Florida Law Review

Volume 64 | Issue 5 Article 2

10-17-2012

Triangulating Judicial Responsiveness: Automated Content Analysis, Judicial Opinions, and the Methodology of Legal Scholarship

Chad M. Oldfather

Joseph P. Bockhorst

Brian P. Dimmer

TRIANGULATING JUDICIAL RESPONSIVENESS: AUTOMATED CONTENT ANALYSIS, JUDICIAL OPINIONS, AND THE METHODOLOGY OF LEGAL SCHOLARSHIP

Chad M. Oldfather[*]
Joseph P. Bockhorst[**]
Brian P. Dimmer[***]

Abstract

The increasing availability of digital versions of court documents, coupled with increases in the power and sophistication of computational methods of textual analysis, promises to enable both the creation of new avenues of scholarly inquiry and the refinement of old ones. This Article advances that project in three respects. First, it examines the potential for automated content analysis to mitigate one of the methodological problems that afflicts both content analysis and traditional legal scholarship—their acceptance on faith of the proposition that judicial opinions accurately report information about the cases they resolve and courts' decisional processes. Because automated methods can quickly process large amounts of text, they allow for assessment of the correspondence between opinions and other documents in the case, thereby providing a window into how closely opinions track the information provided by the litigants. Second, it explores one such novel measure—the "responsiveness" of opinions to briefs—in terms of its connection to both adjudicative theory and existing scholarship on the behavior of courts and judges. Finally, it reports our efforts to test the viability of automated methods for assessing responsiveness on a sample of briefs and opinions from the United States Court of Appeals for the First Circuit. Though we are focused primarily on validating our methodology, rather than on the results it generates, our initial investigation confirms that even basic approaches to automated content analysis provide useful information about responsiveness, and generates intriguing results that suggest avenues for further study.

∗ Professor, Marquette University Law School. Thanks to Fred Bloom, Mary Clark, Amanda Frost, Michael Gerhardt, Mitu Gulati, Renee Lettow Lerner, Andrew Martin, Neomi Rao, Lori Ringhand, Ryan Scoville, Jay Tidmarsh, Robert Vaughn, and Steve Vladeck for their feedback on earlier drafts, as well as to the participants in a workshop at Marquette University Law School and the other panelists and audience members at the panel on "New Empirical and Theoretical Work on Judging and the Judicial Process" at the 2010 Southeastern Association of Law Schools (SEALS) conference.

** Assistant Professor, Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee.

*** Member, Wisconsin Bar.

INTRODUCTION

The American legal process has always been document-intensive.[1] Litigation occurs primarily through the submission of written briefs and often reaches its final resolution via a written judicial opinion. Legal scholarship has long reflected the centrality of the written word, albeit

1. *See* Suzanne Ehrenberg, *Embracing the Writing-Centered Legal Process*, 89 IOWA L. REV. 1159, 1178–85 (2004) (tracing the development of the American legal system's writing-centered nature and attributing it to the relatively vast geography of the United States coupled with the lack of trained barristers in the early days of our legal system).

in a limited way. In its classic form, it focuses overwhelmingly, and often exclusively, on judicial opinions.[2] This is understandable. Until recently, judicial opinions have constituted the only readily available source of documentary raw material for scholars.

Yet traditional legal scholarship's reliance on judicial opinions is its potential Achilles' heel. Such work takes on faith that opinions accurately reflect not only the court's reasoning, but also the facts and other features of the disputes that the opinions resolve.[3] If this is incorrect, and if opinions do not reliably provide an accurate report, then scholarship that relies entirely upon them may fail to perceive what is truly taking place, and thereby serve as an unreliable guide to its subject.[4]

Over the past several decades, however, a greater range of documents has become available, providing access to the litigants' perspective on the cases that reach the courts. One can now obtain electronic versions of opinions and, to an increasing degree, the parties' briefs through commercial services such as Westlaw and Lexis, as well as through courts' websites. At the same time, the power and sophistication of computational techniques of textual analysis have increased as well. These techniques have most famously been used to explore disputed questions of authorship, ranging from the *Federalist Papers* and some of Shakespeare's works to e-mails connected with the founding of Facebook.[5] It is hardly surprising that researchers have

---

2. *See* Mark A. Hall & Ronald F. Wright, *Systematic Content Analysis of Judicial Opinions*, 96 CALIF. L. REV. 63, 66 (2008) ("The traditional legal scholarly enterprise relies, like literary interpretation, on the interpreter's authoritative expertise to select important cases and to draw out noteworthy themes and potential social effects of decisions.").

3. *See id.* at 95–96. We discuss this point at greater length below. *See infra* Part I.

4. *See, e.g.*, Ann Juliano & Stewart J. Schwab, *The Sweep of Sexual Harassment Cases*, 86 CORNELL L. REV. 548, 559 (2001) ("The judicial opinion is the judge's story justifying the judgment. The cynical legal realist might say that the facts the judge chooses to relate are inherently selective and a biased subset of the actual facts of the case."); Robert P. Burns, *The Lawfulness of the American Trial*, 38 AM. CRIM. L. REV. 205, 219 (2001):

> The rhetoric of appellate opinions is designed, in part, to reflect the conception of the Rule of Law that is expressed in the Received View. Only hypothetical facts, or facts that are "found" by a court, lose the morally significant uncertainty and the normative multivalence surrounding virtually all "facts" in the trial court, and, I might add, in the world. The temptation to recount such "facts," by choices of characterization and inclusion with the legal norms and the preferred outcome in mind is almost irresistible. The expected unity of the opinion demands it. And so it is no surprise that lawyers, even appellate lawyers, often believe that the account of the facts provided by appellate courts is deeply unfair.

5. *See* Ben Zimmer, *Decoding Your E-Mail Personality*, N.Y. TIMES, July 23, 2011, http://www.nytimes.com/2011/07/24/opinion/sunday/24gray.html.

started to apply them to legal documents.[6]

Our core project in this Article is to introduce a methodological approach that, we contend, promises to shed some light on whether scholars' faith in the accuracy of judicial opinions is misplaced, as well as to illuminate a range of other questions relating to judicial performance and institutional design. We develop a specific measure that employs computational methods to assess—or, if you will, "triangulate"—the relationship among briefs and opinions.[7]

We call the characteristic under study "judicial responsiveness." In brief, the concept of responsiveness originates from the idea that the judicial role is, and for the most part ought to be, fundamentally reactive.[8] Reduced to its essence, the notion stems from the recognition that the judicial system exists primarily to provide a peaceful means of resolving disputes. From this, the argument runs, it follows that courts should focus primarily on addressing the parties' disputes, and should doing so on the terms by which the parties themselves conceive of them. If, for example, the parties regard their dispute as turning on the proper application of the case of *Smith v. Jones*, one would thus expect the court hearing their case to resolve it primarily with reference to *Smith v. Jones*. This is not, of course, to suggest that the court must always restrict itself to *Smith v. Jones*. As Amanda Frost has pointed out, the judicial system serves ends other than dispute resolution, such that it will often be appropriate for a court to draw on a broader range of material than what the parties have placed before it.[9] It might be that

---

6. *See infra* Section I.B.

7. We are not the first to apply computational methods to judicial opinions. *See generally* Stephen J. Choi & G. Mitu Gulati, *Which Judges Write Their Opinions (And Should We Care)?*, 32 FLA. ST. U. L. REV. 1077 (2005) [hereinafter Choi & Gulati, *Which Judges Write Their Opinions*]; Michael Evans et al., *Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research*, 4 J. EMPIRICAL LEGAL STUD. 1007 (2007).

8. For the classic articulation of this view, see generally Lon L. Fuller, *The Forms and Limits of Adjudication*, 92 HARV. L. REV. 353 (1978). Fuller's model "calls for the judiciary to assume a passive role pursuant to which judges restrict themselves as much as possible to reacting to the parties' arguments." Chad M. Oldfather, *Defining Judicial Inactivism: Models of Adjudication and the Duty to Decide*, 94 GEO. L.J. 121, 140 (2005) [hereinafter Oldfather, *Defining Judicial Inactivism*]; *see also* STEPHAN LANDSMAN, THE ADVERSARY SYSTEM: A DESCRIPTION AND DEFENSE 2 (1984):

> The adversary system relies on a neutral and passive decision maker to adjudicate disputes after they have been aired by the adversaries in a contested proceeding. He is expected to refrain from making any judgments until the conclusion of the contest and is prohibited from becoming actively involved in the gathering of evidence or the settlement of the case.

9. *See generally* Amanda Frost, *The Limits of Advocacy*, 59 DUKE L.J. 447 (2009) (defending the courts' practice of addressing claims and arguments that the parties have not raised).

Smith v. Jones must be read in light of other cases that the parties have overlooked, or perhaps even that some issue prior to the application of *Smith v. Jones*, such as jurisdiction, will end up driving the court's decision.[10] But there remains a basic obligation—though undoubtedly contestable in its particulars[11]—to grapple with what the parties have put before the court, such that even a court that grounds its decision elsewhere should address the question of why *Smith v. Jones* does not govern.[12] Our aim here is not to resolve these normative disputes, though we hope and expect that our methodology will generate results that will help ground them.

Measures of judicial responsiveness are potentially valuable in at least four broad respects regardless of one's preference for judicial passivity. First, and at the most basic level, they can inform our understanding of how the judiciary works by allowing for assessment of differences among courts and judges at both the same, and different, levels of the judicial hierarchy and over time. Because, for example, an appellate court's institutional role is different from a trial court's, we would expect to see a different relationship among briefs and opinions at the two levels.[13] Courts facing different docket pressures may vary as well. In addition, investigations of responsiveness might inform debates

10. *See id.* at 462–63 (examining how courts use jurisdictional issues to drive their decisions); *id.* at 463–67 (outlining various methods courts use to address arguments not raised by the litigants).

11. On one view, judicial decision-making that fails to be appropriately responsive constitutes "judicial inactivism." Chad M. Oldfather, *Remedying Judicial Inactivism: Opinions as Informational Regulation*, 58 FLA. L. REV. 743, 745 (2006) [hereinafter Oldfather, *Remedying Judicial Inactivism*]. There is plenty of anecdotal support for the suggestion that courts at least occasionally disregard their obligation to address the parties' contentions. *See, e.g.*, *id.* at 762, 774 & nn.151–52. And the increasing institutional pressures faced by most courts, primarily as a result of rising caseloads, have resulted in a situation in which there is arguably a greater likelihood of such "judicial inactivism," whether through inadvertence or a more conscious cutting of corners. *Id.* at 745. Indeed, some commentators have suggested that the sorts of behavior associated with inactivism has become epidemic. *See, e.g.*, William M. Richman & William L. Reynolds, *Elitism, Expediency, and the New Certiorari: Requiem for the Learned Hand Tradition*, 81 CORNELL L. REV. 273, 274–97 (1996) (examining shortcuts that courts are increasingly taking in the decision-making process and their impact on the quality of justice obtained by litigants).

12. As Judge Richard A. Posner has put it, "For the judge, the duty to decide the case (and with reasonable dispatch) is primary. He does not choose his cases, or the sequence in which they are presented to him, or decree a leisurely schedule on which to decide them." Richard A. Posner, *Tribute to Ronald Dworkin and a Note on Pragmatic Adjudication*, 63 N.Y.U. ANN. SURV. AM. L. 9, 12 (2007). For an effort to develop the contours of this duty see also Oldfather, *Defining Judicial Inactivism*, *supra* note 8, at 160–81.

13. Trial courts are, in general, more focused on the resolution of disputes, while appellate courts place comparatively greater emphasis on the refinement and development of legal standards. Because appellate courts must cast their gaze more broadly, we might expect to see less responsiveness in their opinions.

over the extent to which ideology and other nonlegal factors drive judicial decision making. All else being equal, greater responsiveness is consistent with there being less space for the operation of ideology.[14] Second, assessments of responsiveness can inform more normatively oriented scholarship, such as work attempting to assess judicial quality or critiquing the device of unpublished opinions.[15] Third, this line of research might yield payoffs to advocates. To the extent that it becomes possible to know specifics about what triggers greater responsiveness—such as, for example, whether the filing of a reply brief has an effect—lawyers will be able to adjust their efforts accordingly.[16] Fourth, as we have already alluded, studies of responsiveness might mitigate one of legal scholarship's methodological problems, namely that of taking on faith that judicial opinions accurately reflect the cases they describe. An appropriately crafted inquiry into the extent to which opinions appear to be "products" of the parties' briefs can provide evidence on the question of whether this faith is warranted.[17]

Although the relationship among the parties and the court stands at the heart of the judicial process, it has historically been difficult to assess systematically. In part, this is a product of the conceptual difficulties involved in determining precisely what the responsiveness obligation entails in any given case. As suggested above, sometimes a properly oriented court should focus on *Smith v. Jones*, while other times it will be appropriate, and even necessary, for the court to look beyond a particular case. There are practical difficulties as well. The measurement of a court's responsiveness in a given case requires nearly as much effort as was required to generate the court's decision in the first instance. The evaluator must first come to an understanding of the particulars of the parties' arguments. She must then measure whether the court has engaged with those arguments and whether its decision is, in a meaningful sense, a product of those arguments (and thus a decision that resolves the parties' dispute rather than some simulacrum). That process in turn raises at least two barriers to large-*n* research: the labor-intensive nature of the evaluation makes it impractical, and the subjectivity of the process introduces significant concerns about inter-

14. To elaborate, the hypothesis here is that a court that issues a highly responsive opinion will have left less space for the operation of ideology than a court that does not tether its analysis to the arguments and authorities in the parties' briefs. This is not to deny that ideological or other non-legal factors might drive such a decision. It is instead simply to assert that the limits imposed by responsiveness are real, and to posit that in the aggregate a court that limits the range of materials it offers in justification of its decisions will take fewer inputs (and thereby fewer improper inputs) into account in reaching its decisions than a court that does not so limit itself.

15. *See infra* Section I.B.

16. *See infra* Section I.B.

17. *See infra* Section I.B.

coder reliability.[18]

Our final aim, then, is to explore whether computational methods can overcome these barriers. Enlisting computers rather than humans to "read" and code opinions and other documents will enable researchers to analyze large amounts of information in short periods of time, and to do so with no need to worry about consistency from one reader to the next. Using a set of briefs and opinions from the First Circuit, we have investigated two automated measures of judicial responsiveness both of which avoid the practical difficulties associated with manually assessing responsiveness, both of which employ a notion of the similarity between briefs and opinions. The first involves assessing document similarity through analysis of textual content of briefs and opinions. The second utilizes a similar methodology applied to citations to authority; that is, we assessed the extent to which opinions cite to the same legal authorities as relied upon by the parties in their briefs. In order to test the validity of these measures, we also undertook the sort of full-scale assessment of a set of cases outlined in the preceding paragraph, reviewing the briefs and opinions in depth and coding them for responsiveness.

Our primary focus was on establishing the viability of automated measures of responsiveness. A comparison of the results of our automated and manual assessments suggests both the validity of an automated approach and avenues for potential refinement. Other results were also intriguing. For example, reply briefs in our sample scored substantially lower in terms of responsiveness than principal briefs. And the court's citation practices show surprisingly little overlap between authorities cited in briefs and those cited in opinions. It is unclear what to make of this—one could equally tell a story of a court admirably exercising independent judgment or of a court failing to meet its obligations to the litigants (or perhaps even to the law). The truth is probably somewhere in between. Either way, the results provide further support for the conclusion that the investigation of responsiveness promises to generate useful insights.

The remainder of this Article proceeds as follows. Part I provides an overview of prior efforts to apply automated content analysis to legal documents and attempts to situate those efforts within the larger project of content analysis. As our brief survey reveals, past research has focused on questions relating to the authorship of judicial opinions,[19] to the refinement of quantitative empirical research,[20] and to the exploration of the relationship between party briefs and judicial

18. *See* Evans et al., *supra* note 7, at 1008–09.

19. *See infra* Subsection I.B.1.

20. *See infra* Subsection I.B.2.

opinions.[21] This work remains in its early stages but promises to facilitate new types of inquiry into old questions, as well as enabling new types of research into the behavior of courts and litigants. We contend that a broader form of content analysis—one that is made considerably more practicable through the use of automated methods—offers the potential to mitigate one of the methodological problems of both content analysis and traditional legal scholarship. Specifically, inquiry into the relationship among opinions and the briefs and other documentary components of a case can provide a means for assessing whether judicial opinions are consistently faithful in their reporting of the facts and arguments in the cases they resolve.

Part II makes the case for measuring responsiveness as a component of broader scholarly efforts to understand courts and judges. Responsiveness may be valuable in its own right, as a characteristic of legitimate adjudication. It may also assist in addressing various questions of institutional design and process, such as those relating to the effects of caseload pressures and the role of ideology in judging, as well as in efforts to assess judicial quality. Part III relates the methodology and results of our initial investigation of responsiveness, using a set of cases from the First Circuit, and employing methods that analyze the correspondence among briefs and opinions using both textual similarity and citation overlap. That work provides initial confirmation of the reliability and validity of automated methods of measuring responsiveness. Finally, we conclude and offer our thoughts on future directions that our research might take.

I. THE USES AND POTENTIAL USES OF AUTOMATED CONTENT ANALYSIS

Our broad topic is automated content analysis, which is of course a subset of the larger domain of content analysis. In a recent article in the *California Law Review*, Professors Mark Hall and Ronald Wright consider the prospect of content analysis as "a uniquely legal empirical methodology."[22] At its heart, the method is straightforward and charts a middle ground between traditional and empirical legal scholarship. "[A] scholar collects a set of documents, such as judicial opinions on a particular subject, and systematically reads them, recording consistent features of each and drawing inferences about their use and meaning."[23] The result is to combine traditional scholarship's textual engagement[24]

21. *See infra* Subsection I.B.3.

22. Hall & Wright, *supra* note 2, at 64. The backdrop for their analysis is their assessment that legal scholarship does not have its own unique empirical methodology, tending instead to borrow social scientific techniques, with mixed results. *Id.* at 63–64.

23. *Id.* at 64.

24. "This method comes naturally to legal scholars because it resembles the classic scholarly exercise of reading a collection of cases, finding common threads that link the

with the methodological rigor of quantitative empirical analysis. Properly conducted, content analysis involves systematic selection and coding of cases, often followed by statistical analysis of the coding.[25] The method's rigor and social-scientific overtones arise primarily through the prospect of replicability.[26] If other researchers would reach the same conclusions were they to read and analyze the same cases, then the authority for the project's results lies in the method rather than in the researcher.[27]

Much of modern legal scholarship, however, falls into two other categories.[28] The first, which we will somewhat loosely refer to as "traditional legal scholarship," has as its hallmark close attention to judicial opinions. The scholar starts with a basic legal question, such as "How should the Fourth Amendment apply to e-mails?" The core of the scholar's effort to answer the question in this form of scholarship consists of the close scrutiny and detailed analysis of, in this case, past Fourth Amendment decisions, particularly those generated by the United States Supreme Court. Much of the reasoning is analogical, with the author working to show that there are pertinent ways in which e-mail is, or is not, analogous to the situations addressed in previous cases. She may draw on other disciplines, such as history, political theory, or psychology, but the work remains rooted in the content of judicial opinions. This sort of work proceeds based on a number of typically unstated assumptions, including acceptance of the propositions that legal rules and doctrine operate as meaningful guides to and restraints on judicial decision making and that opinions accurately reflect the rules and doctrine that the court viewed as governing its decision.

The second category of scholarship falls under the banner of "empirical legal studies." This sort of work, which has slowly migrated from political science departments into legal academia, focuses on criteria that can be observed and quantified.[29] Rooted in legal realism and rational choice theory, it views judicial decision making as largely the product of political attitudes and as being as driven by ideology as legislative voting.[30] In its most basic form, the variables taken into

---

opinions, and commenting on their significance." *Id.*

25. *See id.* at 79–85.

26. *See id.* at 64.

27. *See id.* at 66.

28. Note that we have said "much" and not "all" or even "most." The array of work that appears in law reviews these days is far too varied to fit into these two categories.

29. This work's intellectual roots include legal realism, economic rational choice theory, and the behavioralist movement in political science. *See* ALBERT SOMIT & JOSEPH TANENHAUS, THE DEVELOPMENT OF POLITICAL SCIENCE: FROM BURGESS TO BEHAVIORALISM 177–78 (1967).

30. The most basic form is the attitudinal model:

account in the analysis are (1) judges' ideology, measured via some objective proxy such as party of the appointing president,[31] and (2) the ideological valence of a jurist's vote in a given case, also measured in a reductionist, objective way, such that, for example, any vote in favor of a criminal defendant will be regarded as liberal.[32] The research is quantitative in nature, using large-*n* studies and statistical methodology. Scholarship produced using these methods has established, at a minimum, that there is a relatively strong correlation between ideology and judicial behavior as measured in these ways.[33]

A. *Content Analysis*

Content analysis stands as something of a hybrid of these two methodologies.[34] It reflects both traditional legal scholarship's attention to texts and empirical legal studies' systematization. Although content analysis was not recognized as a genre prior to the publication of their article,[35] Hall and Wright found 134 law review articles published

---

> The attitudinal model represents a melding together of key concepts from legal realism, political science, psychology, and economics. This model holds that the Supreme Court decides disputes in light of the facts of the case vis-à-vis the ideological attitudes and values of the justices. Simply put, Rehnquist votes the way he does because he is extremely conservative; Marshall voted the way he did because he was extremely liberal.

JEFFREY A. SEGAL & HAROLD J. SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED 86 (2002).

31. More recent work has incorporated refinements such as including the political party of the judge's home-state senators into the measure of ideology. *See, e.g.*, Micheal W. Giles et al., *Picking Federal Judges: A Note on Policy and Partisan Selection Agendas*, 54 POL. RES. Q. 623, 624 (2001).

32. For example, much of this work is based on databases created by political scientist Harold Spaeth.

> Each case is given either a liberal or conservative code based on the nature of the prevailing party. So, for example, Spaeth codes cases involving criminal defendants as liberal if the defendant wins and conservative if the government wins; cases involving federal taxation, on the other hand, are coded as liberal if the government wins and conservative if the taxpayer prevails. Spaeth is—quite deliberately—uninterested in the content of the opinions.

Carolyn Shapiro, *Coding Complexity: Bringing Law to the Empirical Analysis of the Supreme Court*, 60 HASTINGS L.J. 477, 485 (2009).

33. *See, e.g.*, Jeffrey A. Segal, *Judicial Behavior*, in THE OXFORD HANDBOOK OF LAW AND POLITICS 26–28 (Whittington et al. eds., 2008); CASS R. SUNSTEIN ET AL., ARE JUDGES POLITICAL?: AN EMPIRICAL ANALYSIS OF THE FEDERAL JUDICIARY app. at 152 (2006); Frank B. Cross, *Collegial Ideology in the Courts*, 103 NW. U. L. REV. 1399, 1400 (2009).

34. *See* Hall & Wright, *supra* note 2, at 64.

35. They even noted:

between 1956 and 2006 that used content analysis.[36] Within that period, they found such studies being published with increasing frequency, a development they attribute in part to the availability of computerized legal databases and statistical software.[37] Studies employing the methodology have ranged across a broad swath of subject areas as well as "focus[ing] on questions of legal methods, judicial decision making, and statutory interpretation."[38] The work has appeared "in the very best law journals"[39] and seems "somewhat more likely to generate discussion and citation than law review articles more generally."[40]

Hall and Wright do not position themselves as unqualified advocates for the use of content analysis in legal scholarship. They instead regard it as providing another useful perspective, and aim to identify and weigh its benefits and drawbacks, and to generate a set of "best practices" to be used in the implementation of this approach.[41] Systematic content analysis, like any process that involves a process of categorization and coding, entails a certain amount of reductionism and glossing over of nuance.[42] And it is best employed in contexts where each document under assessment is entitled to equal weight, for the simple reason that the method is ill-suited to adequately account, for example, for cases with disproportionate influence within a body of law.[43] What results is a methodology that "can augment conventional analysis by identifying previously unnoticed patterns that warrant deeper study, or sometimes by correcting misimpressions based on ad hoc surveys of atypical cases."[44] Thus, while it generates results that are more objective, in the sense that others should be able to replicate them, and broad, because the methodology can more easily cover large swaths of cases, it tends toward shallowness, "trad[ing] the pretense of ontological certainty for a more provisional understanding of case law."[45]

> In project after project, legal researchers reinvent this methodological wheel on their own. The two of us, for instance, each learned how to do content analysis on the fly, feeling at first as if we each discovered something new until we learned that we had each done the same thing independently. We see now that many of our colleagues share the same sense of having found their own way.

Id. at 74–75.

36. *See id.* at 72 tbl.1.
37. *See id.* at 69–70.
38. *Id.* at 73.
39. *Id.* at 70.
40. *Id.* at 74.
41. *See id.* at 100–20.
42. *See id.* at 82–83.
43. *See id.* at 83–84.
44. *Id.* at 87.
45. *Id.*

Content analysis's primary focus on judicial opinions acts as both an advantage and a limitation. The advantage comes in that opinions matter in their own right. Prototypical empirical work relies on proxy measures for assessing the ideology of judges and case results. Because the content of judicial opinions matters both to the parties in a given dispute and to those for whom knowledge of the law is important, opinions themselves constitute a significant form of judicial behavior rather than standing as a proxy for some underlying phenomenon. Of course, opinions might also serve as proxies for underlying behavior—we care about the motivation behind opinions, and the extent to which the reasons provided in an opinion are the "real" reasons behind a decision.[46] This leads to perhaps the most significant limitation of content analysis—its ultimate dependence on the documents under study, which will most often be judicial opinions. For an analysis of the contents of judicial opinions to yield useful results, it must be the case that the opinions meaningfully reveal something about whatever is under study. Put differently, the reading and analysis of opinions will provide insight into the factors that drive decision making only to the extent that opinions actually relate the factors that in fact drive decision making. In this regard, consider two types of goals one might have in analyzing the content of opinions. The first is to learn something about the opinions as opinions. One could, for example, focus on the use of certain rhetorical strategies or otherwise analyze how judges choose to justify their decisions.[47] In this type of inquiry the sole concern is on the text, and not on some underlying phenomenon as to which the text is a mere window.[48] One can consider a court's use of a rhetorical device in an opinion without needing to make any assumptions about whether the court was, for example, sincere in using it.

---

46. *See generally* Micah Schwartzman, *Judicial Sincerity*, 94 VA. L. REV. 987 (2008). For other works directly addressing the topic of judicial candor, see generally Scott Altman, *Beyond Candor*, 89 MICH. L. REV. 296 (1990); Scott C. Idleman, *A Prudential Theory of Judicial Candor*, 73 TEX. L. REV. 1307 (1995); Robert A. Leflar, *Honest Judicial Opinions*, 74 NW. U. L. REV. 721 (1979); David McGowan, *Judicial Writing and the Ethics of the Judicial Office*, 14 GEO. J. LEGAL ETHICS 509 (2001); David L. Shapiro, *In Defense of Judicial Candor*, 100 HARV. L. REV. 731 (1987); Martin Shapiro, *Judges as Liars*, 17 HARV. J.L. & PUB. POL'Y 155 (1994); Nicholas S. Zeppos, *Judicial Candor and Statutory Interpretation*, 78 GEO. L.J. 353 (1989).

47. For some recent examples, see generally Keith Cunningham-Parmeter, *Alien Language: Immigration Metaphors and the Jurisprudence of Otherness*, 79 FORDHAM L. REV. 1545 (2011) (analyzing the use of immigration metaphors); Julie A. Oseid, *The Power of Metaphor: Thomas Jefferson's "Wall of Separation Between Church and State*," 7 J. ASS'N LEGAL WRITING DIRS. 123 (2010); Louis J. Sirico, Jr., *Failed Constitutional Metaphors: The Wall of Separation and the Penumbra*, 45 U. RICH. L. REV. 459 (2011).

48. This is not to suggest that style is divorced from substance. Metaphors, for example, are not simply ornamentation, but also shape, sometimes insidiously, the legal standards they are used to describe. *See, e.g.*, Sirico, *supra* note 47, at 459.

Most often, however, the researcher has a second type of goal in mind, which is to understand something for which the opinions are, in a sense, a proxy. That is, the content analyst views opinions as law in the Holmesian sense of informing predictions about what courts will do.[49] An opinion's value in this regard depends to a great degree on the correspondence between what it says and the judge's actual reasons for deciding the case.[50] As Hall and Wright recognize, a problem arises—for both content analysis and traditional legal scholarship—in that opinions might not consistently reflect those reasons.[51] This might be a product of cognitive limitations, because a judge might be unaware of, or unable to articulate fully, all of the relevant components of his decisional process.[52] It might be a product of insincerity or deceit, in that the opinion provides reasons that the judge recognizes are not the true factors motivating or explaining her decision.[53] Or it might result from a natural tendency to want to provide a strong justification for a decision already reached, such that the opinion highlights those aspects of the case that support the decision while minimizing those that do

---

49. *See* Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457, 461 (1897) ("The prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law."); *see also* K.N. LLEWELLYN, THE BRAMBLE BUSH: ON OUR LAW AND ITS STUDY 14 (4th prtg. 1973):

> But if I am right, finding out what the judges *say* is but the beginning of your task. You will have to take what they say and compare it with what they *do*. You will have to see whether what they say matches with what they do. You will have to be distrustful of whether they themselves know (any better than other men) the ways of their own doing, and of whether they describe it accurately, even if they know it.

50. As Professor Frederick Schauer points out, there is tension between the positions of Holmes and Llewellyn on this point:

> [I]f we were to undertake a statistical analysis of 'the law' in order best to engage in the process of predicting future legal outcomes, we would, in some form or other, look to identify the variables that had the greatest predictive value. These variables might, as Holmes suspects, be the variables of legal doctrinal categorization. But whether the variables were in fact what Holmes suspected—and desired—would be an empirical question, and it might turn out, as Llewellyn suspected to the contrary, that they were variables not likely to be identified from the opinions of the courts that reached those decisions.

Frederick Schauer, *Prediction and Particularity*, 78 B.U. L. REV. 773, 783–84 (1998).

51. Hall & Wright, *supra* note 2, at 100 ("The major limitation of content analysis—a limit that applies equally to traditional interpretive methods—is that one cannot treat as accurate and complete the facts and reasons given in opinions. Therefore, researchers should be cautious about the meanings they attach to observations made through content analysis.").

52. *See* Chad M. Oldfather, *Writing, Cognition, and the Nature of the Judicial Function*, 96 GEO. L.J. 1283, 1305–08 (2008).

53. *See* sources cited *supra* note 46.

not.[54] Whatever the reason, the consequence is that opinions might provide an incomplete or misleading picture of the decisional behavior they purport to reflect, such that content analysis-based efforts to understand or predict judicial behavior will merely reflect any systematic disconnect between opinions' depiction of law and decisional processes and their actual operation in practice.[55] Thus, if one wishes to gain an accurate understanding of what courts do, inquiry focused solely on opinions will generate a potentially incomplete and misleading picture.

Of course, this is not a fatal flaw. As alluded to above, the same difficulty arises in traditional legal scholarship. The fact that neither method is perfect does not mean that they cannot generate useful results. Moreover, the process of content analysis can itself incorporate steps designed to check for correspondence between the facts reported in the opinion and the actual facts of the case.[56] A researcher with enough information about a case could independently measure the extent to which an opinion accurately depicts the underlying dispute. Such an analysis could involve comparison of the parties' briefs to the opinion, or it might extend more broadly to include the analysis of all or portions of the record as well as lower court opinions.[57]

---

54. Professor Dan Simon has explored this phenomenon in connection with his analysis of the seeming disconnect between Justice Cardozo's opinions, which give a "distinct sense of obvious correctness" and are "cast in the mold of formalism," and his off-bench descriptions of the judicial process, which depict the judge as faced with tasks that "are complex, difficult, and replete with clashes between seemingly irreconcilable opposites." Dan Simon, *The Double-Consciousness of Judging: The Problematic Legacy of Cardozo*, 79 OR. L. REV. 1033, 1043, 1046 (2000). Professor Simon attributes this not to conscious duplicity, but to "the fact that closure is a naturally occurring cognitive phenomenon that accompanies mental tasks of the kind involved in legal decision-making." *Id.* at 1065. "[E]ven in the face of complex, difficult, underdetermined tasks, people ultimately experience their decisions as being solidly determined by the arguments and thus singularly correct." *Id.*; *see also* Dan Simon, *Freedom and Constraint in Adjudication: A Look Through the Lens of Cognitive Psychology*, 67 BROOK. L. REV. 1097, 1100–01 (2002).

55. *See* Hall & Wright, *supra* note 2, at 99.

56. *Id.* at 97–98. Hall and Wright report that several of the studies they looked at incorporated such steps via close readings of opinions or comparison of appellate majority opinions with trial court or dissenting opinions. *See id.* at 97–98 & nn.139–40. Those they reference include Robert A. Hillman, *Questioning the "New Consensus" on Promissory Estoppel: An Empirical and Theoretical Study*, 98 COLUM. L. REV. 580 (1998); Joseph A. Ignagni, *U.S. Supreme Court Decision-Making and the Free Exercise Clause*, 55 REV. POL. 511 (1993); Kimberly D. Krawiec & Kathryn Zeiler, *Common-Law Disclosure Duties and the Sin of Omission: Testing the Meta-Theories*, 91 VA. L. REV. 1795 (2005); Reed C. Lawlor, *Fact Content Analysis of Judicial Opinions*, 8 JURIMETRICS J. 107 (1968); Richard A. Posner, *A Theory of Negligence*, 1 J. LEGAL STUD. 29 (1972); Mark J. Richards & Herbert M. Kritzer, *Jurisprudential Regimes in Supreme Court Decision Making*, 96 AM. POL. SCI. REV. 305 (2002).

57. *See, e.g.*, Richard A. Posner, *Judicial Biography*, 70 N.Y.U. L. REV. 502, 522 (1995) ("No evaluative study of an individual judge is complete until his opinions are compared with

Notice, however, that this sort of inquiry seems likely to depart from the level of rigor and systematization associated with content analysis. Well-conducted content analysis aims for replicability and uses coding categories that can be consistently applied across a body of cases.[58] An assessment of the fit between a court's depiction of the facts and the actual facts seems inevitably to require the exercise of judgment because the researcher must determine whether, for example, a court's failure to mention what the researcher believes to have been a key fact fell outside the proper bounds of the court's discretion. In other words, the researcher in such a situation necessarily chooses to substitute her own view of what full candor and sincerity would require, with that view being a product of contestable judgment calls about the significance of certain facts in light of applicable legal standards rather than something that can be made by reference to objective criteria.

The process described in the preceding paragraph sounds more like traditional legal scholarship in its combination of deep textual engagement and normative evaluations. In addition to reintroducing the problem of subjectivity, this sort of deep comparison would be (as we can attest based on the efforts described below)[59] incredibly time- and labor-intensive. Determining whether a court was faithful to the record and the parties' arguments would take at least as much time as was required to reach the initial decision. Addressing these problems requires the development of proxy measures. Just as political scientists have relied on measures such as party of appointing president as a stand-in for more nuanced measures of judicial ideology, so might we seek such measures for assessing the correspondence between opinions' depiction of cases and the underlying reality.[60] To be sure, even such a measure would not remove all difficulties. An opinion's fidelity to facts might be normatively desirable in its own right, as might responsiveness to the arguments in the parties' briefs.[61] But even these are proxies for what some might regard as the true underlying concern: the extent to which an opinion fully and accurately reflects the court's decisional process.[62] That, of course, will remain known only to the judge, and even then only to the extent that true self-knowledge is possible.

Still, proxy-based measures of correspondence or responsiveness would provide at least a tentative answer to the claim that content

---

the lawyers' briefs. This is necessary in order to determine not only the judge's 'value added' but also his scrupulousness with respect to the facts of record and the arguments of the parties.").

58. *See* Hall & Wright, *supra* note 2, at 105–09.

59. *See infra* Subsection III.D.1.

60. *See infra* Section III.D for an exploration of the use of word and citation counts as such proxies.

61. *See infra* Part II for discussion and development.

62. *See supra* notes 49–50 and accompanying text.

analysis-based and traditional legal scholarship are suspect based on their assumptions about the accuracy with which opinions relate facts and reasoning.[63] If the results of studies employing such measures suggest high correspondence between opinions and the remaining corpus of text in a case, that would suggest that opinions faithfully reflect the application of legal standards to cases. While it is unlikely that any such analysis could provide conclusive proof as to any given court's level of faithfulness, comparative analyses of courts would allow at the very least for relative assessments. As we develop below, automated content analysis holds great promise in addressing these methodological difficulties.

B. *Automated Content Analysis*

Adjudication, especially at the appellate level, is an almost entirely text-based practice. Even the spoken portions of lower-court proceedings are reduced to text in the form of a transcript of the proceedings. There are exceptions, primarily photographs and video recordings, but for the most part an appellate case—or at least the visible manifestations of an appellate case—consists of a collection of words. These words have become increasingly accessible and manipulable over the past several decades. Westlaw, Lexis, and other databases have of course long provided access to judicial opinions and, to a lesser extent, briefs. Courts themselves are slowly making more information available electronically.[64] It is conceivable, and perhaps inevitable, that court records, including transcripts and documentary evidence, will be readily available electronically in the relatively near future.

It should accordingly come as no surprise that scholars have begun to apply computational methodologies to the analysis of adjudication. The ability to "read" opinions through automated content analysis software offers the prospect of a different, and in some respects deeper,

63. Professor Pamela Corley made a similar observation in reporting her use of plagiarism software to examine the correspondence between briefs and majority opinions at the Supreme Court level:

> If the justices are motivated to reach legally sound decisions, they are likely to be influenced by the persuasiveness of legal argumentation. Thus, the arguments presented to the Court in the briefs are part of a legal model of decision making, one in which a quality argument can persuade the justices to interpret precedent in a particular way and to develop new legal rules, both of which affect decision making in future cases.

Pamela C. Corley, *The Supreme Court and Opinion Content: The Influence of Parties' Briefs*, 61 POL. RES. Q. 468, 468–69 (2008) (citations omitted).

64. *See generally* Lynn M. LoPucki, *Court-System Transparency*, 94 IOWA L. REV. 481 (2009) (explaining the federal court system's move towards electronic records).

exploration of the behavior of individual judges and justices. Early efforts fall into three general categories: (1) studies designed to explore questions concerning the authorship of judicial opinions; (2) those employing computational methods as refinements to the traditional machinery of empirical legal studies; (3) and those exploring the relationship between briefs and judicial opinions. None of these is unique in the sense that they all have roots in prior programs of scholarship. Each nonetheless holds out the promise of a unique perspective on the judicial process.[65]

1. Authorship of Judicial Opinions

Questions concerning authorship of judicial opinions, such as which judges write their own, which justice is the primary author of a per curiam opinion, and whether one judge consistently ghost wrote for another, are inherently interesting to those who pay attention to courts. As Professors Stephen Choi and Mitu Gulati point out, there are also scholarly payoffs to such inquiries. Conceiving of the judiciary as presenting two levels of agency problems (judges as agents of the polity, and clerks as agents of the judges), Choi and Gulati contend that information about authorship could "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."[66] Knowledge about whether and to what extent a judge is involved in the opinion-writing process, they suggest, can be used as part of a comprehensive assessment of judicial quality.[67]

Choi and Gulati also note the usefulness of the information to ongoing scholarly efforts to understand judicial behavior. A judge's tendency toward authorship rather than editorship might serve as an explanatory variable with respect to a variety of factors such as "voting patterns, citation rates and styles, invocation rates, publication patterns, independence levels, and choices about styles of argument (for example,

65. There is a sense in which this research is incredibly rudimentary. Most forms of automated content analysis put computers to work more or less as "dumb clerks," albeit ones that are incredibly fast and accurate. Robert L. Stevenson, *In Praise of Dumb Clerks: Computer-Assisted Content Analysis*, in THEORY, METHOD, AND PRACTICE IN COMPUTER CONTENT ANALYSIS 4 (Mark D. West ed. 2001). "The most promising change in content analysis is the ability to search massive quantities of materials instantly. While this may reduce the depth of analysis, it increases dramatically the breadth of a study. By itself, this is enough to praise the computer's value as a dumb clerk." *Id.* at 5.

66. *See* Choi & Gulati, *Which Judges Write Their Opinions*, *supra* note 7, at 1083.

67. As discussed below, Professors Choi and Gulati have explored the concept of measuring judicial quality in a series of recent articles. *See infra* notes 68–77 and accompanying text.

whether one prefers the use of multifactor balancing tests)."[68] There may be payoffs in terms of institutional design as well. They suggest that if, for example, analysis reveals that judges tend to rely more heavily on their clerks to draft opinions in certain subject areas, that could help shape our views as to the relative desirability of specialized courts.[69]

Choi and Gulati also acknowledge that information about judges' involvement in the opinion writing process has potential downsides. One objection to their inquiry proceeds from the view that it does not matter whether a judge is the primary author of an opinion, perhaps on the ground that the judge's job is primarily that of deciding, with the justification for the decision being of substantially less importance.[70] Of course, that view stands in contrast to the assumption, outlined above, underlying both content analysis and traditional legal scholarship to the effect that judicial opinions accurately reflect law and judicial decision-making. A second, and more significant, objection concerns the problem of imperfect measurement, which creates the possibility that some judges will be inaccurately categorized as editors when they are really authors.[71] In their study, Choi and Gulati used various tests from computational linguistics in an effort to determine which federal court of appeals judges write their own opinions.[72] The basic premise underlying this sort of inquiry is that at least some writers have stylistic fingerprints, which reveal themselves in patterns of word usage.[73] Choi

68. Choi & Gulati, *Which Judges Write Their Opinions*, *supra* note 7, at 1090.

69. *See id.*

70. *See id*. at 1094–95. Choi and Gulati are (rightly, in our view), underwhelmed by this objection. Even if one remains skeptical of the proposition that judicial opinions accurately report judicial reasoning, it seems unlikely to be the case that opinions tell us nothing useful. To be functional, a precedent-based system seemingly requires not only that written opinions exist, but that they be given authoritative weight. *See* James Boyd White, *What's an Opinion For?*, 62 U. CHI. L. REV. 1363, 1366 (1995) ("Rough prediction, then, and with it a certain kind of argument, might be possible in [a system without judicial opinions], but the invocation of the past as authority is a different matter and seems to require the existence of the judicial opinion, or something like it."). Moreover, "when we are in the pit of actual application, we will discover that it is not what the Supreme Court held that matters, but what it *said*." Frederick Schauer, *Opinions as Rules*, 53 U. CHI. L. REV. 682, 683 (1986) (reviewing BERNARD SCHWARTZ, THE UNPUBLISHED OPINIONS OF THE WARREN COURT (1985)).

71. *See* Choi & Gulati, *Which Judges Write Their Opinions*, *supra* note 7, at 1095–96. The objection as we have characterized it is not quite the same as what Choi and Gulati report the commentators to their article have made. That objection was that there is a better source of information concerning judicial authorship, namely the judges themselves. As Choi and Gulati convincingly argue, however, there are plenty of reasons to think that judges will not be entirely forthcoming on the question. *See id.*

72. *See id.* at 1103–08.

73. *Id.* at 1099. Although efforts to determine authorship using the methods on other types of text date back to the 1930s, their use has expanded with increases in computational power. For a listing of significant words, see *id.* at 1101 & n.68.

and Gulati posited that, while individual judges are unlikely to have discernible styles, there is likely to be a difference between judicial style and law clerk style.[74] The more that a judge's opinions manifest inconsistent stylistic markers, they reason, the less likely that judge is to be the author of the opinions.[75] Although their use of generic computational linguistics methodologies applied to opinions without regard to subject matter failed in the sense of not being able to identify the three judges (Boudin, Easterbrook, and Posner) whom they knew to be among those who write their own opinions, they experienced somewhat greater success when they modified their methodology by controlling for subject matter and taking account of features such as citation practices[76] and average length of opinion.[77]

In another recent study,[78] Professors Jeffrey Rosenthal and Albert Yoon, employing methods similar to those used in projects analyzing the authorship of the *Federalist Papers*[79] and Shakespeare's plays,[80] investigated the commonly held understanding that Supreme Court Justices have in recent decades placed growing reliance on their law clerks in the opinion-writing process.[81] Their methodology involved examination of frequencies of the use of "function words"—common words the usage of which can constitute something of an authorial fingerprint independent of subject matter.[82] Similar to Choi and Gulati, they posited that a Justice whose writing style showed greater variability was likely to have delegated a greater portion of opinion-writing responsibility to clerks.[83] Their results were generally consistent with prior understandings of the extent to which various individual Justices have relied on their clerks, as well as with the proposition that such reliance has increased over time.[84] They were also able to predict opinion authorship with a relatively high degree of accuracy.[85]

---

74. *Id.* at 1102. Specifically, they posit that judges are likely to be more confident in their analyses than clerks and that, as a result, judge-written opinions will be shorter and include fewer citations and footnotes. *Id.*

75. *See id.* at 1103.

76. *See id.* at 1111–13.

77. *See id.* at 1116–20.

78. Jeffrey S. Rosenthal & Albert H. Yoon, *Judicial Ghostwriting: Authorship on the Supreme Court*, 96 CORNELL L. REV. 1307 (2011).

79. *See generally* FREDERICK MOSTELLER & DAVID L. WALLACE, INFERENCE AND DISPUTED AUTHORSHIP: THE FEDERALIST (1964) (using statistical methods to determine authorship of the *Federalist Papers*).

80. For an extensive survey of sources, see Choi & Gulati, *Which Judges Write Their Opinions*, *supra* note 7, at 1097–98.

81. Rosenthal & Yoon, *supra* note 78, at 1311–12.

82. *Id.* at 1312.

83. *Id.*

84. *Id.*

85. *See id.* at 1337.

Finally, Russell Smyth and his colleagues applied these methods to investigate longstanding rumors of ghostwriting on the High Court of Australia.[86] Various evidence had suggested that Sir Owen Dixon authored a number of opinions issued under the names of his colleagues Sir Edward McTiernan and Sir George Rich.[87] Smyth's team concluded, with a high degree of confidence, "that about four per cent of McTiernan's judgments and 18 per cent of Rich's judgments were very likely authored by Dixon."[88] They argue that their findings are not merely of value as matters of "historical curiosity," but because they shed light on questions of judicial ethics and the reliability of attributions of authorship.[89]

2. Refining Empirical Legal Studies

As noted above, the bread and butter of quantitative empirical legal studies has been work that examines the relationship between judicial ideology and decision making.[90] As this work has evolved, scholars have refined it primarily by reworking measures of judicial ideology. For the most part, however, decisions continue to be coded in terms of binary, liberal/conservative categories.[91] As Professor Michael Evans and his colleagues have pointed out, automated content analysis holds out the promise of enabling considerably more nuanced coding of the results of decision making.[92] It also offers value because of its efficiency, transparency, and replicability.

Initial efforts to use computational methods have relied primarily on a program called Wordscores. A basic description of the method is as follows:

> The process begins with the selection of "reference" (training) texts, written with a known position along a dimension of interest (e.g., ideology, policy issue field, etc.). The Wordscores program then generates a word frequency matrix for every word (feature) in the reference texts. Based on the relative frequencies of each word in the reference texts and the values assigned to those documents, word scores are then calculated to represent the association between words and each document. . . . Finally, text scores

86. *See generally* Yanir Seroussi, Russell Smyth & Ingrid Zuckerman, *Ghosts from the High Court's Past: Evidence from Computational Linguistics for Dixon Ghosting for McTiernan and Rich*, 34 U.N.S.W. L.J. 984 (2011).

87. *Id.* at 985–86.

88. *Id.* at 1003.

89. *Id.* at 987.

90. *See supra* notes 29–33 and accompanying text.

91. *See, e.g.*, Shapiro, *supra* note 32, at 485; Evans et al., *supra* note 7, at 1020–22.

92. Evans et al., *supra* note 7, at 1020–21.

> are computed for unread, uncharacterized "virgin" texts (the test examples), characterizing them with respect to the reference documents. The score given to each virgin text is simply the average of all word scores for all scored words within the text.[93]

Early studies have shown the promise of the methodology, but also make it apparent that much work remains to be done. In one study, Professors Kevin McGuire and Georg Vanberg used Wordscores in an effort "to extract valid policy positions from the text of written opinions for a series of decisions in the areas of religion and search and seizure."[94] They analyzed Supreme Court opinions dealing with three issue areas, having concluded that it was necessary to confine the inquiry to specific issues because discussions of different issues will use different language.[95] In each area, they used two Supreme Court opinions (of reasonably clear ideological valence) as reference texts, and then scored a series of other opinions (the general ideology of which they also knew beforehand).[96] They found that the method was unreliable when applied to both majority and dissenting opinions; in other words, it could not accurately distinguish between liberal and conservative opinions.[97] It did, however, do a reasonably good job of marking the relative position of opinions within groups of exclusively liberal or conservative opinions.[98]

In a similar study, Michael Evans and his colleagues undertook to assess

> the performance of the Wordscores and Naïve Bayes methods at analyzing U.S. Supreme Court litigant and amicus curiae briefs. Specifically, we examine the ability of the two approaches to (1) accurately classify the ideological position of the various legal briefs, (2) identify words from those briefs that are distinctive to opposing ideological positions in enhancing interpretive analysis, and (3) detect patterns in language usage over time by advocates on a single issue.[99]

---

93. *Id.* at 1014. More, including the papers describing and implementing the methodology, is available at http://www.tcd.ie/Political_Science/wordscores/index.html.

94. KEVIN T. MCGUIRE & GEORG VANBERG, MAPPING THE POLICIES OF THE U.S. SUPREME COURT: DATA, OPINIONS, AND CONSTITUTIONAL LAW 2 (2005), *available at* http://www.unc.edu/~kmcguire/papers/McGuire_and_Vanberg_2005_APSA_Paper.pdf.

95. *Id.* at 14.

96. *See id.* at 15–28.

97. *See id.* 29–30 & n.12.

98. *See id.* at 28.

99. Evans et al., *supra* note 7, at 1023.

With respect to the first, the best methods accurately characterized briefs between 80–90% of the time.[100] "These results present evidence for the ability of automated content analysis techniques to classify the ideological positions of legal texts and point to the utility of computational techniques in general."[101]

In an early effort to take these sorts of inquiries beyond measurements of ideology, Professors Robert Howard and Joseph Smith attempted to test the Supreme Court's receptiveness to originalist arguments.[102] They used Wordscores and compared the results it generated against the coding of an existing database.[103] Their use of Wordscores was more limited than that of other researchers. Rather than assessing all the words in the documents they analyzed, they simply used the program to count the frequencies of four phrases.[104] In general, they concluded that "computers can characterize legal briefs, and that these characterizations are comparable to those of human coders."[105]

3. Exploring the Relationship Between Briefs and Opinions

Some work in both the traditional and empirical genres has sought to assess the impact of briefs and other forms of advocacy on judicial decision making. Work from a traditional, doctrinal perspective tends to involve a close, qualitative reading and comparison of briefs to opinions.[106] Empirical projects attempt quantification. As is generally true of the empirical research described above, this work has proceeded without engaging with the content of the briefs. So, for example, one major study measures the influence of amicus curiae briefs[107] primarily by assessing whether the presence of such briefs bears a relationship to the outcome in a case.[108] Another researcher has conducted a number of studies in which he has focused on such factors as a lawyer's prior

100. *See id.* at 1028.

101. *Id.*

102. *See* ROBERT M. HOWARD & JOSEPH L. SMITH, THE NEXT FRONTIER IN LEGAL ANALYSIS: COMPUTER-AIDED CONTENT ANALYSIS OF LEGAL TEXTS 8–9 (2008), *available at* http://www.scribd.com/doc/36768599/Howard-Smith-APSA-08.

103. *See id.* at 8–12.

104. *Id.* at 8.

105. *Id.* at 14.

106. *See, e.g.*, Clay Calvert, *Punishing Public School Students For Bashing Principals, Teachers & Classmates In Cyberspace: The Speech Issue the Supreme Court Must Now Resolve*, 7 FIRST AMEND. L. REV. 210, 247 (2009).

107. These are briefs that are not filed by the parties to the case, but rather by "friends of the court"—that is, groups that have an interest in the resolution of the legal issue before the court and that seek to provide the court with input on aspects of the issue beyond what the parties themselves are likely to provide.

108. *See* Joseph D. Kearney & Thomas W. Merrill, *The Influence of Amicus Curiae Briefs on the Supreme Court*, 148 U. PA. L. REV. 743, 749 (2000).

experience having a case before the Supreme Court or the lawyer's performance at oral argument (the latter based on the grades that Justice Blackmun gave to the advocates who appeared before the Court while he was on it).[109]

Two studies have used automated methods in an effort to measure the influence of briefs on the Supreme Court. In the first, Professor Pamela Corley used plagiarism software to compare the briefs with majority opinions issued in the 2002–2004 terms.[110] She set the software to search for phrases of six words or more and text strings of 100 characters or more, as well as to skip over citations and to identify phrases separated by up to two nonmatching words (so as to identify minor edits).[111] The result of this inquiry was to reveal a surprising degree of overlap between briefs and opinions. "The mean percentage of the majority opinion directly borrowing from the appellants' and respondents' briefs was 10.1 (standard deviation of 5.7) and 9.4 (standard deviation of 5.4), respectively."[112] In contrast, running the same comparison between thirty randomly selected opinions from her data set and the opinions in ten percent of the cases cited in those cases generated a mean plagiarism rate of 1.1%.[113] Her further analyses revealed that opinions borrowed a greater percentage of briefs that she determined were of high quality or were ideologically compatible with the Court, or in cases that were not politically salient.[114] Case complexity, in contrast, bore no relation to the level of borrowing.[115]

Kevin McGuire and his colleagues used Wordscores to test the hypothesis that briefs to the Supreme Court will target the "median Justice"—that is, the Justice who is at the ideological center of the Court.[116] Advocates will target this Justice because his vote will be necessary to win a majority of the Court, and as a consequence, they posit that opinions authored by the median Justice will be more likely to reflect the arguments made in the winning brief than will those authored

109. *See* Andrea McAtee & Kevin T. McGuire, *Lawyers, Justices, and Issue Salience: When and How Do Legal Arguments Affect the U.S. Supreme Court?*, 41 LAW & SOC'Y REV. 259, 263 (2007). In similar fashion, Epstein, Landes, and Posner examined the relationship between the results in Supreme Court cases and the questioning of counsel at oral argument. *See* Lee Epstein, William M. Landes & Richard A. Posner, *Inferring the Winning Party in the Supreme Court from the Pattern of Questioning at Oral Argument*, 39 J. LEGAL STUD. 433, 437 (2010).

110. Corley, *supra* note 63, at 469.

111. *Id.* at 471.

112. *Id.* at 472.

113. *Id.*

114. *Id.* at 474.

115. *Id.*

116. KEVIN T. MCGUIRE, GEORG VANBERG & ALIXANDRA B. YANUS, TARGETING THE MEDIAN JUSTICE: A CONTENT ANALYSIS OF LEGAL ARGUMENTS AND JUDICIAL OPINIONS 3–4 (2011), *available at* http://www.unc.edu/%7Ekmcguire/papers/targeting_median.pdf.

by other Justices.[117] Taking the insight one step further, they posited that the similarity between the majority opinion and the arguments in the winning brief should decrease in proportion to the authoring Justice's distance from the ideological median of the Court.[118] Their results were consistent with these hypotheses. They found that, in general, opinions were more similar to the brief of the prevailing party than the losing party, and that, although their sample size was relatively small, "there clearly appears to be [a] positive relationship between a justice's ideological proximity to the Court's median and the similarity of her opinions to the winning parties' briefs."[119]

As we have suggested above, we believe that content analysis directed toward checking for correspondence among briefs and opinions would provide an at least partial answer to the concern that opinions do not accurately reflect either the facts of the cases being decided or the court's underlying decisional processes.[120] It seems reasonable to expect that the degree of correspondence between opinions and briefs will, at least as a general matter, increase along with the extent to which opinions accurately reflect the facts and arguments actually presented in the underlying dispute. In our adversary system, the briefs in an appellate case are the primary conduit through which the court gets its information about the dispute, and the other information in the record appears there only as a product of the adversaries' efforts.[121] Substantial departures are thus at least suggestive of the conclusion that the court is resolving what might be characterized as a different case than the one put before it.

While the question of candor[122]—the extent to which opinions reflect the court's true reasoning—is trickier, at the margins at least one would expect an opinion grounded in the arguments made by the parties to be an opinion that accurately reflects a decision actually grounded in those arguments and the authorities invoked in support of them. This will not be an absolute, or even necessarily strong, relationship. Most everyone would resist the notion that the content of documents is the sole determinative factor in a judicial decision.[123] In many, perhaps most, cases the judge will bring her own understanding of an area of law to a case, as well as her own intuitions about what the correct decision is. As a result, things extraneous to the content of a brief will often matter, such as the subject matter of the case, the identity of the

117. *Id.* at 3.
118. *See id.* at 6.
119. *Id.* at 13.
120. *See supra* notes 56–57 and accompanying text.
121. *See supra* notes 8–12 and accompanying text.
122. *See supra* note 46 and accompanying text.
123. *See supra* note 33 and accompanying text.

brief's author, the politics of the issues involved, the specific facts of a case, and so on, and may render the role the documents play relatively minor. There are differing intuitions as to the roles that judicial opinions should and do play in legal analysis. Although the official story is that judicial opinions provide a more or less accurate window into the rationale underlying a decision, the cynic would contend that opinions are merely after-the-fact efforts to give cover to a decision already made, possibly on other grounds, and that they do not reliably tell us anything useful about the process of judicial reasoning.

II. THE CASE FOR MEASURING JUDICIAL RESPONSIVENESS

We have argued in the preceding Part that research comparing briefs and other litigation documents to judicial opinions holds the promise of ameliorating, at least partially, one of the methodological shortcomings of content analysis by enabling us to determine whether judicial opinions accurately reflect the cases they discuss. In this Part, we expand on that insight by discussing more specifically the measurement of what we call "judicial responsiveness." The discussion unfolds in two Sections. Section A justifies the measure by tracing out the contours of the normative case for responsiveness as a feature of legitimate adjudication. Although we attempt in this Article to remain agnostic concerning the ultimate validity of the normative case in its particulars, it nonetheless makes sense to outline it in order to develop an appreciation for the centrality of responsiveness to the adjudicative process. Section B considers measures of responsiveness as a useful window into larger efforts to study the judiciary. Information about the relative responsiveness of courts and judges can help inform both academic inquiry into the nature and processes of judging as well as research directed toward questions of institutional design.

A. *Responsiveness as a Normatively Desirable Feature of Adjudication*

Lon Fuller argued, in his classic article *The Forms and Limits of Adjudication*, that the defining characteristic of adjudication lies not in the attributes of the judge, but rather in that it is a process based in reasoned argumentation.[124] On this view, the key to legitimacy in adjudication is not whether the judge is impartial, learned in the law, or otherwise possessed of some specific attribute. Nor is it the case that one is engaged in "adjudication" simply because one has resolved a dispute. Instead, Fuller contended, legitimate adjudication can take place only "within an institutional framework that is intended to assure to the disputants an opportunity for the presentation of proofs and

124. *See* Fuller, *supra* note 8, at 363.

reasoned arguments."[125] For Fuller, then, party participation in the decision making process is essential to legitimacy.

Indeed, Fuller took the point a step further, arguing that legitimacy requires that the judge should strive to render a decision as completely as possible on the grounds the parties have argued.[126] This is so, in part, for the simple reason that the parties would, if they knew that the judge were going to base his decision on some ground mentioned by a party in passing or not at all, address their arguments differently. But Fuller also suggested that something more fundamental is at work. The logic of a system that depends on party participation also demands that the resulting decisions be responsive to the specific contentions raised as part of that participation.[127] While it may never be possible for a court to base its decision purely on what the parties have put before it, Fuller argued that

> this is no excuse for a failure to work toward an achievement of the closest approximation of it. We need to remind ourselves that if this congruence is utterly absent—if the grounds for the decision fall completely outside the framework of the argument, making all that was discussed or proved at the hearing irrelevant—then the adjudicative process has become a sham, for the parties' participation in the decision has lost all meaning.[128]

The idea that courts should base their decisions on the grounds offered by the parties has appeal beyond the realm of legal philosophy. To understate the point, practicing lawyers dislike it when courts resolve issues on grounds not raised by the parties, recharacterize the arguments raised by the parties, ignore certain arguments (or components of arguments), and the like.[129] This animosity is perfectly understandable. The lawyer's role within the adversary system calls for the presentation

125. *Id.* at 365.
126. *See id.* at 364.
127. *See id.*
128. *Id.* at 388.
129. One appellate advocate put the point as follows:

> Appellate advocates hope that the appellate court will address, somewhere in the opinion, all issues that the parties have raised. The failure to do so suggests that the court reviewed the matter so quickly that it missed an issue or saw the issue but then forgot to address it in the written opinion. This apparent lack of care undermines confidence in the outcome. It does so for both sides, although it is particularly difficult for the losing side to accept a decision when the court failed to discuss all issues.

Mary Massaron Ross, *Reflections on Appellate Courts: An Appellate Advocate's Thoughts for Judges*, 8 J. APP. PRAC. & PROCESS 355, 362 (2006).

of the best arguments that the lawyer can conjure up on his client's behalf. If he is the one pressing a claim, he has come up with what she understands to be the best grounds for concluding that that claim should be successful.[130] A court's failure to engage with those grounds will be, at a minimum, disappointing.

Of course, to say that the scope of judicial decision making should be driven primarily by the parties' arguments is not to say that it should be entirely so. There are reasons why a court can and should depart from strong responsiveness. As Professor Amanda Frost points out, courts' lawmaking responsibilities often counsel in favor of departing from the precise terms of the parties' arguments when necessary to preserve the coherence and integrity of legal standards.[131] Commentators such as Abram Chayes and Owen Fiss have observed that courts must sometimes account for the interests of parties who are not involved in the immediate lawsuit, but who will nonetheless be affected by its resolution.[132] Even Fuller recognized that performance of the judicial role—indeed, what he recognized as exemplary performance—will occasionally involve the judge seeing things that the parties did not see, "bring[ing] to clear expression thoughts that in lesser minds would have remained too vague and confused to serve as adequate guideposts for human conduct" or "devis[ing] a solution that will reconcile and bring into harmony interests that were previously in conflict."[133] At least in some instances, then, strong responsiveness might constitute a failing rather than a virtue, stemming from a lack of

130. This is no doubt a somewhat idealized conception of the advocate's role. There are certainly some advocates who come to the court with a dispute the way one would approach a wise elder—that is, seeking insight from the court in addition to the application of logic. For example, in an argument a few years ago before the United States Court of Appeals for the Seventh Circuit, counsel for a criminal defendant attempted an argument in the face of a U.S. Supreme Court decision that he acknowledged was contrary to his position. *See* United States v. Johnson, 123 F. App'x 240 (7th Cir. 2005). His pitch to the court, in substantial part, consisted of the expression of his hope that the court could find a way to distinguish the case. Audio of the argument is available at http://www.ca7.uscourts.gov/tmp/J20L0KAH.mp3.

131. *See* Frost, *supra* note 9, at 501.

132. *See* Abram Chayes, *The Role of the Judge in Public Law Litigation*, 89 HARV. L. REV. 1281, 1311–12 (1976); Owen M. Fiss, *The Supreme Court, 1978 Term—Foreword: The Forms of Justice*, 93 HARV. L. REV. 1, 24–26 (1979).

133. Lon L. Fuller, *An Afterword: Science and the Judicial Process*, 79 HARV. L. REV. 1604, 1619 (1966). Though Fuller regards the judicial ideal as involving as little of the judge's predispositions as possible, he is under no illusions that reality can reflect this ideal:

> It would be foolish to assert that when judges are engaged in solving problems all of their personal attitudes and values become dissipated in a bright glow of objectivity. The final solution may well be skewed in one direction or another by something that may be termed a personal or collegial predilection.

Id.

effort or imagination rather than exemplary performance of the judicial role. Across the run of cases, however, some relatively strong version of responsiveness would seem most consistent with prevailing conceptions of proper judging.

B. *Responsiveness as a Window into Questions of Institutional Design and Process*

Quite apart from whether responsiveness is, in some relatively unqualified sense, necessary to legitimate adjudication, its study can benefit other strands of scholarly and practical inquiry. First, and at the most basic level, the study of responsiveness can help us to understand how the judiciary works. One might expect, for example, that courts at different levels of the judicial hierarchy would exhibit differing levels of responsiveness. Because courts have more responsibility for law development the farther one moves up the judicial hierarchy, research would likely show decreasing levels of responsiveness in higher courts. Such techniques might also enable studies assessing whether, as is often contended, caseload pressures have affected the manner in which judges do their work.[134] Research examining briefs and opinions from different time periods might show that the relationship between courts and adversaries has changed as judges face greater workloads and have delegated increasing responsibility to law clerks.[135]

In addition, large-scale implementation of a responsiveness measure will potentially provide results that can inform ongoing debates concerning the role of ideology in judicial decision making. Most of the quantitative empirical work focusing on the judiciary has been, and continues to be, of the sort that stems from the "attitudinal" and "strategic" models of judging developed by political scientists.[136] Stated generally, the focus of that work is on assessing the correlation and potential causal relationship between judges' ideological preferences and their decision making.[137] Many in the legal academy (and in the legal world more broadly) have resisted that work's suggestion that ideology drives decision making, and have sought to demonstrate that more traditionally legal factors explain the bulk of judicial behavior.[138]

---

134. *See, e.g.*, Richman & Reynolds, *supra* note 11, at 274–75.

135. For an overview of both phenomena and consideration of the potential consequences, see RICHARD A. POSNER, THE FEDERAL COURTS: CHALLENGE AND REFORM 124–59 (1996) [hereinafter POSNER, FEDERAL COURTS].

136. *See* SEGAL & SPAETH, *supra* note 30, at 312–26 (evaluating the attitudinal model); LEE EPSTEIN & JACK KNIGHT, THE CHOICES JUSTICES MAKE 10–11 (1998) (reviewing the strategic model).

137. *See id.* at 10.

138. For a discussion of these critiques, see Brian Z. Tamanaha, *The Distorting Slant in Quantitative Studies of Judging*, 50 B.C. L. REV. 685, 737–39 (2009).

Assessments of responsiveness would potentially shed further light on this debate, at least insofar as one accepts the proposition that a highly responsive decision is less likely to be the product of ideology than a relatively unresponsive decision. In this regard our work dovetails with prior work done by Stephen Choi and Mitu Gulati, who investigated for political bias in judges' citation practices.[139] Their working theory was that judges have considerable freedom in choosing what authorities to cite, such that a federal circuit judge's choice to cite an opinion authored by a judge from another circuit might reveal underlying biases that the result-focused inquiry of most empirical research might overlook.[140] Our theory is that the extent to which a judge exercises that freedom by citing authorities other than those relied upon by the parties also tells us something significant about that judge's tendencies. A judge who focuses primarily on the authorities offered by the parties arguably leaves less room for her ideological or other biases to manifest themselves.

Second, assessments of responsiveness can inform more normatively oriented scholarship, such as debates over questions of process and institutional design. For example, a key component of the debate over the device of the "unpublished" opinion is the suggestion that such opinions are justifiable because they involve the creation of no law, and thus need only to speak to the parties. A measure of responsiveness would allow for assessment of whether unpublished opinions actually are, as this justification suggests, relatively more focused on the parties' contentions than their published counterparts.[141]

A measure of responsiveness also might be added to the mix of factors employed in recent efforts to assess judicial quality, and could serve as a basis for comparisons of courts and individual judges.

---

139. Stephen J. Choi & G. Mitu Gulati, *Ranking Judges According to Citation Bias (As a Means to Reduce Bias)*, 82 NOTRE DAME L. REV. 1279, 1281 (2007).

140. *See id.* at 1286–87.

141. It might also reveal that the nature of the responsiveness that appears in unpublished opinions is different from that in published opinions. As one commentator has articulated the justifications:

> A principal justification for unpublished rulings is that they take less time to prepare than do published opinions. An extensive opinion is said not to be needed if the law to be applied is straightforward or if a case is heavily fact-specific and thus is of minimal or narrower applicability. Because unpublished opinions are primarily directed to the parties rather than a larger audience, the statement of facts, which are known to the parties, can be truncated. Also, the law need not be elaborated, with only enough analysis provided to demonstrate to the parties that consideration has been given to the legal issues.

Stephen L. Wasby, *Unpublished Decisions in the Federal Courts of Appeals: Making the Decision to Publish*, 3 J. APP. PRAC. & PROCESS 325, 333–34 (2001).

Perhaps the most prominent example of the recent work on judicial quality is a series of articles by Stephen Choi and Mitu Gulati, in which they endeavor to evaluate the performance of federal appeals court judges.[142] Choi and Gulati based their evaluation on a combination of measures designed to assess productivity, quality, and independence, including, respectively, the number of opinions written by each judge, the frequency with which each judge's opinions are cited by other judges, and the extent to which each judge disagreed with her colleagues who were appointed by presidents from the same party.[143]

Choi and Gulati's work generated a significant response.[144] Although many commentators were generally positive about the idea of attempting to assess judicial quality empirically, most also offered up critiques of the methodology. Some of these critiques paralleled those directed at ideologically focused work—for example, that there are qualitative dimensions to judging that simply cannot be captured by quantitative measures.[145] Other critics emphasized what they perceived as incompleteness in the measures, whether because they regarded the specific phenomena that Choi and Gulati investigated as not sufficiently reflective of the underlying traits they attempted to measure,[146] or more generally on grounds that Choi and Gulati's set of underlying traits provided an incomplete picture of judicial quality.[147] Coupled with these assertions of incompleteness is the concern that quantitative measurement of judicial quality will skew judicial behavior, as judges work to maximize their performance along the measured dimensions, perhaps to the detriment of the less easily quantifiable aspects of effective judging.[148]

Finally, this line of research may yield insights that are useful to practicing lawyers, and to those who teach advocacy. One can imagine, for example, large-scale analysis of the relationships among briefs and

142. *See, e.g.*, Stephen J. Choi & G. Mitu Gulati, *Mr. Justice Posner? Unpacking the Statistics*, 61 N.Y.U. ANN. SURV. AM. L. 19 (2005); 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); Stephen Choi & Mitu Gulati, *A Tournament of Judges?*, 92 CALIF. L. REV. 299 (2004).

143. *See* Choi & Gulati, *A Tournament of Judges?*, *supra* note 142, at 305–10.

144. *See, e.g.*, Frank B. Cross & Stefanie Lindquist, *Judging the Judges*, 58 DUKE L.J. 1383, 1384–85 (2009). The work also served as the focal point for a symposium issue of the *Florida State University Law Review*. *See* Steven G. Gey & Jim Rossi, *Empirical Measures of Judicial Performance: An Introduction to the Symposium*, 32 FLA. ST. U. L. REV. 1001, 1002–03 (2005).

145. Gey & Rossi, *supra* note 144, at 1004.

146. *See* Cross & Lindquist, *supra* note 144, at 1388–93.

147. *See, e.g.*, Lawrence B. Solum, *A Tournament of Virtue*, 32 FLA. ST. U. L. REV. 1365, 1389, 1397–98 (2005) (criticizing specifically Choi and Gulati's technique for undermining the rule of law, excluding certain variables, and lacking transparency).

148. *See* Cross & Lindquist, *supra* note 144, at 1395–96.

opinions generating information about the relative utility of briefing practices and approaches. It may tell us something about whether reply briefs matter, or whether response briefs should place relatively greater emphasis on engaging with the opponent's arguments or developing their own. It could also facilitate quantitative assessment of lawyering skills, such as enabling assessment of the relative quality of public defenders and private counsel in criminal appeals, or comparisons of specialists and non-specialists.

III. AN INITIAL INVESTIGATION OF RESPONSIVENESS IN THE FIRST CIRCUIT

Despite the potential payoffs, the concept of judicial responsiveness remains understudied—especially so if one regards it as central to the entire endeavor of adjudication. One of us has in previous work explored various dimensions of courts' responsiveness obligations, ranging from an effort to define the contours of those obligations (with the failure to meet them constituting "judicial inactivism")[149] to the exploration of various ways in which judicial processes and structures create or fail to create incentives for courts to be responsive.[150] Underlying this work is an understanding that courts often fall short of Fuller's ideal, even when that ideal is moderated to take account of other legitimate considerations that might drive judicial behavior away from its fully responsive version. Yet, as is characteristic of much legal scholarship touching on the judicial process, that understanding is based largely on anecdotal evidence derived from personal experience, lore gathered from lawyers, and the occasional judicial admission that things occasionally get swept under the rug (always by other judges, of course).[151] To date, no one has rigorously investigated the extent to which courts and judges are responsive to the advocacy before them.

There are at least two reasons for this lack of developed evidence. One is that the necessary information has historically been difficult to obtain. Court opinions have been readily available at least since the rise of West Publishing, but only recently has it become easy to access electronic versions of the briefs submitted in a large range of cases. The second is that measuring judicial responsiveness, as is the case with

149. *See* Oldfather, *Defining Judicial Inactivism*, *supra* note 8, at 123.

150. *See* Oldfather, *Remedying Judicial Inactivism*, *supra* note 11, at 749–58.

151. *See, e.g.*, POSNER, FEDERAL COURTS, *supra* note 135, at 165 (noting that "the unpublished opinion provides a temptation for judges to shove difficult issues under the rug in cases where a one-liner would be too blatant an evasion of judicial duty"); Patricia M. Wald, *The Rhetoric of Results and the Results of Rhetoric: Judicial Writings*, 62 U. CHI. L. REV. 1371, 1374 (1995) ("I have seen judges purposely compromise on an unpublished decision incorporating an agreed-upon result in order to avoid a time-consuming public debate about what law controls.").

"thick" measures of judicial output more generally,[152] is, as noted above, both labor intensive and subject to concerns about coding reliability. This Part reports the methodology and results of our initial efforts to employ automated methods to assess responsiveness in a sample of cases from the First Circuit Court of Appeals.

A. *The Sample of Cases*

We analyzed a sample of thirty cases in which opinions were issued by the First Circuit Court of Appeals in 2004 (the specific cases are listed in the Appendix). The sample was selected from the total set of such cases decided by the First Circuit in 2004 via a two-step process. First, we identified the cases for which briefs from both parties were available on Westlaw.[153] That returned a list of ninety-seven cases.[154] Second, we selected every third case to analyze, except where the case that would otherwise be selected was inappropriate (such as, for example, where it involved third parties), in which case we moved to the next case and resumed the pattern of selecting every third case.[155] Of the thirty opinions in the sample, twenty-seven are "published" opinions, and twenty-one affirmed the lower court's ruling.[156] The briefing in fifteen of the cases included a reply brief; the other fifteen

152. It is a problem that pervades the assessment of judicial output more generally. Because it is so difficult to assess the quality of a judicial decision, we tend to place a lot of emphasis on process and on qualities of the judge such as impartiality. *See* Evans et al., *supra* note 7, at 1010; *see also* RICHARD A. POSNER, HOW JUDGES THINK 3 (2008).

153. More specifically, the subset is limited to cases in which there is a primary brief from each party and at most one reply brief. Excluded were cases in which there were amicus briefs, cases involving more than two parties, and cases in which more than one reply brief was filed. Once the methodology is perfected, these sorts of variations would make good independent variables in a sufficiently large study.

154. Our query to West resulted in a partially satisfactory explanation for how it is that the briefs for some cases but not for others are available:

> Some reasons include: (1) Access—some courts will not provide us with briefs to certain cases, for various reasons. For briefs that the courts have online, this is the primary reason why we do not have every brief. (2) Availability—some briefs (especially older briefs) are not available through the courts online. (3) Resources—for some briefs that are not available online, West would (and may still) send someone to the court to scan copies of briefs for later addition to Westlaw; in many courts it would be too time consuming to copy every brief they had on file.

That doesn't clarify much, but it's clear that the subset of this set of cases for which both briefs are available on Westlaw is a nonrandom sample. E-mail from Matthew Singewald, Academic Account Manger, West, a Thomson-Reuters Co. (June 2, 2009, 11:41 CST) (on file with author).

155. The result is that this, too, is a nonrandom sample.

156. We coded an opinion as an affirmance only when it affirmed the lower court's decision in all respects. All other results were coded as reversals.

did not. Because the sample of cases for which Westlaw makes briefs available is presumably nonrandom (as is our subset of that sample), we recognize that it is inappropriate to attempt to generalize from our findings to any conclusions about the responsiveness of the First Circuit across a broader range of cases. We do, however, contend that this sample of cases provides a basis for testing the validity of the measures of responsiveness that we proposed.

B. *Assessment One—Manual Coding*

The first stage of our analysis of our sample of cases involved manual content analysis and coding. That proceeded as follows. With respect to each of the thirty cases, we first assessed the arguments made by the parties. This involved drawing on the statement of issues, summary of argument, and argument sections of each brief, with the focus on identifying the thrust of the argument and the principal authorities upon which the parties relied. After doing this for both parties' briefs, the next step entailed a comparison of the issues, arguments, and authority presented by the parties in their briefs to the issues, arguments, and authority discussed by the court in its opinion. Although this assessment necessarily required the exercise of judgment, it involved some relatively concrete steps such as searching the opinions for specific words and phrases, as well as citations to authority that played a prominent role in the parties' briefs.

The next step was to categorize the opinion in terms of its responsiveness to the issues and arguments presented by the parties. We broke responsiveness down into three basic categories into which the court's analysis with respect to each issue in its opinion could be placed:[157]

> (i) *Strongly responsive*—A strongly responsive analysis addresses the issues on the parties' terms, relies almost exclusively on the universe of authority they present, and grapples with the arguments they make. A strongly responsive analysis thus proceeds from the same fundamental conception of the nature of the issue as is held by the parties and manifests itself in an opinion that the parties would regard as having fully addressed their proofs and arguments.
>
> An example of an opinion coded as strongly responsive is *Redondo Construction Corp. v. Puerto Rico Highway &*

157. The theoretical justification for these categories can be found in Oldfather, *Defining Judicial Inactivism*, *supra* note 8, at 164, 168–75 & n.202.

Transportation Authority.[158] The defendant Authority appealed the district court's denial of its claim of Eleventh Amendment immunity as an arm of the state.[159] The parties regarded the case as being governed by the court's decision in a prior case,[160] and the court accepted that conception of the dispute and engaged in an analysis that falls within the parameters created by the parties' arguments and positions.[161] In all, the process maps out fairly well onto idealized notions of what the appellate process should look like.

A case that is less a classic example of the appellate process but that is still coded as being strongly responsive is *Sullivan v. Neiman Marcus Group, Inc.*,[162] which was an appeal from a grant of summary judgment in favor of an employer on a claim that plaintiff was fired for having a disability.[163] The trial court based the grant of summary judgment on its conclusion that there was no evidence based on which a reasonable fact finder could conclude that the plaintiff was fired for having a disability, rather than based on the defendant's rational belief that the plaintiff possessed alcohol and was intoxicated on the job.[164] The appellant contested that ruling on the trial court's terms.[165] The appellee addressed that argument, and also offered an alternative ground for affirming the trial court.[166] The First Circuit based its affirmance on the alternative ground.[167] Despite the fact that the appellant did not file a reply brief, and consequently did not address the alternative ground, the opinion was coded as strongly responsive because the court did not depart from the framework put before it by the litigants.

(ii) *Weakly responsive*—In a weakly responsive analysis, the court addresses the parties' arguments, but offers a justification for its decision that departs in some

158. 357 F.3d 124 (1st Cir. 2004).

159. *Id.* at 125.

160. *Id.* at 126 (citing Fresnius Med. Care Cardiovascular Res., Inc. v. Puerto Rico & the Caribbean Cardiovascular Ctr. Corp., 322 F.3d 56 (1st Cir. 2003)).

161. *See id.* at 126–28.

162. 358 F.3d 110 (1st Cir. 2004).

163. *Id.* at 114.

164. *Id.*

165. *Id.*

166. *Id.*

167. *Id.*

meaningful way from strong responsiveness. Weak responsiveness can manifest itself in differing forms, including the following:

> (a) The court might *recast* the proofs and arguments of the parties, as by, for example, reaching a conclusion about the significance of a precedent or authority that departs from those proposed by the parties. For example, in *Quaak v. Klynveld Peat Marwick Goerdeler Bedrijfsrevisoren*,[168] the court, in considering the standards to be employed in considering whether to grant an injunction concerning an international proceeding, read a pair of cases discussed by the parties to require a "gatekeeping inquiry" that neither party had identified.[169] In this situation the analysis is strongly responsive except insofar as the court uses the authorities provided by the parties to step outside the parameters of the argument that the parties have set.
>
> (b) The court might rely on *additional* authority in reaching its decision, apart from the authorities that the parties identify as governing the case. Here again, the court's analysis might generally be strongly responsive, but for its resort to some authority not identified by either of the parties in support of a material portion of its analysis. For example, in *Correia v. Hall*,[170] the court considered, among other claims, a habeas petitioner's argument that the trial judge's comments during trial suggested pique at the petitioner's decision to demand a jury trial, and rejected the argument based on authority that neither party had cited.[171]
>
> (c) The court might rely on *alternative* authority, addressing the issue on the same general terms that the parties have identified, but concluding that its resolution is governed by an authority that neither party has identified. An example here is *Gulf Coast Bank & Trust Co. v. Reder*,[172] in which the defendant-appellant argued

168. 361 F.3d 11 (1st Cir. 2004).

169. *Id.* at 18; *see also* Brief for Appellant Quaak v. Klynveld Peat Marwick Goerdeler Bedrijfsrevisoren, 361 F.3d 11 (1st Cir. 2004) (No. 03-2704); Brief for Appellee Quaak v. Klynveld Peat Marwick Goerdeler Bedrijfsrevisoren, 361 F.3d 11 (1st Cir. 2004) (No. 03-2704).

170. 364 F.3d 385 (1st Cir. 2004).

171. *Id.* at 391–92; *see also* Brief for Appellant Correia v. Hall, 364 F.3d 385 (1st Cir. 2004); Brief for Appellee Correia v. Hall, 364 F.3d 385 (1st Cir. 2004) (No. 03-1203).

172. 355 F.3d 35 (1st Cir. 2004).

> that the trial court erred in granting the plaintiff's motion for judgment on the pleadings under Rule 12(c) of the Federal Rules of Civil Procedure.[173] The parties argued over whether the standard to be applied to a Rule 12(c) motion is the same as that applied to a Rule 12(b)(6) motion.[174] The court, in concluding that the Rule 56 summary judgment standard applied, relied on a case that neither party cited.[175] Here, although the court stayed within the broad parameters of the issue on which the parties focused, its invocation of alternative authority resulted in an analysis that tends more toward a nonresponsive, sua sponte resolution than toward strong responsiveness.

One should note that the three subcategories identified above can be regarded as involving increasingly greater departures from strong responsiveness. A court that recasts the authority relied upon by the parties remains within the contours of the dispute as the parties conceive of it, while a court that relies on additional authority has stepped outside of that framework. A court that relies on alternative authority has, in turn, taken an additional step outside the parameters set by the parties. The court in each instance resolves what can still be characterized as the same issue, but in a way that departs in an increasingly significant way from the framework within which the parties have presented the case.

(iii) *Nonresponsive*—A court's analysis can be nonresponsive in two primary ways. First, it could resolve the case based on issues and authorities not presented by the parties, as by raising sua sponte what the court concludes is a dispositive issue. Second, the court could simply fail to address an issue. The only case in our sample that was coded as entirely nonresponsive was *Olick v. John Hancock Mutual Life Insurance Co.*,[176] a two-paragraph unpublished, per curiam opinion that characterizes the

173. *Id.* at 37.

174. *See* Brief for Appellant at 10 Gulf Coast Bank & Trust Co. v. Reder, 355 F.3d 35 (1st Cir. 2004) (No. 03-1963); Brief for Appellee at 4 Gulf Coast Bank & Trust Co. v. Reder, 355 F.3d 35 (1st Cir. 2004) (No. 03-1963).

175. *Gulf Coast*, 355 F.3d at 38; *see also* Brief for Appellant Gulf Coast Bank & Trust Co. v. Reder, 355 F.3d 35 (1st Cir. 2004) (No. 03-1963); Brief for Appellee Gulf Coast Bank & Trust Co. v. Reder, 355 F.3d 35 (1st Cir. 2004) (No. 03-1963).

176. 106 F. App'x 736 (1st Cir. 2004).

> appellant's motions as partly moot (a contention that does not appear in the appellee's brief),[177] and to the extent not moot, then "misplaced."[178] While a charitable interpretation of the opinion would be that it provides evidence of the court's engagement with the parties' arguments, that interpretation requires drawing substantial inferences about the court's reasoning process, and the opinion itself does not reflect the sort of responsiveness envisioned by Fuller.

As the discussion suggests, the concept of responsiveness as it manifests itself in any given opinion tends to be nuanced and multifaceted. Particularly within the category of weak responsiveness it is not unusual to see an analysis of a single issue in which the court departs from strong responsiveness in more than one way. There is, without question, considerable reductionism involved in placing the analysis of a specific issue within a single category, and the categories themselves suggest brighter lines than reality provides. To the extent that it is even appropriate to consider responsiveness as merely a one-dimensional concept, it undoubtedly ought to be regarded as continuous rather than a discrete variable. But efforts to code it in that fashion would only magnify the concerns about reliability discussed above.

Finally, we also coded for the extent to which the court provided elaboration of its analysis underlying the resolution of each issue it decided. Conceptually, although it is not simply a measure of length, elaboration is meant to capture an aspect of the opinions that is more quantitative than responsiveness.[179] But there is undoubtedly some overlap. We placed the court's analysis of each issue into one of four categories: (1) *full elaboration*, which tends toward an idealized form of appellate decision making, in which the court provides a factual background and relatively detailed explanation of its analysis, including reference to and further analysis of applicable authorities; (2) *mixed elaboration*, in which the court departs in a material way from full elaboration, yet provides more than a summary disposition of an issue coupled with a citation to authority; (3) *minimal elaboration*, in which the court provides at most a cursory discussion of the issue, coupled with a citation to authority and a conclusory assertion that the authority resolves the issue; and (4) *no elaboration*, in which the court either fails to speak to the issue at all or simply asserts its resolution with no citation to authority. Though perhaps to a lesser extent than is true of

177. *See* Brief for Appellee Olick v. John Hancock Mut. Life Ins. Co., 106 F. App'x 736 (1st Cir. 2004) (No. 03-2350).

178. *Olick*, 106 F. App'x at 738.

179. For a more complete definition, see Oldfather, *Defining Judicial Inactivism*, *supra* note 8, at 164, 175–80.

the responsiveness determination, characterizing the nature of a court's elaboration with respect to its elaboration of a given issue likewise involves the exercise of some judgment.

C. *Assessments Two and Three—Automated Content Analysis and Coding*

We investigated two types of automated approaches for quantification of responsiveness. These methods differ in the types of evidence considered. One approach uses the textual content of a case's opinion and briefs. This method estimates responsiveness by the *cosine similarity* between opinion and brief documents. This widely used document-similarity measure has been successfully applied to document classification, information retrieval, and other natural language processing tasks.[180] The second approach is based on citation patterns in the opinion and briefs. Both methods involve measuring various aspects of the overlap among the documents.

We preprocessed all brief and opinion hypertext documents in our corpus in the same way. We first extracted the text and citations. A document's citations are the external addresses referenced in *anchor* tags, which appear in browsers as clickable links. Because Westlaw provides links only for other materials available within Westlaw, a small percentage of citations will not be extracted. Since for typographic reasons the Westlaw encoding often has multiple anchor tags for the same citation instance, we disregarded citation occurrence quantities. That is, for purposes of subsequent processing we only recorded which citations are present and absent in a given document. A document's text can be thought of as the words that are visible when the hypertext document is viewed in a web browser. We eliminated typographic markup—for example, font size and italics—and tokenized the document into a word sequence. We defined a word to be a contiguous string of alphanumeric (a–z, A–Z, 0–9) or underscore (_) characters. We converted all words to lowercase and stem words to their roots using Porter's method.[181]

After all documents were processed as described, we removed common or so-called "stop" words, which we assume are uninformative. We define a stop word to be any stemmed word present in at least one brief or opinion of all thirty cases. The remaining stemmed non-stop words, which we refer to as "terms," make up the corpus vocabulary. The vocabularies of the "argument only" and

180. *See, e.g.*, RADA MIHALCEA ET AL., AM. ASS'N FOR ARTIFICIAL INTELLIGENCE, CORPUS-BASED AND KNOWLEDGE-BASED MEASURES OF TEXT SEMANTIC SIMILARITY (2006), *available at* http://www.cse.unt.edu/~rada/papers/mihalcea.aaai06.pdf.

181. *See* M.F. Porter, *An Algorithm for Suffix Stripping*, 14 PROGRAM 130 (1980).

"argument + facts" corpora (described below) comprise 9,442 and 10,422 terms, respectively.

To compute similarity we first represented each document by its "term frequency, inverse-document frequency" (TF-IDF) vector, a common and useful representation for text processing tasks.[182] TF-IDF represents a document as a vector whose length is equal to the number of terms in the vocabulary.[183] Thus, each term *t* is associated with one dimension in the TF-IDF vector. Let x_A be the TF-IDF vector of document *A.* Then, element $x_A(t)$, which denotes the importance of term *t* to document *A,* is the product of a term-frequency factor and an inverse-document frequency factor:

$$x_A(t) = tf(A,t) \times idf(t)$$

$$tf(A,t) = \frac{\text{Number of occurences of } t \text{ in } A}{\text{Total number of total terms in } A}$$

$$idf(t) = \log\left(\frac{\text{Total number of documents in corpus}}{\text{Number of documents that contain } t}\right)$$

The cosine similarity between documents *A* and *B* is the cosine of the angle between their TF-IDF vectors x_A and x_B.

To compute similarity based on citation patterns, we generated a score indicating the percentage of authorities that were cited in the opinion that were also cited in the brief or set of briefs noted, which we designate as "*% Responsive*." We also generated a score indicating the percentage of authorities cited in the brief or set of briefs in question that were also cited in the opinion, which we designate as "*% Responded*."

We applied these techniques to the opinions and three versions of the briefs. The three versions of the briefs were: (1) the entire document, including tables of authorities and contents, as well as West-generated front and back matter; (2) versions from which the tables of contents and authorities, as well as any sections pertaining to the court's jurisdiction (which was not in controversy in any of the cases), were removed, as was all West-generated front and back matter; and (3) versions including only the standard of review and argument sections (including any sections designated as a summary of argument). The

182. GERARD SALTON & MICHAEL J. MCGILL, INTRODUCTION TO MODERN INFORMATION RETRIEVAL 30, 63 (1983). For a general overview of stemming, TF-IDF, and other natural language processing issues, see generally CHRISTOPHER D. MANNING ET AL., AN INTRODUCTION TO INFORMATION RETRIEVAL (2009).

183. *See* MANNING ET AL., *supra* note 182, at 119.

primary difference between the second and third versions was the removal of the statement of facts.

D. *Results and Analysis*

1. Manual Coding

Our primary purpose in manually coding the documents was to generate a baseline by which to assess the validity of our automated assessments. Even so, the results are interesting in their own right. Because the cases varied in terms of the number of issues presented, we coded each issue presented individually, and assigned responsiveness scores to cases by averaging the scores across all the issues. The court considered sixty-two issues over the thirty cases. Broken down in terms of the overall responsiveness of the analysis, the distribution of how the issues were considered is presented in Table 1.

Table 1: Breakdown of Responsiveness—All Issues

Opinion Type	Strongly Responsive	Weakly Responsive	Nonresponsive
Published	16	33	7
Unpublished	1	1	4

Considering only the thirty-four issues as to which the court's analysis was weakly responsive to the parties' contentions, the types of weak responsiveness in which the court engaged is displayed in Table 2 (note that the total is greater than thirty-four because for some issues the analysis fell into two categories).

Table 2: Subcategories of Weak Responsiveness

Alternative Authority	Additional Authority	Recasted Authority
23	8	7

We used two alternative methods for scoring cases for responsiveness. In the first, which we will call *categorical responsiveness* (CR), we assigned a score of 1, 2, or 3 for each issue as to which the court's analysis was coded as, respectively, nonresponsive, weakly responsive, and strongly responsive. Summary statistics for categorical responsiveness appear in Table 3.

Table 3: Summary Statistics for Categorical Responsiveness (CR)

	Percentiles	Smallest		
1%	1	1		
5%	1.4	1.4		
10%	1.5	1.5	Obs	30
25%	2	1.5	Sum of Wgt.	30
50%	2		Mean	2.257333
		Largest	Std. Dev.	.5830002
75%	3	3		
90%	3	3	Variance	.3398892
95%	3	3	Skewness	-.1002565
99%	3	3	Kurtosis	2.054464

Our second method, which we will call *incremental responsiveness* (IR), involved assigning different values to the different forms of weak responsiveness, as discussed above. This entailed use of a five-point scale, with values of 1 through 5 assigned to nonresponsiveness and strong responsiveness, respectively. Within the category of weak responsiveness, we assigned the value 2 to issues where the court relied on alternative authority, 3 to issues for which the court relied on authority in addition to that offered by the parties, and 4 to issues with respect to which the court relied on the authority provided by the parties but recast that authority. The underlying assumption, as discussed above, is that the ordering of this coding reflects in a rough way the relative extent of their departures from strong responsiveness. Summary statistics for incremental responsiveness appear in Table 4.

Table 4: Summary Statistics for Incremental Responsiveness (IR)

	Percentiles	Smallest		
1%	1	1		
5%	1.4	1.4		
10%	1.75	1.5	Obs	30
25%	2	2	Sum of Wgt.	30
50%	3		Mean	3.257667
		Largest	Std. Dev.	1.315847
75%	5	5		
90%	5	5	Variance	1.731453
95%	5	5	Skewness	0.1100383
99%	5	5	Kurtosis	1.66367

We also assigned scores based on how cases were coded for elaboration, scoring 4 for full elaboration, and 3, 2, and 1, respectively, for mixed, minimal, and no elaboration. Summary statistics for elaboration appear in Table 5. As was the case with responsiveness, elaboration scores for cases involving more than one issue were generated by averaging the scores for all issues.

Table 5: Summary Statistics for Elaboration

	Percentiles	Smallest		
1%	1	1		
5%	1.25	1.25		
10%	1.4	1.4	Obs	30
25%	2	1.4	Sum of Wgt.	30
50%	2.5		Mean	2.401667
		Largest	Std. Dev.	.7057013
75%	3	3		
90%	3	3	Variance	.4980144
95%	3	3	Skewness	-0.094912
99%	4	4	Kurtosis	2.386428

2. Document Similarity

As noted above, we applied our automated coding methodologies to three versions of the briefs. The averages, ranges, and standard deviations of the document similarity scores are depicted in Table 6.

Table 6: Averages, Ranges, and Standard Deviations of Document Similarity Scores

Version of Briefs	Appellant Briefs to Opinion	Appellee Briefs to Opinion	Reply Briefs to Opinion	Appellant + Reply Brief to Opinion	All Briefs to Opinion	Appellant + Reply Briefs to Appellee Briefs
Entire Document	68.6 27.1 – 86.6 StD: 11.8	70.4 50.3 – 90.6 StD: 10.5	56.3 17.9 – 77.9 StD: 14.2	67.1 23.5 – 84.8 StD: 14.9	73.1 34.5 – 90.5 StD: 11.4	76.7 49 – 90.2 StD: 10.5
Facts + Argument	68.6 28.4 – 86.3 StD: 12.1	70.1 42.4 – 90.6 StD: 11.0	58.7 17.9 – 77.5 StD: 14.7	67.9 24.3 – 85 StD: 15.2	73.5 35.9 – 88.7 StD: 11.1	74.5 38.9 – 90.2 StD: 12.0
Argument Only	70.0 37.0 – 88.6 StD: 11.1	70.6 39.7 – 89.4 StD: 11.1	60.9 30.9 – 77.6 StD: 12.6	69.0 34.6 – 86.5 StD: 13.6	75.6 46.5 – 88.5 StD: 9.9	71.5 35.3 – 90.2 StD: 12.3

These numbers demonstrate that excising portions of the briefs had surprisingly little effect on document similarity scores. Even so, we have chosen to conduct our analysis using the similarity scores generated using the briefs edited to include the arguments only. We base this decision on our intuition that the argument sections of the briefs will contain the key components of the parties' arguments (including references to those facts that they deem significant), coupled with the suggestion in the data implying that set of similarity scores captures both greater responsiveness and a greater disjunction between the two sides' arguments.

3. Citation Analysis

Working with the versions of the briefs referenced above in which all portions other than argument sections were excised, we assessed the relationship between the briefs and the opinions in terms of authorities upon which both relied. For both the "%Responsive" and "%Responded" measures discussed above, Table 7 shows the averages, ranges, and standard deviations of these scores.

Table 7. Averages, Ranges, and Standard Deviations of Citation Analysis Similarity Scores

Appellant Brief– %Responsive	Appellant Brief– %Responded	Reply Brief– %Responsive	Reply Brief– %Responded	Appellant + Reply– %Responsive	Appellant + Reply– %Responded
18.2 0.0 – 53.3 StD: 13.2	22.9 0.0 – 78.6 StD: 21.6	16.4 0.0 – 57.1 StD: 18.4	24.6 0.0 – 66.7 StD: 27.0	23.6 0.0 – 57.1 StD: 15.4	20.0 0.0 – 50.0 StD: 16.7
Appellee Brief– %Responsive	Appellee Brief– %Responded	All Briefs– %Responsive	All Briefs– %Responded		
26.6 0.0 – 68.4 StD: 16.8	20.4 0.0 – 59.1 StD: 14.2	35.0 0.0 – 68.4 StD: 16.8	16.3 0.0 – 41.7 StD: 11.1		

4. Analysis

a. The Viability of Automated Assessments of Responsiveness

Our primary goal in this Part is to explore the general question of whether automated content analysis can effectively substitute for human assessment of the relationship among briefs and opinions and, more specifically, the viability of our two automated methodologies as approaches to the study of judicial responsiveness. The results are encouraging. Although room for refinement remains, our investigation demonstrates that relatively basic methods of analyzing document similarity can provide insight, across a run of cases, into whether a court consistently and deeply engages with cases on the terms in which the parties have argued them. In addition, some of our specific findings are interesting in their own right, and suggest avenues for further study.

One of the virtues of an automated approach is that it maximizes reliability.[184] Our task, then, is to establish that document similarity scores and citation analysis similarity scores serve as valid measures of responsiveness.[185] In this regard there is, we believe, considerable intuitive appeal to both measures. Documents grappling with the same proofs and arguments seem likely to use the same words, such that one would expect considerable overlap among the briefs and opinion in a case where the parties and the court approach the same issue in fundamentally the same way. In cases where the parties and the court have differing ideas about what is at stake, in contrast, one would

184. "Reliability is concerned with questions of stability and consistency—does the same measurement tool yield stable and consistent results when repeated over time." *QMSS e-Lessons: Validity and Reliability*, COLUMBIA CTR. FOR NEW MEDIA TEACHING AND LEARNING, http://ccnmtl.columbia.edu/projects/qmss/measurement/validity_and_reliability.html (last visited Feb. 13, 2012).

185. "Validity refers to the extent we are measuring what we hope to measure (and what we think we are measuring)." *Id.*

expect to find greater divergence in word usage. Thus, it seems probable that, to use something of an extreme example, a case in which the court raises sua sponte an issue that it determines is dispositive of the entire case will be a case in which there is relatively little overlap between the words used in the briefs and the words used in the opinions. In short, a responsive opinion would seem likely to have a higher textual similarity score when compared to the briefs than would a nonresponsive opinion. And while there will undoubtedly be individual instances in which that is not the case—because, for example, a court disposes of the entire case based on its resolution of a single issue and thereby renders the remaining issues moot—these expectations seem reasonable when applied to opinions in the aggregate.

In similar fashion, it seems likely that legal documents engaged with the same doctrines and arguments will refer to the same authorities. Indeed, since authorities are at the very core of the sorts of "proofs and arguments" that Fuller referred to,[186] and thus are at the very core of the responsiveness that he placed at the center of adjudicative legitimacy, a court's resort to the same authorities as relied upon by the parties seems almost necessarily to be coextensive with a responsive analysis. Put in terms of the simple example we used above, if the parties and the court all conceive of the case as governed by *Smith v. Jones*, then one would expect the briefs and opinions to refer to *Smith v. Jones* and other cases and materials concerning the scope and proper application of *Smith v. Jones*.[187]

Our manual coding of the thirty cases in our sample allows us to assess these measures of responsiveness based on more than mere intuition. Using Stata,[188] we calculated the pairwise correlation (Pearson) of all variables. Since we compute *p*-values for a number of different correlations, we must be mindful of multiple hypothesis testing issues when testing for statistical significance. We assessed statistical significance of observed *p*-values in the context of multiple hypotheses using quantile–quantile (Q–Q) plots. Figure 8 shows the Q–Q plot of *p*-values from correlations of responsiveness scores of the "argument only" documents with the manually coded responsiveness scores. We have four types of briefs (appellant, appellee, appellant+reply and appellant+reply+appellee), and three kinds of responsiveness measures (%Responsive, %Responded, and cosine similarity), which provide a total of twelve measures of responsiveness. For each of these twelve measures we computed correlations and *p*-values with our two manual responsiveness codes (MR and IR), yielding the twenty-four *p*-values

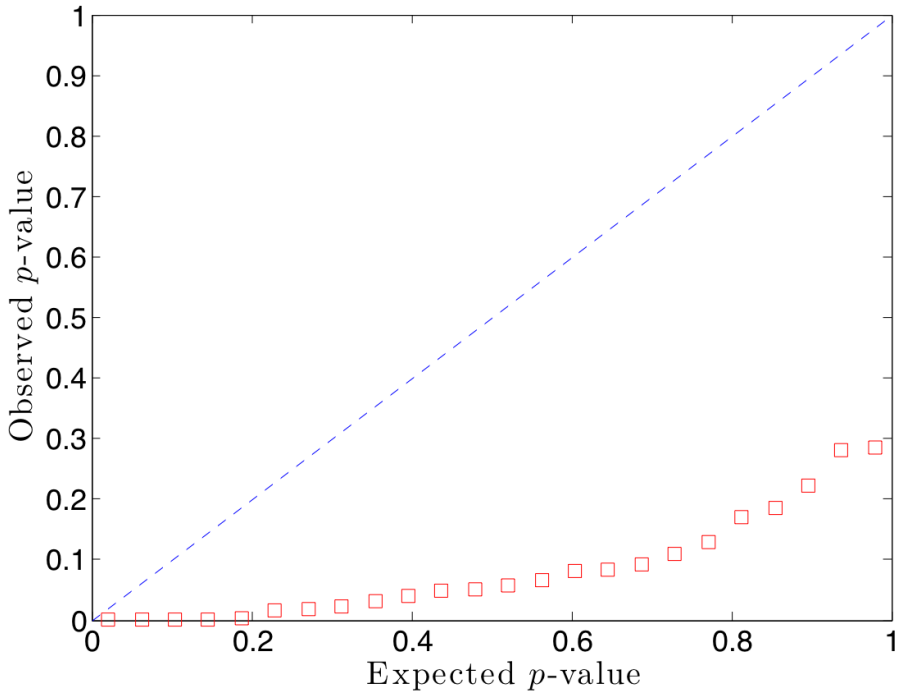
186. Fuller, *supra* note 8, at 367.

187. *See supra* notes 9–12 and accompanying text.

188. Stata is a popular statistical software package. *See* STATA, http://stata.com (last visited June 4, 2012).

plotted in the figure. Under the null hypothesis of no correlation we would expect the observed *p*-values to lie close to the diagonal line from (0,0) to (1,1). As eleven of the twenty-four *p*-values are < 0.05 and five of the twenty-four are < 0.01, we clearly observe substantial deviation from the diagonal, and thus conclude that the deviations of the observed correlations from 0.0 are statistically significant.

Figure 8: Correlations with Manually Coded Responsiveness Scores



Among the correlations with a *p*-value of .05 or less are those for the relationships set forth in Table 9.

Table 9: Correlations with *p*-value < .05

Variable 1	Variable 2	Correlation Coefficient	*p*-value
Incremental Responsiveness	Appellee Briefs/Opinion Textual Similarity	0.44	0.0150
Incremental Responsiveness	Appellant+Reply Briefs/Appellee Briefs Textual Similarity	0.37	0.0414
Categorical Responsiveness[189]	All Briefs–%Responsive Citation Similarity	0.63	0.0002
Incremental Responsiveness[190]	All Briefs–%Responsive Citation Similarity	0.69	<10e-4
Incremental Responsiveness	All Briefs–%Responded Citation Similarity	0.40	0.0287
Categorical Responsiveness	Elaboration	0.69	0.0002
Incremental Responsiveness	Elaboration	0.67	<10e-4
All Briefs–%Responsive	All Briefs/Opinion Textual Similarity	0.47	0.0091

Several conclusions follow from these numbers. Beginning with the positive, the results suggest the validity of our citation analysis as a measure of responsiveness. The strongest correlation in the table is between the percentage of authorities cited in an opinion that were also cited in a brief (All Briefs–%Responsive) and our Incremental Responsiveness (IR) assessment. That same measure of authorities is also correlated, though not quite as strongly, with our Categorical Responsiveness measure. IR is also correlated with the percentage of authorities cited in any brief that are also cited in the opinion (All

189. Categorical Responsiveness was also correlated in a significant way with Citation Analysis–%Responsive (Appellant + Reply) (0.3774/0.0398), and Citation Analysis–%Responsive (Appellee) (0.4273/0.0185).

190. Incremental Responsiveness was also correlated in a significant way with Citation Similarity–%Responsive (Appellant + Reply) (0.4159/0.0223), and Citation Similarity–%Responsive (Appellee) (0.5153/0.0036).

Briefs–%Responded). Although, as we discuss below, the absolute percentages involved in these measures seem strikingly low, the analysis confirms our intuition that a more responsive analysis will tend to involve greater reference to the authorities cited by the parties. In short, the results support the conclusion that citation analysis is a valid measure of judicial responsiveness.

While the results show that our measure of textual similarity has promise, they also suggest that further refinement is in order. Two textual analysis scores—for the similarities between appellee briefs and the opinions, and for the similarity among the appellant-side and appellee-side briefs—were correlated with our IR measure. Both results comport with our intuitions: the former because one would anticipate similarities between briefs and a responsive opinion, and the latter because a dispute in which the parties' arguments are more tightly bound to one another seems more likely to be one in which a court will address their arguments in a responsive manner. Perhaps the most promising result in this regard is the final one displayed in Table 9, which is the correlation between All Briefs–%Responsive and All Briefs/Opinion–Textual Similarity. What this suggests is that there is greater textual similarity among the briefs and opinions in cases in which the parties and court base their analyses on the same authority. Beyond that, however, none of the document similarity scores produced a significant correlation with either of our manually generated measures of responsiveness.

These are, of course, complex relationships, and our data suggest certain concerns of which to be mindful as we develop this line of research. For example, the extent to which a court can be responsive in a way that will register as such using these measures will sometimes turn on whether the parties have a common conception of their dispute. If the parties disagree about the nature of the issues in a case, a court could easily issue an opinion that, while responsive, would score low on both of our measures. For example, if one party offers an argument that the court accepts, and that, when resolved, renders the remainder of the issues before the court moot, then the opinion will likely fail to register as responsive under either of our automated methodologies even though the court may have resolved the case in an entirely appropriate way. In similar fashion, the strong correlation between our coding for elaboration and both of our manually coded responsiveness measures might give us pause. Although our measurement of elaboration was intended to capture more than mere length of opinion, the correlation suggested the need to ensure that our measures were not testing largely for length. We found no significant correlation between length of opinion (measured with stop words excluded) and either IR or All Briefs–%Responsive. This is consistent with the possibility that our

measures of elaboration and responsiveness were ultimately assessing different aspects of the same underlying phenomena. There was, however, a significant correlation between opinion length and our textual analysis document similarity scores for all briefs compared to the opinions.[191]

b. Suggestions from the Results of Our Sample

Although they are not the focus of our study, the results of our automated coding as applied to our sample of cases are intriguing in their own right and suggest topics for future investigation. Considering first the textual similarity scores, the range of scores (presented in Table 6) strikes us as quite large. With a large enough set of cases, it might be possible to uncover relationships between, for example, the outcomes in cases and their responsiveness to appellants or appellees. (We found no significant correlation between results and any of our measures of responsiveness within our thirty cases.) Courts and judges could also be compared in terms of the extent to which their opinions are similar to the briefs. Further refinements could account for variances based on subject matter or litigant characteristics. The development of appropriate baselines—beginning with average similarity scores for a large set of unrelated document pairs—would enable normative judgments about performance.

Another potentially noteworthy finding concerns the difference in scores between principal and reply briefs, with the latter scoring substantially lower regardless of the form of the briefs analyzed. This is true even when we consider only the subset of cases involving reply briefs. In those cases, the average textual similarity score for the comparison between reply briefs and opinions is 60.9, while it is 68.2 and 71.4 for appellant and appellee briefs, respectively. It is difficult to know what to make of this, particularly given the correlation between document length and similarity scores, but the results suggest that further inquiry might reveal interesting information concerning the nature and utility of reply briefs.

The results of our citation analysis are interesting because they show that, within this sample of cases, the First Circuit did not typically restrict itself to the universe of authorities cited by the parties, and indeed, failed to refer to the bulk of the authorities mentioned in the briefs. On average, only 35% of the authorities cited in the court's opinions were among those cited by the parties, and the court cited just over 16% of the authorities referenced in the briefs. These numbers undoubtedly understate responsiveness to some degree because our methodology does not give greater weight to authorities to which the

191. More specifically, the correlation coefficient is 0.39, with a *p*-value of 0.03.

parties make multiple references. Even so, it is striking how little overlap there is between the parties' citations and the court's.[192] This suggests that judges have, and exercise, a considerable amount of discretion in choosing which precedent to follow; this is, at the very least, consistent with attitudinal descriptions of judicial behavior.[193] Here, too, the measure offers an intriguing avenue for exploring the relative performance of courts and judges, not only broken down in terms of the factors we identify above, but also to account for the nature of the authorities to which references were made.

IV. NEXT STEPS AND CONCLUSION

Efforts at using computational techniques to analyze judicial opinions and other legal documents remain in their early stages, and past efforts have yielded mixed results.[194] Even so, automated content analysis will undoubtedly play a greater role in legal scholarship in coming decades. The combination of greater availability of information in electronic formats coupled with increased sophistication of computational techniques should enable increasingly varied and sophisticated investigations. Our aim in this Article has been simply to establish the value of the methodology, and to demonstrate that resort to basic methods of automated content analysis can provide useful information about the relationship among the briefs and opinions in a case. We conclude by outlining potential refinements to our methodology, as well as potential avenues for the use of automated content analysis more generally.

With respect to our study, while our measures show promise, they also suggest the possibility of superior computational approaches to quantifying responsiveness. One foreseeably productive approach is to

192. Indeed, the First Circuit in our sample cited an even lower portion of the cases referenced in the briefs than prior research comparing briefs and opinions in the U.S. Supreme Court found. *See* William H. Manz, *Citations in Supreme Court Opinions and Briefs: A Comparative Study*, 94 LAW LIBR. J. 267, 294 (2002) (finding that "[s]lightly more than 25% of Supreme Court decisions cited in the briefs also appeared in the opinions" and "[a]pproximately 25% of the Court's case citations did not appear in any of the briefs").

193. *See* Frank B. Cross et al., *Citations in the U.S. Supreme Court: An Empirical Study of their Use and Significance*, 2010 U. ILL. L. REV. 489, 527–28; Frank B. Cross, *Chief Justice Roberts and Precedent: A Preliminary Study*, 86 N.C. L. REV. 1251, 1277 (2008) (noting "the value of examining the briefs of the parties as a cue for evaluating the Court's citation practices"). As Justice Cardozo noted over ninety years ago, "in a system so highly developed as our own, precedents have so covered the ground that they fix the point of departure from which the labor of the judge begins." BENJAMIN N. CARDOZO, THE NATURE OF THE JUDICIAL PROCESS 19–20 (1921).

194. *See* Choi & Gulati, *Which Judges Write Their Opinions*, *supra* note 7, at 1121 (finding that computational techniques somewhat correlates with authorship); Evans et al., *supra* note 7, at 1036 (finding that computational techniques "hold[] great promise" for future research).

develop scores that are a function of both citations and text. Since text and citations are disjoint sources of evidence, and as both measures correlate positively with our manually annotated responsiveness scores, it is probable that appropriately combined scores would correlate more strongly with our manual annotations than do scores based on a single type of evidence. The key, of course, is to intelligently combine text and citation evidence. Supervised learning methods, such as multiple regression and support vector machines, are designed to learn functions such as this. These approaches induce a mathematical function mapping inputs (text and citation attributes) to outputs (responsiveness) from a training set of input/output pairs. An advantage of the learning approach is that since the mapping function is learned automatically, more (and more complex) attributes can be readily incorporated with little overhead. If, for example, we wished to distinguish among different types of cited authorities (state law, United States Code, previous case, etc.), we could create attributes for each citation type, and use supervised algorithms to learn how to synthesize a case's attribute values into responsiveness scores. Other attributes that can be considered under the supervised methodology include term-specific weights, for example, to identify specific words indicative of responsiveness, and term patterns within the context of sentences, paragraphs, sections, and other higher-level document elements.

More broadly, we believe that automated content analysis holds out the promise of expanding the scope of topics for research. The ability to compare large numbers of briefs and opinions can facilitate the exploration of not only the behavior of different actors in the system[195]—various types of courts and judges and litigants—but also, as the capacity for digitizing archival material improves, changes in those actors' behavior over time.

---

195. Michael Evans and his co-authors outline the following possibilities:

> If the textual inputs and outputs [of the various actors in the legal system] can be reliably and meaningfully quantified, then a variety of innovative research questions can be addressed. What explains the ideological positions of the briefs submitted by litigants to a case? Are they influenced by positions taken by today's median justice in his or her opinion in a previous case? How do litigants' positions compare to the positions taken by amicus curiae? Do different types of interest groups submit more or less ideologically extreme amicus curiae briefs? How do repeat players' positions vary over time? Under what conditions (e.g., case salience, coalition size, type of opinion, position/clarity of relevant precedent) do justices articulate extreme or moderate positions? Do lower court opinions exhibit ideological shifts in response to change in Supreme Court policy? Can litigant success be explained by the positions taken in their briefs?

Evans et al., *supra* note 7, at 1020–21.

These efforts promise to lead us to a greatly enhanced understanding of the workings of the judiciary and the legal system more generally. A related hope underlying much of this work is that a relatively well-developed understanding might support efforts at prediction. That is, we might expect that textual analysis—perhaps supplemented by information external to the text, such as judicial ideology or the identity of counsel—will develop to the point where we have such a refined ability to account for the factors presented by a new case that we can predict to a great degree of accuracy how it will be resolved. Recent efforts at prediction have attained a relatively high degree of predictive accuracy within limited domains,[196] but much work remains.[197]

Part of that work—and it is work that might ameliorate somewhat the perceived gap between what legal academics do and what might benefit legal practitioners—can be done through automated content analysis. It will remain, at least until computers are able to "read" and comprehend text, an imperfect mode of inquiry, another tool in the scholar's toolbox, rather than a replacement for what has come before it.[198] We believe that inquiries of the sort we engage in above will, particularly when expanded to take into account the full spectrum of

196. For example, Professor Kevin Ashley and his colleague Stefanie Brüninghaus compared several computerized prediction methods applied to the same set of 184 trade secret misappropriation cases drawn from both federal and state courts over a several-decade period. Kevin D. Ashley & Stefanie Brüninghaus, *Computer Models for Legal Prediction*, 46 JURIMETRICS 309, 333, 337 (2006). Their models ranged from 57.6% to 91.8% accurate in their predictions. *Id.* at 338. The Supreme Court Forecasting Project, using a model based on six observable case characteristics, managed to predict the outcome of cases in the Supreme Court's 2002 term at a 75% rate of accuracy, as contrasted with a 59.1% rate for a panel of experts. *See* Theodore W. Ruger et al., *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking*, 104 COLUM. L. REV. 1150, 1151–52, 1154 & n.19 (2004).

197. Ashley and Brüninghaus conclude that "problems of representing textual cases for purposes of prediction are still major hurdles, and most prediction approaches have not been able to explain predictions in terms of legal reasons that are meaningful to legal practitioners." Ashley & Brüninghaus, *supra* note 196, at 310. As Professor Frederick Schauer has pointed out, in the context of relating the views of Karl Llewellyn, this may be because many of those reasons are not of the sort that are, in a formal sense, legally meaningful:

> Llewellyn did not deny that there were regularities in law. Nor did he deny that those regularities might facilitate the process of predicting future legal outcomes. He did, however, deny that those regularities were regularly captured by the generalizations typically referred to as "legal doctrine," and thus claimed that legal doctrine did not reflect empirical regularities, and that legal regularities were reflected by categorizations that did not resemble traditional legal doctrine.

Frederick Schauer, *Prediction and Particularity*, 78 B.U. L. REV. 773, 782 (1998).

198. *See generally* Richard Esenberg, *A Modest Proposal for Human Limitations on Cyberdiscovery*, 64 FLA. L. REV. 965 (2012).

information available in a case, enable considerably more informed assessments of whether judicial opinions tell an accurate story about the cases they resolve. Although some portions of the process will remain shielded from view, the result will be a considerably broader and more nuanced picture of how the judiciary works.