for the excessive numbers of footnotes and the deadening style of current opinions but also stating that part of the blame may also lie with computerization).

71. See Wahlbeck et al., *supra* note 42.

72. See *id.* at 170-72; see also sources cited *supra* note 42.

containing clerk identifiers on bench memoranda and draft opinions, to see whether they could detect the “fingerprints” of the clerks in the opinions with which the two Justices were involved.⁷³ The greater the involvement of the Justice in the actual drafting, they hypothesized, the less likely it would be that the fingerprints of the clerks would be detectable.⁷⁴ In order to perform the detection task, they used eight different frequency measures.⁷⁵ These were average footnote length, average sentence length, average word length, word length diversity, sentence length diversity, footnote frequency, type-token ratio, and the once-word rate.⁷⁶ Consistent with their initial hypothesis, Wahlbeck, Spriggs, and Sigelman found that the fingerprints of judicial clerks are clearer in Marshall’s opinions than in Powell’s opinions. This implied that Powell had a greater hand in authoring his own judicial opinions.⁷⁷

We attempt to adapt the authorship methodology used in computational linguistics to the task of ranking authorship rates for judges. We note, however, several important points. Ideally, we could perform the following two step methodology:

Step 1: Run the authorship methodology on a set of authentic writing samples from a particular judge to serve as the baseline of comparison.

Step 2: Compare the judge’s judicial opinions against the authentic writing samples to determine whether the judge in fact authored the opinions. The greater the discrepancy between the opinions and the authentic writing samples, the less likely the probability that the particular judge authored the opinions.

For most judges, there is unlikely to be an available set of authentic samples of the judge’s own writing that can be used to calculate baseline frequencies.⁷⁸ We therefore are unable to determine how many of a particular judge’s opinions the judge actually wrote.

Nonetheless, at least two uses for the authorship methodology are possible in judge-related authorship studies. First, we can make a comparison of a subset of a particular judge’s opinions to each other.

73. Wahlbeck et al., *supra* note 42, at 168, 178-82.

74. *See id.* at 174.

75. *Id.* at 176-77.

76. *Id.* *Type-token* ratio is “the number of different words in an opinion (types) as a percentage of the total number of words in the opinion (tokens).” *Id.* at 176. *Once-word* refers to “the relative frequency of words that appear exactly once in an opinion.” *Id.*

77. *Id.* at 182-83.

78. For a subset of judges such as former law professors, we can find articles that they authored. There are two problems with using these articles as the basis for opinion testing. First, article writing is a different genre from opinion writing; and second, law review editors are involved in article writing. That said, since it is these same law review editors who often become law clerks, maybe opinion styles and law review article styles are more similar than we think.

Assuming that consistency in the sentence patterns and other aspects of an author's style is positively correlated with authentic authorship, one can then rank the judges *relatively* against one another. If the judge authored all the opinions, presumably the opinions will receive a high same-authorship score. If the judge did not author all of the opinions, then the set of opinions will receive a low same-authorship score.

This variation measure across a set of opinions for which a particular judge is named as the author, it should be cautioned, will provide a high score for a judge who delegates all her work to a single, permanent law clerk.⁷⁹ It would also likely give higher scores to judges who imposed their styles on the writing process by doing heavy editing as opposed to doing any of the actual writing. What we would be measuring, then, is best described as the relative degree of various judges' involvement in the writing process.

Second, we can use the authorship methodology to make relative comparisons of how authorship for a particular judge has changed over time. We may not know if Judge A authors her opinions or not. But we can see if the level of authorship (regardless of the starting point) has declined or increased over time. Consider Supreme Court Justices. One hypothesis is that as individual Justices age, they play a diminishing role in writing their own opinions or, alternatively, in monitoring the work product of their clerks.⁸⁰ We could compare the authorship score for a particular Justice when in her fifties against the authorship score when in her eighties. If the hypothesis is correct, one would expect to see a relative drop in the same-authorship score.

In this Essay we discuss only the first application of the authorship methodology to ranking judges (the judge versus judge comparison). Doing a full-scale ranking of any meaningful sample of judges is beyond the scope of this project—and our research budgets. We do, however, briefly report the results of some preliminary tests and discuss the limitations of our methodology as well as possible adjustments to improve on our work. We leave for another paper the comparison of judges relative to their past selves.⁸¹

79. Such a result is unlikely, nonetheless, where the sample of opinions for a judge are drawn across multiple years (and thus involve multiple different sets of clerks).

80. See WARD & WEIDEN, *supra* note 3 (manuscript at ch. 5, at 11) (reporting evidence on how the authorship patterns of Justices such as Rehnquist and Blackmun changed over their time on the Court, with both Justices delegating more of the opinion writing to clerks as they gained seniority).

81. We are in the process of collecting these data now.

B. Ranking Judges

To test authorship rates, we start with a set of generic tests commonly used in other studies of authorship. We do not control for different types of documents in the generic tests but take instead a randomly selected set of opinions for a judge and compare them to one another. We label these generic tests our “black-box” tests because of the lack of subject matter controls.

The two black-box tests that we use are recognized authorship-testing methods—only two out of a variety of tests available.⁸² We describe them briefly before discussing the results of the tests. We rely primarily on the GZip compression technique, which compresses documents based on the similarities in the basic linguistic building blocks of the two files.⁸³ The GZip algorithm looks for repeated phrases within a threshold of the last 8000 characters of text analyzed (a number that can be made larger, depending on the specific program). The longer and more frequent the repeated phrases, the higher the score the algorithm accords to the text; the end result is greater compression.

82. Among the many authorship attribution techniques are reduction of authorship style to a single defining number, Markov chains, cumulative sums, and syntactic annotation. On these various methods, see Dmitri V. Khmelev & Fiona J. Tweedie, *Using Markov Chains for Identification of Writers*, 16 LITERARY & LINGUISTIC COMPUTING 299 (2001); Michael L. Hilton & David I. Homes, *An Assessment of Cumulative Sum Charts for Authorship Attribution*, 8 LITERARY & LINGUISTIC COMPUTING 73 (1993); George K. Barr, *The Cusum Mechanism—A Review of Analysing for Authorship by Jill M. Farrington*, 6 EXPERT EVIDENCE 43 (1998); Harald Baayen et al., *Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution*, 11 LITERARY & LINGUISTIC COMPUTING 121 (1996); and Tony Honoré, *Some Simple Measures of Richness of Vocabulary*, 7 ASS'N LITERARY & LINGUISTIC COMPUTING BULL. 172 (1979).

83. In using and applying the GZip method, we relied on the computer scientists at In-App software company in Trivandrum, India—Satish Babu, M.C. Jayakrishnan, and R.V. Suchithra—who consulted with colleagues at the computer science department at Kerala University to develop a simple application of the GZip method for our project. Should readers be interested in more details on the GZip application, please email questions to: Satish Babu, sb@inapp.com; M.C. Jayakrishnan, jayan@inapp.com; or R.V. Suchitra, suchi@inapp.com. Our use of the compression method to test authorship is simplistic. For discussions of far more sophisticated applications of the compression methodology to test for authorship that are beyond the mathematics and statistics skills of the two authors here, see, for example, Dario Benedetto et al., *Language Trees and Zipping*, 88 PHYSICAL REV. LETTERS 048702-1 (2002); Eibe Frank et al., *Text Categorization Using Compression Models*, in PROCEEDINGS: DCC 2000 DATA COMPRESSION CONFERENCE 555 (James A. Storer & Martin Cohn eds., 2000); D.V. Khmelev, *Disputed Authorship Resolution Through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Texts*, 7 J. QUANTITATIVE LINGUISTICS 201 (2000); O.V. Kukushkina et al., *Using Literal and Grammatical Statistics for Authorship Attribution*, 37 PROBS. INFO. TRANSMISSION 172 (2001); and William J. Teahan & David J. Harper, *Using Compression-Based Language Models for Text Categorization* (2001), available at <http://citeseer.ist.psu.edu/teahan01using.html> (last visited Feb. 11, 2005). Simphile, a program from Geneffects, is a commercial implementation of author recognition, using the same technique. The methodology has also reportedly been used for gene sequence matching in bioinformatics and authorship of music and art. For more information, see GENEFFECTS, SIMPHILE, at <http://www.geneffects.com/simphile> (last visited Feb. 8, 2005).

Compression serves as a metric for redundancy or “entropy”⁸⁴ and, as a result, is a metric for comparing authorship. The compression for a combined file containing an English document combined with an Italian one, for example, will be significantly less than that of a combined file containing two similar length documents in English.⁸⁵ Along those lines, research suggests that the compression of files from the same authors will be significantly greater than that of files from different authors (due to the greater number of repeated phrases).

We can calculate a compression score for any two documents (labeled *A* and *B*) as follows:

$$\text{Compression Score} = P/Q$$

where:

$Size(\cdot)$ = The size in bytes of the compressed file

$P = Size(\text{Document } A) + Size(\text{Document } B)$

$Q = Size(\text{Documents } A \text{ combined with } B)$

P is always greater than *Q*. If documents *A* and *B* are written by the same author, we assert that *P* will be much greater than *Q*. Put another way, if documents *A* and *B* display many similarities, compressing the combination of the two documents will result in significant space savings when compared with the compression of each document separately. On the other hand, if the documents are written by different authors, *P* will, in the extreme, approach *Q* (as the compression program will find fewer common phrases between the documents). If documents *A* and *B* are completely different, then combining the two will result in no additional compression.

For each of the ninety-eight judges in our sample, we collected four text samples of 8000 characters each. Text samples were chosen at random from opinions at least 32,000 characters in length that were written between 1998 and 2000.⁸⁶ In our black-box tests, we did not control for the subject matter of the opinion. A particular judge may have four criminal law opinions or, alternatively, a set of crimi-

84. For a discussion on the meaning of entropy in the context of authorship, see Benedetto et al., *supra* note 83, at 048702-1 to 048702-2.

85. *See id.* at 048702-2.

86. The actual selection of files to run the tests was made at random by the programmers at InApp, with no interference from us other than the request that the selection be made randomly (using a random number generator to select four opinions for each judge). The only possible bias in the selection of the opinions is in their size, in that the opinions in our selection pool had to have a minimum size of 32,000 characters (because that size would generally yield at least 8000 characters after cleaning). This was not a problem in terms of majority opinions because most judges have a large number of opinions of 32,000 characters. But when we attempted to run the program on dissenting opinions we ran into the bias problem because most judges had very few dissenting opinions of 32,000 characters or more. For that reason, we did not attempt to do a full study of the differences between dissents and majority opinions, something that could potentially yield interesting results. *See infra* note 89 (discussing how the comparison of scores on dissents versus those on majority opinions could serve as a test of the methodology).

nal, constitutional, commercial, and securities-regulatory opinions. As we discuss below, however, this lack of control for the type of opinion may introduce genre-specific “noise” in our results. The compression score for four criminal law opinions from a judge who does not write her own opinions may be greater (leading to a higher same-authorship score) than the score for four opinions from different areas of the law for a judge who does in fact write her own opinions.

Since the goal was to test authorship style, we eliminated all portions of the documents that were not a product of this style. That meant removing all West headnotes, citations, and quotes from each document. Commonly used words such as “and” and “the” were also eliminated.⁸⁷ Cleaning the documents was the most time-consuming task in this project because it had to be done manually. It took us 500-plus hours of research-assistant time to clean the pieces of text that satisfied our conditions for each of the ninety-eight judges in our sample. Cleaning the documents is not strictly required but does reduce the level of noise in the same-authorship score results, since, for example, two judges with dramatically different levels of “true” authorship may receive the same GZip score if the compression of the words “and” and “the” swamp all other differences between the documents.

We then made pairwise comparisons of these four pieces of text (labeled *A1*, *A2*, *A3*, and *A4* below) and generated a four-by-four matrix for each judge with *P/Q* scores for each of the pairs:

<i>A1, A1</i>	<i>A1, A2</i>	<i>A1, A3</i>	<i>A1, A4</i>
<i>A2, A1</i>	<i>A2, A2</i>	<i>A2, A3</i>	<i>A2, A4</i>
<i>A3, A1</i>	<i>A3, A2</i>	<i>A3, A3</i>	<i>A3, A4</i>
<i>A4, A1</i>	<i>A4, A2</i>	<i>A4, A3</i>	<i>A4, A4</i>

Our goal was to measure the degree of variation in the writing style (here, measured by compression levels as determined using the GZip program). One measure of this variance is provided by the eigenvalue of the matrix.⁸⁸ The higher the eigenvalue, the higher the probability of sole authorship.

87. This is not to suggest that these commonly used words might not be useful with a different type of authorship test. One could, for example, run frequency tests on the relative uses of these common words, because different authors likely have different styles with respect to these words as well (some authors likely use more of them than others). One rationale for using measures of these “function” words (such as “and,” “the,” and “of”) is that the rate of their use is a function of the unconscious or habitual element of writing and, therefore, a better indicator of true authorship. For discussions, see J.F. Burrows, *Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style*, 2 LITERARY & LINGUISTIC COMPUTING 61 (1987), and David L. Hoover, *Statistical Stylistics and Authorship Attribution: An Empirical Investigation*, 16 LITERARY & LINGUISTIC COMPUTING 421 (2001).

88. The four-by-four matrix of compression scores for each judge’s four opinions is symmetric, since *P/Q* will be the same for the pairs [*A1, A2*] and [*A2, A1*]. By definition, a symmetric square matrix possesses *n* real eigenvalues (where *n* is the order of the ma-

To test the effectiveness of the GZip methodology for determining judicial authorship, we identified a set of judges who, a priori, are known to write their own opinions—our “test” judges. An informal survey of more than two dozen judges and law clerks, who requested that we not name them, produced a high degree of agreement on the names of three federal circuit court judges who regularly author their opinions: Richard Posner, Frank Easterbrook, and Michael Boudin.⁸⁹ More specifically, we were told that Posner drafts every word of his opinions; that Easterbrook may allow his clerks to draft one or two opinions a year but drafts the remainder himself; and that Boudin uses a combination of very heavy editing of clerk drafts and self-authorship. To reiterate, all three are reputed to fall on the high end of the scale of self-authorship. Assuming this informal information to be correct, the scores for these judges should be among the highest. If the GZip methodology is effective in identifying judges who author their own opinions, our three test judges should rank highly among our sample of ninety-eight judges. We do not perform the alternative test of identifying several judges reputed *not* to write their own opinions and assessing whether the GZip methodology succeeds in assigning a poor ranking to such judges. Unfortunately, for purposes of our tests, the judges and clerks we surveyed were reluctant to identify those least inclined to author their opinions.

We report in Table 1 below the ranking of the top fifteen judges out of our sample of ninety-eight judges based on the eigenvalue score from our GZip test. Higher eigenvalues correlate with a greater likelihood of self-authorship. In addition, we report the ranking of our three test judges.

trix). Eigenvalues have several applications. In this context, we use it as a single metric that measures the variability among the elements of the matrix. We borrowed the reasoning for using eigenvalues from factor analysis, where they are used to measure variability.

89. A fuller study would need to use additional controls, particularly if there is doubt about whether Posner, Easterbrook, and Boudin actually author the major portion of their own opinions. Two possible other controls are checking whether the authorship tests return higher scores for (a) dissents than for majority opinions, because informal information suggests that dissents are more personal to the judge and, we assume, more likely to be self-authored and (b) judges in the distant past than for judges of today, because judges in the distant past are reputed to have written a far larger fraction of their own opinions since they did not face the caseload pressures of current judges. We did not use either of these two controls because of inadequacies in our dataset. On the points about dissents being more personal and judges in the past writing more of their own opinions, see Blomquist, *supra* note 6, at 86-92 (discussing the characteristics of dissent styles); Richman & Reynolds, *supra* note 8, at 278-79 (discussing how the increases in caseloads from the time of Learned Hand to the present have resulted in increased levels of delegation of the opinion-writing task); Frankel, *supra* note 3 (bemoaning the passing of the Leaned Hand style of opinion writing, where there was little delegation of the writing task to the law clerks); *cf.* Wahlbeck et al., *supra* note 42, at 168 (pointing out that the conventional view that judges in the past single-handedly crafted all their own opinions whereas today’s judges lean heavily on their clerks is an oversimplification, because there is evidence suggesting that clerks have long played a significant role in the opinion-drafting process).

TABLE 1
RANKING BASED ON GZIP METHODOLOGY

RANK	JUDGE	CIRCUIT	OPINIONS EIGENVALUE
1	Randolph, A. Raymond	DC	169.01
2	Ginsburg, Douglas H.	DC	160.29
3	Manion, Daniel A.	7	150.58
4	Garza, Emilio M.	5	145.91
5	McKee, Theodore A.	3	140.23
6	Moore, Karen Nelson	6	139.44
7	Tjoflat, Gerald Bard	11	137.95
8	Batchelder, Alice M.	6	137.46
9	Calabresi, Guido	2	137.03
10	Flaum, Joel M.	7	135.83
11	Walker, John M., Jr.	2	135.00
12	Gilman, Ronald Lee	6	134.86
13	Dennis, James L.	5	134.82
14	DeMoss, Harold R., Jr.	5	134.57
15	Ebel, David M.	10	134.48
.	.	.	.
44	<i>Easterbrook, Frank H.</i>	7	130.19
50	<i>Posner, Richard A.</i>	7	129.89
97	<i>Boudin, Michael</i>	1	121.81

Somewhat surprisingly (and to our dismay after expending 500 hours of resources), none of our three test judges show up as top scorers on the GZip tests. Boudin, in fact, scores near the very bottom of the judges, and only Easterbrook is in the top half of judges.

We provide one variation on the GZip methodology. Instead of computing the eigenvalue of the four-by-four matrix of GZip scores, we classify the compression score for any two opinions for a particular judge as either high, middle, or low.⁹⁰ We then look at the difference between the number of high compression scores (high likelihood of sole authorship) and the number of low compression scores (low

90. We take the lower triangular numbers of the four-by-four matrix (the upper triangular numbers can be ignored since the matrix is symmetric, while the diagonal can be ignored since it represents a self-referential relation). The overall range of the P/Q compression numbers across all judges goes from 3.0 to 6.0. We classify each score in the lower triangular portion of the matrix into one of three qualitative groups as follows:

Low	3.0 – 4.2
Medium	4.2 – 4.8
High	4.8 – 6.0

likelihood of sole authorship) while ignoring the middle scores. We predict that judges who self-author their opinions will tend to receive a much greater number of high-compression-score opinions, after netting out all the low-compression-score opinions. Table 2 reports the rankings based on this GZip variation.

TABLE 2
RANKING BASED ON NUMBER OF HIGH - LOW GZIP SCORES

RANK	JUDGE	CIRCUIT	NUMBER OF HIGH COMPRESSION SCORES MINUS NUMBER OF LOW COMPRESSION SCORES
1	Randolph, A. Raymond	DC	3
2	Manion, Daniel A.	7	2
2	Ginsburg, Douglas H.	DC	2
2	Garza, Emilio M.	5	2
2	McKee, Theodore A.	3	2
2	Cole, R. Guy, Jr.	6	2
2	Gilman, Ronald Lee	6	2
2	Sentelle, David B.	DC	2
2	Tashima, A. Wallace	9	2
10	Ebel, David M.	10	1
10	Tjoflat, Gerald Bard	11	1
10	<i>Posner, Richard A.</i>	7	1
13	<i>Easterbrook, Frank H.</i>	7	0
13	O'Scannlain, Diarmuid F.	9	0
13	Rovner, Ilana Diamond	7	0
.	.	.	.
98	<i>Boudin, Michael</i>	1	-5

As with the initial GZip methodology, our test judges fail to score consistently well in the GZip score variation. Posner scores in the top fifteen, coming in tied at number ten with three other judges. Easterbrook also scores highly, tied at thirteenth. Forty-seven other judges, however, are also tied at thirteenth. Boudin is the lowest-ranked judge in terms of authorship. Given the small number of each judge's opinions that we examined, differences among judges in the GZip variation are not great enough to make fine-tuned distinctions among judges. Our first cut at both black-box tests resulted in failure: Neither GZip test produced results consistent with our a priori information that Posner, Easterbrook, and Boudin author their own opinions more than other circuit court judges.

What does the foregoing suggest? First, it may be that the small sample of data that we used for each judge, four randomly selected opinions, was inadequate. We may need to use a larger set of opinions for each judge for the black-box tests to have any traction. Moreover, our opinions are all selected within a narrow time frame (1998 to 2000). A high authorship score for a particular judge may indicate that the same clerk wrote some, if not all, of the four opinions for the judge. Choosing opinions spread out across a longer time frame will reduce the possibility that a high authorship score is due, in fact, to a particular clerk.

Second, the fact is that legal writing in a particular subject matter area contains genre-specific language. Tests such as the GZip algorithm look for similar word patterns. Legal opinion writing is likely to have considerable genre-specific commonality. The vast majority of opinions are likely to mention phrases such as “standard of review” or “summary judgment” or “motion to dismiss.” If there are enough of these common phrases in all the opinions, they may swamp the calculations. Put differently, legal opinion writing as a genre may have such a distinct style that genre-specific tests must be devised for it.

C. White-Box Tests

Our simple black-box tests without subject matter controls failed to distinguish judges based on the degree of authorship. While the black-box tests worked well in comparing generic texts with one another, we suspect that they failed in the judicial context due to the specialized nature of many types of judicial opinions. Opinions of the same genre, for example, may use various forms of jargon and other common phrases that are shared in the opinions of different judges. Even judges who do not self-author their own opinions will receive a high same-authorship score for opinions within the same genre.

Despite the failure of our simple black-box tests, we contend that additional tests geared to controlling for the subject matter of specific opinions may still work to distinguish judges based on the degree of opinion authorship. We call these tests our “white-box” tests. At least three categories of white-box tests are possible based on citation practices, language patterns, and a more nuanced version of the black-box GZip methodologies.

1. Citation Practices

Citations are a key element in the crafting of judicial opinions. Judges who write their own opinions may display specific patterns in the opinions they cite, tending to cite opinions with which they are more familiar. We perhaps acted too hastily in cleaning the opinions used in our black-box tests of the citations to other opinions. The pat-

terns of citations themselves provide crucial information on authorship. Other things equal, two opinions with the same citation patterns (for example, citing the same set of opinions) are more likely to be authored by the same author than two opinions without the same citation pattern.

One type of citation pattern involves self-citations. Landes, Lessig, and Solimine conjecture that judges who write their own opinions are likely to be more familiar with these opinions, which would lead to greater self-citation rates.⁹¹ Law clerks, by contrast, are likely to simply conduct Westlaw searches or draw from the parties' briefs. A higher average number of self-citations per opinion is therefore likely to correlate with a greater likelihood of authorship.

Importantly, self-citation patterns are at least somewhat invariant with the subject matter of an opinion. A judge who tends to cite her own work will cite her own criminal law cases when writing a subsequent criminal law opinion and her own securities law cases when writing a subsequent securities law opinion. Subject matter may, nonetheless, still be important if a judge writes opinions in a particular area more frequently because she will, as a result, cite her own opinions in that area more frequently.

As a quick test of the hypothesis that self-citation rates correlate with authorship, we examine the self-citation patterns using our 1998 to 2000 dataset. We focus in particular on whether the three test judges reputed to author their own opinions are among the top scorers on self-citations. Table 3, *infra*, reports the self-citation rates ranking for our test judges and the top fifteen scorers in our ranking. The self-citation rate is defined as the average number of self-citations (measured from 1998 to 2003) to each judge's opinions written in the 1998 to 2000 sample time period. Because we focus on each judge's opinions from the same three-year period, each judge starts with a similar pool of opinions that they may self-cite. In addition, the use of a pool of opinions from the same time period controls for changes in self-citation patterns over time.

91. See Landes et al., *supra* note 14, at 274 ("It is not implausible that judges who write their own opinions will cite themselves more frequently than judges who do not—if only because they have a greater familiarity with their own prior opinions.").

TABLE 3
RANKING BASED ON SELF-CITATION RATE

RANK	JUDGE	CIRCUIT	AVERAGE SELF-CITATIONS TO EACH OPINION	T-TEST OF DIFFERENCE WITH THE MEDIAN JUDGE* (EQUAL VARIANCES)
1	Selya, Bruce M.	1	2.44	7.00**
2	Wollman, Roger L.	8	1.56	3.93**
3	Lynch, Sandra L.	1	1.48	4.15**
4	<i>Posner, Richard A.</i>	7	1.46	2.81**
5	Clay, Eric L.	6	1.40	4.44**
6	Carnes, Ed	11	1.31	4.02**
7	Garland, Merrick B.	DC	1.31	4.52**
8	Moore, Karen Nelson	6	1.23	3.12**
9	Kelly, Paul J., Jr.	10	1.12	3.13**
10	<i>Easterbrook, Frank H.</i>	7	1.10	2.04**
11	Ebel, David M.	10	1.08	2.73**
12	Coffey, John L.	7	1.03	2.15**
13	Murphy, Michael R.	10	1.03	2.61**
14	Kanne, Michael S.	7	1.02	2.07**
15	Marcus, Stanley	11	1.02	3.21**
.
27	<i>Boudin, Michael</i>	1	0.61	0.89

* The median judge (Mary Beck Briscoe) is chosen as the forty-ninth ranked judge out of ninety-eight total judges.

** Indicates a significance level of 5%. The self-citation rate for Mary Beck Briscoe was 0.35 to each opinion.

In Table 3, we observe that both Posner and Easterbrook are in the top ten judges out of the sample of ninety-eight circuit court judges in terms of self-citation rates. Boudin, on the other hand, is ranked twenty-seventh, although he still is in the top half of all judges. For each judge, we perform a two-sided t-test assuming equal variances between that judge's self-citation rate and the median judge's—Mary Beck Briscoe—self-citation rate. All the top fifteen judges are significantly different at the five percent confidence level from the median judge.

Note that the t-statistic test we perform in comparison with the median judge does not tell us anything about whether Posner's authorship score is significantly different from the next-ranked judge (Clay). When we compare Posner against Clay, we do not find any statistically significant difference. Nevertheless, assuming that authorship is a positive, part of our goal in ranking judges is to incen-

tivize all judges to exert greater effort in authoring their judicial opinions. Even if no statistically significant difference exists between any two particular judges, the ranking will induce judges to exert effort. So long as greater effort increases the likelihood that a judge will rank higher than another judge (even if not with statistical significance), the judge will have an incentive to exert more effort.⁹²

Second, we examine invocation rates. An invocation is a citation whereby the judge is mentioned in the citing opinion by name (other than a perfunctory use of the name as, say, part of a parenthetical indication that the judge authored a particular dissenting or concurring opinion).⁹³ Higher invocation rates for a judge's opinions, we posit, correlate with a higher likelihood that the invoked judge authored her own opinions.

How are invocations related to authorship? We assume that judges have institutional knowledge as to which of their colleagues write their own opinions. Because an invocation represents a special indication of respect to the judge being cited (ordinarily judges are not referred to by name), it is unlikely that the special respect will be given unless the judge in question is one who writes her own opinions.⁹⁴ Put differently, a judge who is known to delegate the majority of her opinions to the clerks

92. Judges may write opinions of different length. The longer the opinion, the greater the likelihood that the opinion will receive a self-citation (due, for example, to the greater amount of analysis in a longer opinion). To control for this possibility, we also ranked the judges in our sample based on the average number of self-citations to each written opinion page. We provide the ranking below.

RANK	JUDGE	CIRCUIT	AVERAGE SELF-CITATIONS TO EACH OPINION PAGE
1	Wollman, Roger L.	8	0.3393
2	Posner, Richard A.	7	0.3283
3	Bruce M. Selya	1	0.3230
4	Easterbrook, Frank H.	7	0.2520
5	Kelly, Paul J., Jr.	10	0.1873
6	Lynch, Sandra L.	1	0.1673
7	Carnes, Ed	11	0.1612
8	Moore, Karen Nelson	6	0.1600
9	Clay, Eric L.	6	0.1443
10	Coffey, John L.	7	0.1378
11	Garland, Merrick B.	DC	0.1343
12	Tacha, Deanell Reece	10	0.1334
13	Kanne, Michael S.	7	0.1331
14	Ebel, David M.	10	0.1296
15	Ripple, Kenneth F.	7	0.1246
.	.	.	.
20	Boudin, Michael	1	0.1016

Note that Posner and Easterbrook do even better than in the self-citation rate per opinion measure. Boudin also does better, but still remains outside the top fifteen.

93. See Choi & Gulati, *supra* note 5, at 58-61 (defining invocations and analyzing judges based on the rates at which their names are invoked in other opinions).

94. See Landes et al., *supra* note 14, at 274 ("It is also not implausible that judges who write their own opinions will be more influential, since their opinions will be more consistent and, if good, then more consistently good than opinions written by law clerks.").

is unlikely to find her judicial colleagues invoking her. They may cite the judge, if her opinion is on point, but will be unlikely to invoke her and thereby give her a special measure of respect.⁹⁵

Invocation rates are at least somewhat robust to the subject matter of opinions. A judge who is held in high regard among other judges will tend to receive invocations for all types of opinions. Nonetheless, subject matter will not be completely irrelevant to the extent that judges do tend to invoke other judges for specific types of opinions (for example, Easterbrook on corporate and commercial-related law). Table 4 reports the results from the invocation rate ranking. A judge's invocation rate is defined as the average number of invocations (measured from 1998 to 2003) of each judge's opinions written in the 1998 to 2000 time period.

TABLE 4
RANKING BASED ON INVOCATION RATE

RANK	JUDGE	CIRCUIT	AVERAGE INVOCATIONS TO EACH OPINION	T-TEST OF DIFFERENCE WITH THE MEDIAN JUDGE* (EQUAL VARIANCES)
1	<i>Posner, Richard A.</i>	7	0.664	6.34**
2	<i>Easterbrook, Frank H.</i>	7	0.442	3.78**
3	Calabresi, Guido	2	0.228	3.08**
4	Wilkinson, J. Harvie, III	4	0.185	2.59**
5	Edmondson, J.L.	11	0.138	1.17
6	Higginbotham, Patrick E.	5	0.124	1.34
7	Luttig, J. Michael	4	0.124	1.14
8	Jones, Edith H.	5	0.109	0.98
9	<i>Boudin, Michael</i>	1	0.096	1.30
10	Walker, John M., Jr.	2	0.095	1.11
11	Clay, Eric L.	6	0.086	0.83
12	Cabranes, José A.	2	0.086	0.92
13	Kleinfeld, Andrew J.	9	0.085	0.86
14	Tjoflat, Gerald Bard	11	0.083	0.55
15	King, Carolyn Dineen	5	0.082	0.77

* The median judge (Terence T. Evans) is chosen as the forty-ninth ranked judge out of ninety-eight total judges.

** Indicates a significance level of 5%. The invocation rate for Terence T. Evans was 0.039 to each opinion.

95. On this special measure of respect that accrues to those judges doing their own work, Justice Brandeis famously said that "[t]he reason the public thinks so much of the Justices of the Supreme Court is that they are almost the only people in Washington who do their own work." DAVID M. O'BRIEN, *STORM CENTER: THE SUPREME COURT IN AMERICAN POLITICS* 116 (5th ed. 2000) (internal quotation marks omitted); see also John G. Kester, *The Law Clerk Explosion*, *LITIGATION*, Spring 1983, at 20, 62.

All three of the test judges score in the top ten in terms of invocation rates. Indeed, Posner and Easterbrook are numbers one and two, respectively, in the ranking. Both Posner and Easterbrook's scores, in addition, are significantly different from the median judge's (Terence T. Evans) score. Posner's and Easterbrook's high t-statistics indicate that they are relatively more likely to self-author compared to the median judge in a statistically significant manner. Some evidence exists that a ranking based on invocation rates may distinguish among judges based on their self-authorship of judicial opinions. Nonetheless, reputation may reflect a judge's long-term self-authorship pattern rather than the degree of authorship in any particular set of contemporary opinions.

Judges who write their own opinions, in addition to being more likely to cite themselves, might also be more likely to cite a smaller number of other judges relative to the average citation pattern. Judges are likely to have a more fine-tuned sense of which of the other judges are worthy of citation than their law clerks. Once a judge identifies her set of preferred other judges, the judge is likely to stick with them in her citation pattern, leading to a relatively low level of variance in citations. On the other hand, where clerks author the opinions, they will not necessarily cite to the same set of judges but may cite to widely differing judges. The variation in the number of different judges cited, therefore, will likely be higher with a judge who delegates extensively compared with a judge who tends to author her own opinions. Because our 1998 to 2000 dataset does not contain information on the specific identities of judges that a particular judge cites, we are unable to test whether variance in citation performs well in distinguishing our a priori set of judges who are reputed to self-author their opinions. We leave this test to another paper.⁹⁶

2. *Subject Matter-Neutral Language Patterns*

The writing of judges who author their own opinions will likely display patterns. While specific word patterns (for example, use of the phrase "habeas corpus") may be subject matter-specific, some patterns may not depend on the subject matter of a judicial opinion. The use of these genre-neutral patterns provides an alternative "white-box" method of testing for authorship.

For example, some judges will tend to write short opinions and others will write longer ones. Some will use extended quotes and others will not. When judges delegate to clerks, however, there will likely be greater variation in the types of opinions because clerks will

96. We are in the process of collecting these data now.

have their own styles. For some fixed number of opinions, a judge who writes her own opinions is likely to have a smaller opinion-size variation than a judge who does not. In our opinion, neither the length of the opinion (or average paragraph or sentence) nor the use of block quotes are necessarily tied to any specific subject matter or area of the law.

For each of the ninety-eight judges in our sample, we calculated the standard deviation of the length of each majority opinion based on published pages in the West Federal Reporter and excluding the summary and West keynotes. The standard deviation of majority-opinion length is for opinions from the 1998 to 2000 time period for each judge. Table 5 reports the judges' rankings based on the standard deviation of majority-opinion length.

TABLE 5
RANKING BASED ON STANDARD DEVIATION
OF MAJORITY-OPINION LENGTH

RANK	JUDGE	CIRCUIT	STANDARD DEVIATION OF MAJORITY-OPINION LENGTH	F-TEST P-VALUE*
1	Loken, James B.	8	1.81	0.0000
2	<i>Posner, Richard A.</i>	7	1.84	0.0000
3	Martin, Boyce F., Jr.	6	1.84	0.0087
4	Wollman, Roger L.	8	1.92	0.0002
5	<i>Easterbrook, Frank H.</i>	7	1.99	0.0002
6	Black, Susan H.	11	2.21	0.0277
7	Arnold, Morris S.	8	2.25	0.0044
8	<i>Boudin, Michael</i>	1	2.33	0.0091
9	Ginsburg, Douglas H.	DC	2.37	0.0267
10	Hawkins, Michael Daly	9	2.51	0.0636
11	Schroeder, Mary M.	9	2.53	0.0586
12	Tashima, A. Wallace	9	2.64	0.0707
13	Walker, John M., Jr.	2	2.64	0.0638
14	Nygaard, Richard L.	3	2.67	0.1243
15	Sentelle, David B.	DC	2.73	0.1146

* The F-test provides a test of the null hypothesis that the standard deviation of opinion length for a particular judge is the same as the standard deviation for the median judge (Deanell Reece Tacha). The standard deviation of opinion length for Deanell Reece Tacha was 3.65.

As with the self-citation rate and invocation rate tests, Posner and Easterbrook score in the top-ten judges. In addition, Boudin—who was in the top ten for the invocation rate test but not for the self-citation rate test—is also ranked among the top-ten judges based on a lower standard deviation of opinion length. The differences between the standard deviation scores for Posner, Easterbrook, and Boudin compared with the score for the median judge (Deanell Reece Tacha) are significant at the one percent level. The standard deviation of the majority-opinion length for the judges successfully distinguished our test judges in terms of both high rank among our sample judges and comparison with the median judge.

We also hypothesize that judges who write their own opinions will produce shorter opinions with fewer quotes. They have a limited amount of time, and a large amount of work, and they are more confident about what they are saying. In Table 6, *infra*, we provide a ranking of our ninety-eight judges according to average length of majority opinions during the 1998-2000 time period. The length of each opinion is based on published pages in the West Federal Reporter excluding the summary and West keynotes. The average printed pages per majority opinion is equal to the total length of all opinions for a particular judge divided by the number of opinions written from 1998 to 2000.

TABLE 6
RANKING BASED ON AVERAGE PAGES PER MAJORITY OPINION

RANK	JUDGE	CIRCUIT	AVERAGE PRINTED PAGES PER OPINION	T-TEST OF DIFFERENCE WITH THE MEDIAN JUDGE* (UNEQUAL VARIANCES)
1	Arnold, Morris S.	8	3.94	-10.14**
2	Easterbrook, Frank H.	7	4.38	-9.47**
3	Posner, Richard A.	7	4.44	-9.51**
4	Loken, James B.	8	4.48	-9.06**
5	Ginsburg, Douglas H.	DC	4.65	-7.22**
6	Wollman, Roger L.	8	4.66	-8.48**
7	Martin, Boyce F., Jr.	6	4.73	-6.84**
8	Schroeder, Mary M.	9	4.93	-6.20**
9	Black, Susan H.	11	5.04	-5.86**
10	Bowman, Pasco M.	8	5.17	-5.91**
11	Murphy, Diana E.	8	5.32	-5.27**
12	Evans, Terence T.	7	5.36	-5.76**
13	Higginbotham, Patrick E.	5	5.65	-4.49**
14	Kozinski, Alex	9	5.70	-3.17**
15	Edmondson, J.L.	11	6.00	-3.08**
.
18	Boudin, Michael	1	6.09	-4.17**

* The median judge (Paul V. Niemeyer) is chosen as the forty-ninth ranked judge out of ninety-eight total judges. Niemeyer had an average majority opinion length of 7.67 pages.

** Indicates a significance level of 5%.

Posner and Easterbrook again appear among the top-ten judges in terms of likelihood of self-authorship. Boudin, however, falls out of the top ten and is ranked number eighteen, but still within the top-twenty judges. The average majority-opinion page length for all three test judges was significantly lower than the average page length for the median judge (Paul V. Niemeyer) as deduced using a two-sided t-test assuming unequal variances.

Other frequency-based tests may prove feasible. A judge writing an opinion will be less likely to footnote her opinion heavily than the law clerk.⁹⁷ Clerks, because of their lower knowledge base, higher level of insecurity, and law review training, are more likely to feel a need to footnote the document. We predict, therefore, that a lower

97. See *supra* note 70 and accompanying text (citing Posner, who hypothesizes that clerks will use more footnotes).

number of footnotes per opinion (or per page) correlates with a higher likelihood of self-authorship.⁹⁸

3. *Revisiting the Black-Box Tests*

We and others may learn from the failure of our initial run at the “black-box” GZip-based tests. In order to take advantage of genre-specific information and to control for genre-specific commonalities across opinions written by different authors, at least three changes to our methodology are possible.

First, to control for the possibility that a particular clerk may dominate a judge’s opinion writing for a specific year, the sample of opinions for each judge should span a relatively long time period—certainly greater than the three years covered by our 1998 to 2000 time period. Increasing the number of opinions for each judge—from four to ten, for example—may also provide traction to the GZip results.

Second, as discussed above, we were too zealous in our desire to remove unnecessary information from our opinions. The informativeness of the opinions (on the issue of authorship) would likely have been greater if we had kept the citation information. Keeping citation information also greatly reduces the work required in preparing opinions for the GZip-based tests.

Third, a study could control for the noise generated from comparing opinions across different areas of the law. Suppose Judge *A* writes her own opinions while Judge *B* does not. If we compared four

98. We used two other less sophisticated methods based on the publicly available Litstat program. For the Litstat tool on the Internet, see Matthew Bielich, *LitStat*, at <http://www.pinionsolutions.com/litstat/> (last visited Jan. 22, 2005). Authors with distinctive styles and patterns are assumed to use similar types of sentence and word structures in all of their writing. Using the Litstat program, we calculated the alphabetical frequency of a text, that is, the frequency with which the words in that text begin with a certain letter of the alphabet. Frequency of commonly used words is an indicator of the specificity of style. Thus, each author is likely to have a set of words that he or she may use more frequently than other authors. A frequency distribution of words could thus be an indicator of the stylistic uniqueness. For each judge we used two samples of text of 8000 characters each. As with our other black-box tests, we did not control for the subject matter of the judicial opinions. As with the GZip test, we cleaned the text. Here, however, we did not take out common words such as “and” and “the,” because the information as to those commonly used words is actually important to these supplementary tests. The alphabetical distribution of the words in the opinions is the output of the tool. This output is copied to a spreadsheet, and the average of the difference of these values for the two opinions is calculated—the lower the average, the higher the probability of sole authorship. None of our three control judges were in the top fifteen judges in terms of probability of sole authorship. We also used the Litstat program to calculate the average sentence length in a fixed amount of text (8000 characters, after the text has been cleaned for citations, quotes, West headnotes and other extraneous information). We predicted that the variance in these frequencies would be smaller among documents authored by the same author than among those authored by different authors. Only Boudin scored in the top fifteen judges (at number two) in terms of probability of sole authorship.

opinions by Judge *A* on diverse areas of the law (including, for example, criminal law, bankruptcy law, administrative law, and securities regulation) against four opinions by Judge *B* on securities regulation, Judge *B* will likely receive a higher same-authorship score. Such a pattern, possible where opinions are randomly selected without regard to subject matter, may introduce noise into the analysis.

A simple control for such noise would be to select randomly opinions from the *same* subject matter area of the law—criminal law, for example. It is possible that the common phrases in criminal law opinions (for example, “habeas corpus”) are so frequent that all judges will receive the identical same-authorship score even with this control. Nonetheless, to the extent that, after taking into account the common baseline vocabulary, judges do have idiosyncratic writing styles, the GZip methodologies should pick up on these differences in determining authorship.

We leave implementing these white-box controls to the black-box tests for the genre-specific nature of opinions for our next paper.

IV. CONCLUSION: OTHER POSSIBLE APPLICATIONS OF AUTHORSHIP TECHNOLOGY

This Essay only touches the surface of the world of technology that might be applied to determine judicial authorship of opinions. The results of the white-box tests are but a preliminary step toward determining authorship. Even assuming that these white-box tests have a significant amount of explanatory power to determine authorship, each of the white-box tests measures something that, at best, correlates with authorship. We suspect that meaningful progress in this area will require a serious collaborative research effort between legal academics and scholars in computational linguistics. Such technology, once developed, has the potential to assist in areas other than research on the judiciary. We list a couple below.

A. *Securities Fraud Complaints*

Congress enacted the Private Securities Litigation Reform Act of 1995 to combat frivolous litigation.⁹⁹ Proponents of the Act argued that plaintiffs’ attorneys literally cut-and-pasted complaints together to file suit against any company that experienced a large drop in stock price, regardless of whether any real evidence of fraud ex-

99. For a discussion of the Private Securities Litigation Reform Act of 1995 and the impact of the Act on both frivolous and meritorious claims, see STEPHEN J. CHOI, DO THE MERITS MATTER LESS AFTER THE PRIVATE SECURITIES LITIGATION REFORM ACT? (N.Y.U. Sch. of Law Law & Econ. Research Paper Series, Research Paper No. 03-04, 2004), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=558285.

isted.¹⁰⁰ Congress intended the Reform Act to force plaintiffs' attorneys to take more care in their filings and provide particularized factual assertions as to why fraud in fact exists.

We may be able to use the authorship methodology to determine how "close" two securities fraud complaints are to one another. If a complaint is really a cut-and-paste job, the complaint is likely to receive a high same-authorship score when compared to other securities complaints from the same law firm. If the Reform Act resulted in more particularized investigation before the filing of a complaint, we predict that a set of pre-Reform complaints from the same law firm would produce a higher same-authorship score as compared to a set of post-Reform complaints from the same law firm.

B. Boilerplate Contract Evolution

The contracts used in a variety of settings such as the corporate bond area are commonly described as boilerplate. In other words, essentially the same standard language gets repeated in every contract in the market. Even if individual drafters do not know what the contract language means, it gets repeated because everyone else in the market is using it. Problems can arise when, in the course of a dispute, the two sides assert different meanings for some piece of ambiguous language whose original understanding has long since been lost.

Authorship testing programs may be used to trace the original source of the later contracts. Finding the source is potentially important for at least two reasons. First, from a practical dispute-resolution point of view, knowing the original understanding might help resolve the current dispute (to the extent the judge decides that the original understanding should govern). Second, from a research point of view, the ability to track the evolution of contract language can help us understand how contract language evolves, how it becomes boilerplate, and what circumstances produce changes in boilerplate.¹⁰¹

100. See James Bohn & Stephen Choi, *Fraud in the New-Issues Market: Empirical Evidence on Securities Class Actions*, 144 U. PA. L. REV. 903, 904-05 (1996) (discussing one plaintiff's attorney's reference to Philip Morris' success in the "toy industry" as evidence of cut-and-paste complaints) (emphasis and internal quotation marks omitted).

101. For an empirical examination of how standardized sovereign bond covenants have evolved over time (not using authorship methodology), see Stephen J. Choi & G. Mitu Gulati, *Innovation in Boilerplate Contracts: An Empirical Examination of Sovereign Bonds*, 53 EMORY L.J. 929 (2004).