Style and Substance on the U.S. Supreme Court

Keith Carlson,
Daniel Rockmore[*]
Allen Riddell[†]
Jon Ashley,
Michael Livermore[‡][*]
[*]To whom correspondence should be addressed.

The U.S. Supreme Court is a singular institution within the American judiciary: it has many unique institutional features such as the ability to select the cases that it will decide, and it plays a unique role in American political and social life. However, while as an institution the Supreme Court is distinctive, it remains recognizably a court of law. The Supreme Court shares certain rituals with other U.S. judicial institutions, such as the black robe and gavel. It also shares many procedures with other courts, including adversarial hearings and restrictions on ex-parte contacts. Perhaps most important, is that its mode of decision making is through case-by-case adjudication, typically in the course of hearing an appeal from a lower court decision. As such, when the Court creates, amends, or clarifies legal obligations, it does so not through directly

[*] Dartmouth

[†] University of Indiana

[‡] Virginia

stated rules (as in a statute or regulation) but through the justificatory documents that accompany a disposition in a particular case, largely embodied in the written opinions of the Court and its Justices.

The judicial opinions of the U.S. Supreme Court serve as among the most important pieces of "data" in understanding the evolution of legal thought and socio-political dynamics in the United States. The judicial opinions issued by the Supreme Court have provided legal scholars (and others) with fodder for analysis since the dawn of the American legal academy. Recent advances in natural language processing and computational text analysis provide new ways to examine and understand the work of the Court. In this chapter, we will discuss our research using these kinds of tools on two sets of questions concerning the Court that are particularly well suited to such analysis. First, we examine trends in writing style on the Court through sentiment analysis and a form of stylometry based on the frequency of function works in a text. Second, we examine whether the Supreme Court has begun to carve out a unique judicial genre by examining how the content of Supreme Court opinions differs from the federal appellate courts that it supervises.

Why Study Style

Judges, lawyers, legal academics, and law students have frequently turned their attention to non-content, stylistic features of legal writing. For example, legal writing courses at American law schools evidence a desire to teach students appropriate writing style, in addition to facilitating a mastery of legal content (Romantz 2003). Practicing lawyers are often called on to persuade through the written word, and stylistic features of a text can contribute to (or detract from) its persuasive force. Guides on legal writing, geared toward law students and practicing attorneys, often pay substantial attention to non-content textual characteristics (Garner 2013). A host of stylistic conventions distinguish legal writing from standard written English, and a lawyer's competence is judged, in part, by the degree to which his or her individual stylistic voice conforms to this particular “professional discourse community" (McArdle 2006, 501).

Judge Richard Posner has defined writing style as “the range of options for encoding the paraphrasable content of a writing" (Posner 1995). Essentially, under Posner’s definition, writing style amounts to the individual imprint that judges leave on their writings, holding the legal content constant. Perhaps in part because judges are individually responsible for drafting their opinions (although there is often a good measure of group editing), there is substantial stylistic variation within judicial writings.

Writing style in judicial opinions is important for a variety of reasons. Style may serve as an indicator of judicial temperament or disposition. Stylistic norms may constrain judicial writing in ways that ultimately affect judicial reasoning, and in turn, legal outcomes. The evolution of writing style may indicate broader substantive trends on the Court. Style can affect the comprehensibility and usability of the law. Finally, style may be deserving of study simply as an empirical feature of an important cultural artifact. In fact, judicial writing style has long been the object of qualitative analysis, with commentators frequently examining (and criticizing) opinions both for basic clarity as well as (sometimes) their literary quality (Ferguson 1990).

There is a nascent movement among legal scholars to bring quantitative tools to bear on the analysis of writing style. As it turns out, an early important application of computational stylistic analysis had something of a legal (or more properly, constitutional) context as (Mosteller and Wallace 1963) brought statistical methods to the problem of identifying the authors of individual Federalist Papers. As for quantitative work directed at opinions, (Little 1998) uses a coding procedure to identify “linguistic devices that obscure" meaning and analyzes Supreme Court cases on federal jurisdiction; (Black and II 2008) examines opinion length over the entire period of the Court’s existence. More recently (Owens and Wedeking 2011) examine “cognitive clarity" in recent Supreme Court cases using the “linguistic inquiry and word court" (LIWC) software package. (Long and Christensen 2013) examines the use of “intensifiers" and readability scoring to test their theory that Justices broadcast weak legal position through use of language. (Johnson 2014) examines readability over time in the Court, comparing Flesch-Kincaid scores in the 1931-1933 and 2009-2011 terms. (Black et al. 2016) use computational tools to examine how Supreme Court writings alter language usage according to context and audience.

In addition to scholarly investigations, computational analysis of judicial writing style has even found its way into pop culture: (Chilton, Jiang, and Posner, n.d.) use token analysis–a measure of sophistication in language use–to compare the vocabulary of several Justices to famous rappers and Shakespeare in a blog post on Slate. They find that Jay Z and most of the Justices have similar vocabulary use, while rapper Aesop Rock and Justice Holmes have exceptionally large vocabulary use, and DMX and Justice Kennedy are on the low end.

In this chapter, we report two stylistic analyses that contribute to the growing judicial stylometry literature. We make use of an original dataset of U.S. Supreme Court opinions: Human researchers conducted a series of “by year" searches on a commercial database to download digitized versions of all Supreme Court cases. All proprietary information was stripped out. A series of iterative human and Python-based analyses then were carried out to separate majority, dissenting, and concurring opinions and to assign an authoring Justice and year to each opinion. Per curiam decisions were removed from the dataset, as were opinions with a file size smaller than one kilobyte. Jonathan Ashley, research librarian at the University of Virginia, was primarily responsible for identifying resources, collecting cases, and providing the markup needed for analysis.

The resulting data covers all opinions for the years 1792 to 2008.[1] Our

[1] We define a “decision" as the set of opinions that relate to a case, identifiable through a citation in the United States Reporter, for example, “347 U.S. 483 (1954)." A decision

data includes 25,407 decisions. We exclude footnotes from our analysis. There are roughly 8,000 dissents and 4,600 concurrences. We have data for 110 Justices: Justices Sotomayor and Kagan were appointed after the end of our study period. We have partial data for Justices who began their terms prior to 2008 but either retired after our study period or remain on the Court. Our analysis was conducted when Justice Scalia was an active member of the Court.

We first conduct a basic sentiment analysis of U.S. Supreme Court opinions and arrive at the interesting conclusion that the Court's language has become decidedly more "grumpy" over the course of the past two centuries. Our second analysis of the Court's writing style is more detailed and attempts to gain purchase on longstanding questions concerning the role of clerks in influencing the work of the Court. For that analysis, data concerning the number of clerks employed in chambers was provided by the Supreme Court Library.

Judicial "Friendliness"

Sentiment analysis is a form of natural language processing, which is a broader field within computer science and computational linguistics focused on human-computer interactions through language. At the heart of sentiment analysis is the concept of sentiment, which is a relation between a person and a target. Simply and intuitively, the sentiment of A toward Target X is whether A likes or dislikes X. Although, in theory, sentiment analysis could distinguish between nuanced emotional flavors, the general tendency in the field (to date at least) has been to reduce sentiment to a single dimension that ranges between positive and negative poles.

can include multiple opinions, including a majority opinion, plurality opinions, and one or more dissents or concurrences. In our data, we do not distinguish majority from plurality opinions.

A leading researcher recently defined sentiment analysis as “the field of study that analyzes people’s opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text" (Liu 2015). Prior work has focused primarily on text relevant commercial “entities"– goods and services like movies, or products for sale on Amazon. The law, broadly understood, is also full of entities of various stripes capable of generating the opinions, sentiments, appraisals, attitudes, and emotions that are amenable to sentiment analysis.

Sentiment can matter a great deal in the law when part of an actual human evaluation, attitude, or affect. A recent example helps illustrate the point. Bowers v. Hardwick was the 1986 decision of the U.S. Supreme Court that upheld a Georgia anti-sodomy statute. In his concurring opinion, Chief Justice Burger wrote separately to “underscore" his views on the matter. To establish the pedigree of anti-sodomy statutes, Burger’s opinion quotes Blackstone’s treatise on the English common law describing, an “‘infamous crime against nature’ [...] an offense of ‘deeper malignity’ than rape, a heinous act ‘the very mention of which is a disgrace to human nature,’ and ‘a crime not fit to be named.’" This language conveys an extraordinary degree of negative sentiment, and was used to justify the rejection of any constitutional protection for the private consensual sexual conduct of same-sex couples. In Lawrence v. Texas, the U.S. Supreme Court revisited this decision, and the shift in sentiment is striking, with Justice Kennedy writing:

[A]dults may choose to enter upon this relationship in the confines of their homes and their own private lives and still retain their dignity as free persons. When sexuality finds overt expression in intimate conduct with another person, the conduct can be but one element in a personal bond that is more enduring.

Further along that line of cases, Obergefell v. Hodges found that the constitution guaranteed access to marriage for same-sex couples. In justifying this decision, Justice Kenney again writing for the majority stated that: The nature of marriage is that, through its enduring bond, two persons together can find other freedoms, such as expression, intimacy, and spirituality. This is true for all persons, whatever their sexual orientation. There is dignity in the bond between two men or two women who seek to marry and in their autonomy to make such profound choices.

It is hard to imagine a more significant shift in sentiment from the Court's characterization of same-sex relationships in Bowers and Obergefell. The changing sentiment in the texts, which presumably mirrors shifting attitudes of the Court's majority, matters both in terms of the change in constitutional status for same sex couples that it accompanied, but also in the expressive function they it serves when communicated on behalf of a major institutional voice in American politics.

The difficulty of the sentiment analysis problem and the appropriate technique to be deployed are related to the degree to which a researcher attempts to zero in on targets within a single document. The simplest approach considers sentiment at the document level, which reduces all of the text within the document to a single sentiment score. Because there is no attempt to determine whether different targets are referenced, this type of analysis might be seen as that which best captures "sentiment" as the overall mood of the author. Take, for example this sentence,

Lousy Day: "I had a lousy day because my commute was blocked up by a terrible accident, and I had to wait around all morning for a boring meeting with my boss." There are different flavors of sentiment (terrible, boring) as well as several targets of negative sentiment (day, commute, accident, meeting) that could be extracted, but it is clear at the document level that the overall mood is pretty sour.

Sophisticated forms of sentiment analysis can attempt to accomplish a more finely grained analysis by either developing more specific and focused measures of sentiment or extracting the targets of sentiments within a document, or both. Imagine that the "Lousy Day" sentence above appeared on Twitter with geotagged information. If a traffic monitoring and predictive service wanted to use real-time social media information to improve its performance, it would be important to extract sentiment concerning some of the targets in the text (commute, accident) while ignoring other irrelevant sentiments.

An important component for sentiment analysis at any level is a “sentiment lexicon" that categorizes words according to the sentiment that they convey. In a sentiment lexicon, some words (e.g. wonderful, intelligent, great) will be classified as positive while others (e.g. terrible, stupid, bad) will be classified as negative. There are two general ways to construct a sentiment lexicon. The first is the thesaurus approach. For this, a researcher starts with some positive and negative seed words that have an obvious valence and then identify associated words. The second is the natural corpus approach. Here, the researcher starts with a some set of documents produced in the real world, like Amazon reviews or (for our purposes) judicial opinions. If there is already metadata related to sentiment–such as how many stars are in the review–this can be used to determine words associated with that metadata. Without this kind of metadata, seed words with known sentiment can be used to identify a larger set of related words based on whether they often co-occur in the corpus with the seed words. As sentiment analysis has become more common, there are also now off-the-shelf lexicons available. These have the obvious benefit of saving time, enabling comparability and replicability as well as reducing concerns that the lexicon is over-fit to data.

For our analysis, we use an off-the-shelf lexicon made publically available by Liu and Hu.[2] In this lexicon, there are roughly 7,000 English words that are characterized as either positive or negative. Some examples of negative words are “admonish" and “problematic"; positive words include “adventurous" and “preeminent." A Python script was programmed to determine for each Justice the total number of negative words and the total number of positive words in opinions he or she authored.

[2] Bing Liu and Minqing Hu, Opinion Mining, Sentiment Analysis, and Opinion Spam Detection, UIC, http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

The numbers of negative and positive words were then each expressed as percentages of the total number of words authored by a Justice. The percentage of negative words was subtracted from the percentage of positive words to generate what we call a “friendliness score."

This analysis—while based on measures of sentiment that have been used in a variety of other contexts—should be approached with a healthy dose of skepticism. Comparing texts over a long time horizon may be problematic for a variety of reasons, including that a text that reads as relatively friendly in one time period may read as downright nasty in another (or vice versa).

With these caveats in place, Figure 1 contains a plot of the “friendliness score" of each Justice across time, with Justices located at their median year of service. We constructed the score by subtracting the percentage positive words from the percentage of negative words. By this measure, over time the Supreme Court has gotten grumpier.

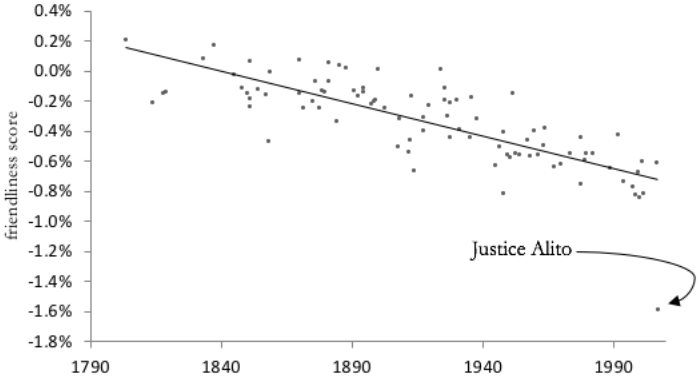

Figure 1. *Sentiment Score by Authoring Justice.*

The obvious time trend is striking and raises a number of interesting questions. Would a bespoke lexicon generate the same results? Do lower courts exhibit the same trend toward negativity? Is the trend driven by the growth of dissents? Is there a larger cultural trend toward grumpiness? These, and other, questions are worthy subjects for future inquiry.

The Role of Clerks

As judicial clerks have become an enduring feature of the operation of the federal courts, the role of these recent law graduates has been the subject of both scholarly and public debate (Peppers 2006). An important empirical predicate to this debate is the belief that clerks play a substantial role in authoring opinions. At least for the Supreme Court, there is a long history of anecdotal evidence supporting the claim that law clerks exert some influence over judicial decision-making. There is also a nascent literature that uses quantitative techniques to address the question of clerk influence over both substance and style. Techniques that have been used include the use of plagiarism software (Sulam 2014), analysis based on the party affiliation of clerks (Peppers and Zorn 2008), compression software (Choi and Gulati 2005), and identified stylistic features of opinions (such as the type-token ratio) (Wahlbeck, II, and Sigelman 2002). In (Rosenthal and Yoon 2011) a method is used similar to the one we described below based on variability of writing style but on a smaller set of data, and with a different strategy to identify clerk influence.

Our stylistic analysis relies on the use of function words in a document to serve as a broad proxy for a range of stylistic characteristics. Function words, such as “the", “a", and “at" do not directly carry content, but instead serve syntactic and grammatical functions. For purposes of the following analysis, the most important characteristic of function words is that they have been found useful as the basis for a stylistic “fingerprint" that can be used for author attribution, and we therefore use it as a proxy for writing style more generally (see e.g., (Hughes, Nicholas J. Foti, and Rockmore 2012). Our study relies on 307 standard function words (or “content-free words" or “CFWs"). The individual occurrences of each CFW can be aggregated to construct feature vectors based on some object of interest. For example, a feature vector can be constructed for each Justice, or each year.

For our analysis, we rely on an intuitive model of the process of drafting and editing judicial opinions on the Supreme Court. One of the peculiar features of the contemporary clerkship is that it is so short, typically lasting a mere year. We take advantage of this clerk turnover as a source of variability. We construct two measures: one a measure of intra-Justice inter-year variability, the other a measure of writing style consistency for the Court as an institution. We then compare our measure of consistency to a set of time periods based on a historical analysis of the role of clerks from (Peppers 2006). (These same groupings were used in (Black and II 2008).)

The first measure of variability that we introduce is centroid distance. We construct feature vectors for each text as well as for the year and calculate the distance between the text vectors and the year vector.[3] This provides a measure of how tightly clustered the Court's style is in a given year: the greater the centroid distance, the bigger the stylistic "spread." There is a clear time trend in our data, with intra-year consistency on the Court increasing over time.[4] To examine whether the overall trend toward greater consistency differed as the institution of the modern clerk developed, we conducted a structural break test on the data. The point of a structural break analysis is to determine whether there has been an underlying shift in the data generating mechanisms, such that the distribution of data from the period after the "break" is systematically different from the distribution prior to the break. We first ran a Chow struc-

[3] Our distance measure is cosine similarity, which is a representation of distance in a multi-dimensional vector space.

[4] For this analysis, we use a simple ordinary least squares regression. For time as a predictor of centroid distance, the p-value is less than 0.01% and the R-squared value is 0.5. We do not report coefficients as the distance measure itself is somewhat difficult to interpret.

tural break test, which is a standard tool to determine whether there are changes in the relationships between time and another variable over different time periods (Hansen 2001; Chow 1960). The Chow test rejects the null hypothesis that there are no structural breaks in the centroid distance data at the Peppers groups dates. For an additional text, we do not hypothesize any dates, and rather use statistical tools to test whether there is a structure break and, if so, the estimated break date.[5] The estimated break date that was returned was 1926, very close to the year that clerks took on a greater substantive role, as indicated by the Peppers group transition from "stenographers" to "assistants."

We also examine inter-year, intra-Justice variability in writing style. For purposes of our analysis, a chamber in a given year can be thought of as a "team" made up of a Justice and several clerks. A team co-produces the opinions in a given year. When clerks turn over, it changes the composition of the team. In chambers with a larger number of clerks that turn over more frequently, there will be a higher percentage of team turnover from year to year.

Although some inter-year stylistic variability can be expected even with a single author, we hypothesize that clerk turnover will decrease inter-year consistency. The dependent variable in our analysis is an inter-year consistency score. To construct the consistency score, we rely on the feature vectors based on the texts authored by a Justice in each year of his or her tenure. To calculate the consistency scores, we calculate the Kullback-Leibler divergence (relative entropy) between each year's vector (interpreted normalized to a probability distribution) and a feature vector based on the remainder of the Justice's writings. We examined the relationship between consistency score and the number of clerks that served in a Justice's chambers over the course of his or her tenure, control-

[5] For this analysis, we used the Supremum Wald test in Stata. For additional background on this test, see (Perron 2006).

ling for time through a quadratic function as well as each Justice's total production in words (under the theory that Justices who produce more words may be more consistent, and there will be less statistical noise between years). For this analysis, we examine the period after 1885, with the introduction of clerks as "stenographers" under the Peppers grouping.

Based on this model, we find the number of clerks is (statistically) significantly negatively related to the consistency score, with the interaction term indicating that clerks have had decreasing influence over time. This temporal effect may be associated with declining marginal influence of an additional clerk as the Court has institutionalized the practice of each Justice having between four and five clerks.

It is worth remembering the difficulty of fully distinguishing the effects of unobserved time-related variables from the effects of clerks. Nevertheless, over the course of the twentieth century, the intra-year stylistic consistency of the Court as an institution has increased, while the inter-year consistency of writing style for individual Supreme Court Justices has declined. Over the same period of time, law clerks have become ever more integrated into the substantive work of the Court. Because the institution of the modern law clerk in the US Supreme Court evolved gradually over time, it is hard to know the degree to which clerks have contributed to changes in writing style, independent from some other set of time-related variables. But, we have some evocative information that provides some evidence at least that the institution of judicial clerks appears to reduce intra-Justice writing style continuity that might otherwise exist while, at the same time reducing the apparent stylistic differences between Justices.

The Judicial Genre

Our second set of analyses move from writing style to the content of U.S. Supreme Court opinions.

One of the classic paradoxes of the American political system is that in a country that purports to hold democratic values dear, final decision making authority on some of the most hotly contested political issues is vested in a body that is almost entirely free from formal democratic accountability. Political scientists and sociologists have offered a variety of theories to explain the counterintuitive fact that the Court enjoys a high level of support by the public even though, from time to time, it reverses the policy choices of democratically elected branches (Gibson and Caldeira 2009). This support can be both "diffuse"– i.e., a "reservoir of favorable attitudes or good will" toward the institution, as well as "specific," which is based on happiness with individual decisions (Easton 1965). Although there is some disagreement on this point, there is evidence that the Court enjoys at least some diffuse support that is resistant to change, even in the face of disagreeable outcomes in individual case (Caldeira and Gibson 1992).

There are a variety of theories about why the Court might enjoy diffuse support. These include the "myth-of-legality" hypothesis, which holds that popular support for the Court is grounded in a widespread misperception that all legal questions can be resolved impartially and dispositively based on the neutral application of relevant law (Scheb and Lyons 2000). A more nuanced version of this hypothesis is that the public accepts that there is a degree of discretion involved in judging, but believes that the Justices exercise their discretion in a principled, public-regarding fashion, rather than strategically to benefit themselves (Gibson and Nelson 2014). A related theory is that the judicial symbols, such as the robe and gavel, help activate a positive frame that predisposes audiences in favor of the Court. In (Gibson, Caldeira, and Spence 2003) this effect is referred to as "positivity bias" and argue that judicial symbols function in this way by signaling the difference between courts and other less favorably perceived official decision makers, such as Congress or agency bureaucrats.

Relatively little empirical work has been done to assess the importance of judicial symbols in affecting perceptions of the Court. One recent study (Gibson, Lodge, and Woodson 2014) examines how exposure to judicial symbols affects the level of support given to the Court and willingness to challenge the Court's rulings. In the study, one participant group was exposed to judicial symbols–a gavel, the Supreme Court courthouse, and the Justices in their robes–while the other was not, and a survey elicited information about their responses to various judicial rulings. In general, the authors found that exposure to these symbols enhanced levels of support, especially for those with relatively less prior awareness of the Court.

At the margins, it may be difficult to distinguish between functional and symbolic characteristics of courts. To the extent that there can be a purely symbolic feature of the judicial role, robe-wearing and gavel-wielding seem like strong candidates. But it is possible that other characteristics of the Court that are more functional in nature could similarly trigger a positivity bias. For example, perhaps the ritual of oral argument serves a similar, positivity bias-triggering function, while at the same time (at least potentially) affecting substantive outcomes.

One of the most obvious distinguishing characteristics of courts is the form of the textual outputs through which their power is exercised and expressed. Judicial opinions are quite different from other textual manifestations of lawmaking, such as the statutes adopted by legislatures or the regulations promulgated by administrative agencies. Statutes and regulations take the form of more or less clearly stated rules, whereas opinions consist of narrative explanations for a decision in a particular case. The practice of issuing judicial opinions is among the most recognizable defining features of courts, especially appellate courts. Symbolically, judicial opinions may serve a role similar to robes or gavels by signaling courts' separation from the political branches.

Whether the practice of issuing recognizably judicial opinions actually reinforces the legitimacy of the Court, and if so what characteristics of opinions are responsible for that effect, are empirical questions that we do not address here. But if issuing judicial opinions helps trigger support in part by demarcating the Court as a judicial (as opposed to political) institution, the ability to do so would be bound up with how well the Court's opinions conform to public expectations of the form more generally. It is possible to think of the judicial opinion, then, as a legitimating genre. By conforming to the norms and conventions of that genre, the Court marks itself as a non-political institution and, in doing so, triggers positive associations in the relevant public that reinforce feelings of support, even for that portion of the public that might disagree with a specific decision. But, if the Court's opinions fail to conform to the judicial genre, then their value in marking the court as a non-political (and therefore more legitimate) institution may be compromised. In the following sections, we discuss our quantitative exploration of whether the Court's opinions do or do not conform to the judicial genre, and whether the degree of the Court's genre conformity has changed over time.

Establishing a Baseline. To investigate how well the Court conforms the judicial genre, we need to establish some baseline for comparison. Rather than attempt to generate an a priori account of the genre (which would doubtless be controversial) we take as a starting place a less controversial judgment about the members of the class of judicial opinions. We identify federal appellate court opinions as a baseline. Starting with this body of documents, we then examine whether the Court's opinions are distinguishable based on their semantic content, without imposing a theory about what is or is not a relevant characteristic.

Our approach can be illustrated through a simple thought experiment. Imagine a hypothetical law student, walking the corridors of a law library. This law student notices on the floor a few pages torn out of the previous year's Federal Reporter. The document lacks information identifying the authoring court. The student tries to guess whether the opinion was written by the Supreme Court or an appellate court. Our hypothetical law student's ability to guess correctly will be related to the distinctiveness of Supreme Court opinions. At one extreme, if Supreme Court opinions were written in Latin while appellate court opinions were written in Greek, the classification task would be trivial.

At the other, if the Supreme Court's docket were selected at random from all appellate court cases, and the Justices employed similar reasoning and writing styles to appellate court judges, it would be extremely difficult to improve on the prior probability estimate based purely on background frequency. If it is relatively easy to distinguish Supreme Court opinions, then on our measure, they depart from the more general genre of judicial opinions, which is defined according to the baseline corpus of appellate opinions.

Considered statically, it would be very difficult to interpret findings of distinctiveness or ease of classification, other than to note departure from pure chance. But a dynamic understanding allows for comparison between time periods. If our hypothetical student is better able to classify cases from 2004 than cases from 1954, it is fair to infer that the Supreme Court has grown more distinctive over time. This conclusion does not imply that the appellate courts have remained steady while the Supreme Court has veered off in uncharted territory. But we can say that the Supreme Court has become more distinctive relative to the appellate courts.

There are two basic mechanisms through which the opinions of the Court may come to be systematically different from those of the appellate courts that it supervises: the certiorari process and the process of opinion drafting. The hierarchal structure of the judiciary generates a vast winnowing of cases and issues before they reach the pages of the U.S. Reports. Each year, roughly one million cases are filed in federal courts.[6] In reporting year 2012, there were 35,302 federal appeals termi-

[6] Federal Judicial Center, The Federal Courts and What They Do 4, http://www.fjc.gov/public/pdf.nsf/lookup/FCtsWhat.pdf/$file/FCtsWhat.pdf (last visited Jan 15, 2017). A substantial portion (70%) of federal cases are bankruptcy filings.

nated on the merits, disposing of a number of cases roughly equivalent to 10% of the non-bankruptcy filings in the federal court (Administrative Office of the United States Courts 2013). The vast majority of these appellate dispositions were not accompanied by a published opinion.From this pool, several thousands of petitions for certiorari were submitted, with the court granting just over one hundred. The Court's control over its docket allows it substantial ability to influence its own agenda.

Given the consequences of the Court's certiorari jurisdiction, it is not surprising that it has long been a subject of study by social scientists and academic lawyers (Tanenhaus et al. 1963; Caldeira, Wright, and Zorn 1999). Based on this prior work, (**Yates201**) concludes that "[a] wealth of judicial politics literature suggests that [J]ustices have an interest in taking on cases that are salient, resolve important legal conflicts, and, in fact, do map well onto [J]ustices' distinct ideological preferences." Because the Court's docket differs in systematic ways from the general pool of appellate cases, we should expect that the opinions the Court issues will be distinguishable based on the unusual characteristics of the underlying cases. If the Court uses its certiorari jurisdiction to focus its attention on certain legal questions (such as constitutional claims or statutory interpretation) while avoiding others (such as family law issues) then its opinions will naturally reflect that emphases in its docket. Purely through the operation of the certiorari process, the body of Supreme Court opinions will reflect the issues that most capture the Court's attention.

The second mechanism that could lead to differences between the Court's opinions and those of the appellate courts is the opinion drafting process. Once certiorari has been granted, a case typically proceeds through merits briefing and oral argument, followed by drafting and editing. During the drafting phase, versions of the majority opinion, and any concurrences of dissents, are circulated within the Court, spurring additional deliberations, occasional vote-shifting, and redrafting and editing. All of these internal operations are governed by both formal rules and entrenched conventions. Most distinctly from the lower appellate courts, the Court always sits as a whole, rather than in panels, so that opinions serve as part of a running conversation among the group that has the potential to create a unique culture, especially during a period when the Court's membership is relatively stable.

There are many ways that the drafting process could lead to the Court producing opinions that are distinct from the appellate courts, even holding the underlying cases constant. The Court has considerable leeway to decide which of the legal questions presented in those cases to explore or emphasize. When the Court grants certiorari, it frequently limits review to specific questions.

Furthermore, Justices have considerable discretion when drafting majority opinions, and even more when authoring dissents or concurrences (when they are less constrained by each other). During this process, the Justices face different incentives than lower court judges, because their decisions cannot be appealed and will serve as the final word on the legal questions that they decide. Justices may, accordingly, be freer in their language, or view themselves as addressing a broader public or posterity rather than a reviewing court. The Justices may also make different choices in the language that they use, perhaps deploying certain rhetorical moves, such as personal anecdotes, metaphor, humor, or colloquialisms, that are less common in the lower courts. Given the unique processes employed by the Court, the distinctive nature of the Court's role and the audience that it addresses, and the peculiar nature of the Justices engage in bargaining, drafting, and editing, it would not be surprising if the types of reasoning or the language that is used in the Court's opinions differ from those that are used in appellate court opinions, even when the set of legal issues is the same.

A Topic Model Approach. It is theoretically possible to carry out the thought experiment discussed above in real life by presenting students with randomly generated snippets of text and asked to classify the documents as issuing from either an appellate court or the Supreme Court. But such an exercise would pose substantial technical and logistical challenges. To avoid these problems, we employ a principled application of statistical computational textual content analysis called topic model analysis. Given a textual corpus, a topic model produces topics, which in the technical topic modeling sense are probability distributions over a vocabulary, where each word in the vocabulary is assigned a non-negative weight (such that all weights sum to one). Each document is in turn summarized as a probability distribution over the topics, creating a compact but descriptive representation of the semantic content of documents. The data generated by the topic model substantially reduces the number of dimensions needed to characterize the content of documents, allowing us to engage in useful statistical analysis.

The highest-weighted words within a topic provide a sense of the subject matter that the distribution represents. For example, in Topic 3 generated by our model, the words “election", “political", “party", and “candidates" are weighted highly, which led us to hand label that topic as “elections." (Topics are generally hand labeled). Thus, the representation of a given document as a distribution over topics summarizes the document as weighted mixtures of intuitively understood themes. These distributions–both of the topics and the words they comprise–are produced as the best fit to an underlying generative probabilistic model for the observed simple word frequencies.

The canonical topic model is a latent Dirichlet allocation (LDA) mixed-membership model. The LDA model posits some number of topics (distributions over the vocabulary) that account for all words observed in a corpus according to the following generative story: for each document in the corpus, a set of topic proportions (or “shares") is drawn from a global probability distribution; then, each word in the document is drawn from a topic distribution in which the topic distribution in question is selected according to the previously mentioned document-specific set of proportions. Topic models are often fit using an iterative algorithm (known as variational approximation) or by using a Markov Chain Monte Carlo approach. In the case of the topic model, the parameters of interest are typically restricted to the topic-word distributions describing the association between topics and words, and the document-topic distributions that describe, for each document, the probability of finding words associated with each topic. See (Blei 2012) for a general overview and formal description of an LDA topic model.

While numerous incremental improvements to LDA topic modeling have emerged in the intervening years, the essence of the original model remains, and the LDA topic model persists as a general industry standard for text analysis and serves, with minor variations, as a building block in more elaborate models of text data.[7] More than a decade after the model's introduction, researchers using topic models and closely related models may be found in almost every field where machine-readable text data is abundant. Topic models are now a familiar part of the methodological landscape in the human and social sciences, from political science to German studies (Quinn et al. 2010; Riddell 2014). The data for our analysis is drawn from Public.Resource.Org, a private not-for-profit corporation that has created a digital version of the Supreme Court and Federal appellate court corpus based on the non-copyrightable information within the Westlaw database.

CourtListener, an effort within the Free Law Project, has augmented the information contained in the bulk resource data and created a user-friendly interface, which is accessible to the public. We relied on CourtListener as the source for all the texts for the Supreme Court and appellate court decisions. The set of Supreme Court documents used in this study includes the opinions associated with all formally decided full opinion cases. There are 7,528 documents in this set. The set of appellate court documents used are all published opinions issued between 1951 and

[7] For our analysis, we use a non-parametric topic model based on the Pitman-Yor Process in place of the traditional Dirichlet distributions. The "hca" software we use is authored by Wray L. Buntine and is open source. See "hca 0.61," Machine Learning Open Source Software, http://mloss.org/software/view/527/ (accessed 20 September 2015).

2007, for a total of 289,550 documents. To reduce the computational burden of fitting the topic model, we randomly selected 25,000 documents from within the appellate court set. In addition to the 25,000 randomly selected appellate court opinion documents, 4,180 appellate court documents that are associated with cases selected for review by the Supreme Court decisions are also included. In total there are 29,180 appellate court opinion documents.

To identify the set of cases that were selected for review by the Court, we gathered information from Lexis/Nexis, which provides “prior history" and “disposition" fields for Supreme Court decisions.

The vocabulary associated with the corpus comprises those words occurring at least 20 times in the entire Supreme Court corpus. There are 21,695 total words in the vocabulary. For purposes of the current analysis, the number of topics is not central to our inquiry, and so we select a convenient number of 100 topics, which is large enough to capture a great deal of the semantic variability of the corpus and small enough to make fitting the topic model computationally straightforward.

To generate the topics, all documents (i.e., both appellate court opinions and Supreme Court opinions) are treated as a single corpus and subjected to the topic model. The top words for the first ten topics generated (the order in which topics appear is not meaningful) are presented in the following table (the labeling was done by the authors):

Our analysis then breaks out three sets of documents: all opinions published in the Federal Reporters during the study period; the subset of those opinions associated with cases that were selected by the Supreme Court for review; and the subset of those opinions that were published in the U.S. Report (i.e., Supreme Court opinions). We build on the motivating thought experiment introduced above through a machine learning algorithm that mirrors the prediction task given to the hypothetical law student. Using only the information contained in the topic distributions, the goal of the algorithm is to predict whether a randomly selected opinion has been drafted by the Court. We ask both whether prediction is

Table 1. *Title of Your Table.*
Caption for for table. Citation for your table.

labels	**top words**
labor	union board labor employees employer nlrb company bargaining relations national local act unfair
family	ms mrs did told husband time testified asked sexual home stated fact received mother daughter
elections	election political party candidates candidate campaign parties primary elections contributions ballot
narcotics	united drug cocaine government cir defendant conspiracy evidence drugs marijuana possession
immunity	immunity officers officer official police law county qualified officials city conduct rights liability
prisons	prison inmates inmate prisoner prisoners officials confinement conditions security jail amendment
procedure	motion district judgment appeal rule order filed summary party appeals judge final fed notice rules
medical	dr medical hospital mental treatment health care patient drug expert patients physician condition
criminal	trial defendant plea guilty indictment united jeopardy criminal double prosecution government
insurance	insurance policy company insured coverage insurer life ins policies liability loss judgment mutual

possible, and whether the corpora of Supreme Court opinions has been growing more distinctive over time. We generate a single metric of distinctiveness based on the predictive accuracy of a logistic classifier algorithm.

This estimate of predictive accuracy has an intuitive interpretation as a measure of distinctiveness. To generate this metric, we begin by limiting the corpus of appellate court and Supreme Court opinions to a single year. Because the number of decisions in each year varies considerably–there are far more appellate court opinions in 2000 than in 1960–we randomly sample year-specific corpora of equal sizes. We then hold out fifty percent of the appellate court and Supreme Court opinions and train a basic logistic regression classification model using the remaining opinions. The only information that the classification model uses is the topic proportions in the opinions. Once the classification model has been fit, we evaluate it for accuracy on the held-out fifty percent. This task is repeated many times, each time randomly sampling the fifty percent of cases that are held out. From this procedure, we construct a distribution of predictive accuracy for the classifier for that year. We repeat these same steps for each year in our sample, providing a means of evaluating whether the distributions change over time.

There is some risk that a na{"ive classifier will become quite good at prediction based on relatively insignificant differences, for example in the usage of a few characteristic words, such as the Justices' names or the courthouse address. A high degree of predictive accuracy, if based on these small differences, would not necessarily imply substantial and meaningful distinctiveness. Use of only the topic model proportions as the basis of prediction reduces this risk. There is a substantial level of aggregation involved in moving from all words to 100 topics, and this aggregation lowers the risk that trivial differences will substantially affect the success of the classifier. The loss of information associated with topic modeling helps reduce the risk of accentuating minor differences. If we find that that the model has an easier time predicting whether opinions are authored by the Court based on topic model proportions alone, we can say with a reasonable degree of confidence that the two corpora are growing more distinct from each other in a meaningful way.

section*The Idiosyncratic Court

Figure 2 displays the results of our analysis. We confirm that the logistic regression classifier performs reasonably well in predicting the difference between appellate court and Supreme Court opinions and we find that prediction is improving considerably over time. The center of the distribution of the accuracy of held-out prediction starts at roughly 80% in the 1950s, but over time increases to over 95% in the 2000s. This is a highly significant result.[8] By the end of the study period, a quite simple classifier, using only topic proportions, achieved nearly perfect prediction. These results quite clearly indicate that Supreme Court opinions are growing more distinctive compared to those in the appellate courts. From this analysis, we know that the mix of topics present in each corpora in each year provides increasing information about the identity of the authoring court, in the sense that the classifier improves over time.

As discussed in above, both case selection and opinion drafting could result in Supreme Court opinions that are distinct from the general-pool appellate court opinions. From prior work, we know that there is an important winnowing effect during the certiorari process, and the cases that come before the Court are far from randomly drawn. At least part of the reason that the Court's opinions are distinctive is that they are based on a non-representative set of cases. It is worth considering whether the contribution of the certiorari process to the distinctiveness of the Court's opinions is growing, declining, or remaining relatively flat. The converse question is whether the opinion-drafting process has changed over time such that, holding the underlying cases constant, the Court's discussion

[8] Pearson product-moment correlation between year and accuracy is 0.79. The error bars for Figures 2 and 3 indicate the range between the 10th and 90th percentiles.

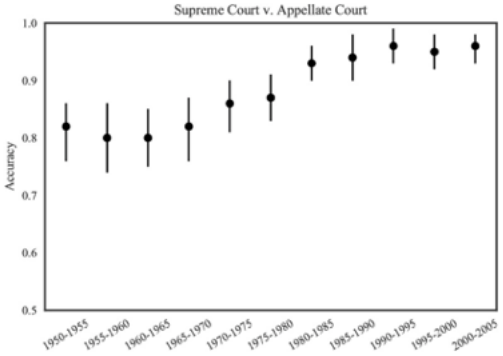

Figure 2. *Prediction of Supreme Court Opinions.*

of those cases has become increasingly distinctive over time.

To investigate these two questions, we carry out the same logistic regression classifier analysis on three different corpora: the set of all appellate court opinions, the set of appellate court opinions associated with cases selected for review, and the Court's opinions. We then carry out two sets of analyses, using the cases selected for review set as an intermediary corpus. We first examine whether the Court is using its certiorari power more aggressively than in the past in the sense of selecting cases that are more distinct from the pool of all appellate court cases. If so, we should find that the performance of the classifier would increase over time. We then examine the opinion-drafting process by analyzing whether the Court's opinions are growing more distinctive vis-á-vis the intermediary corpus of appellate opinions associated with cases selected for review. In essence, this analysis holds the underlying legal issues constant to determine whether the Court is discussing those issues in a more distinctive fashion. The results of these two analyses are reported in Figure 3.

We do not find any evidence that there is any change over time in the representativeness of the group of cases being selected for review. Although this analysis cannot rule out the possibility that a more sensitive

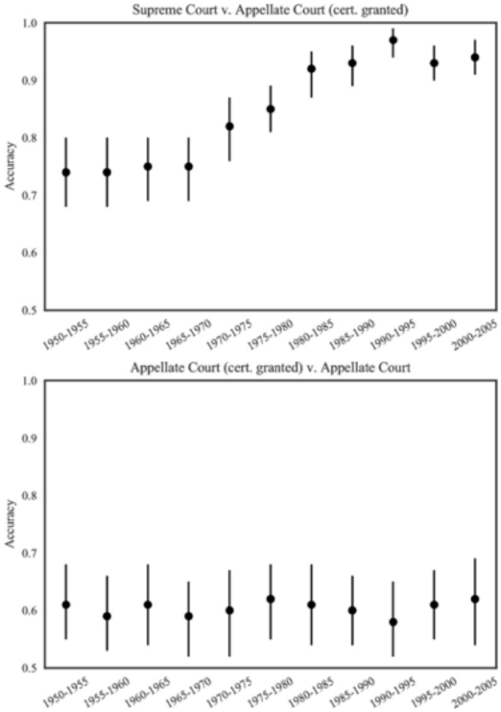

Figure 3. *Prediction of Supreme Court and Appellate Court Opinions.*

textual analysis would identify some temporal change, we fail to find any such effect using the same model that identifies an overall growth in the distinctiveness of the Court's opinions. We can therefore say with confidence that the increasing distinctiveness that we identify is not caused by a change in the level of representativeness in the cases selected for review.

Since the Court's opinions are growing more distinctive, yet the underlying cases selected for review are not, the natural inference is that the Court's opinions must be becoming more distinct from the appellate court cases selected for review. We confirm this conclusion, finding that, when comparing Supreme Court opinions and the intermediary corpus of appellate opinions associated with cases selected for review, the performance of the classifier improves over time. Starting with accuracy centered at roughly 70% in the 1950s, performance increased to well over 90% by the 2000s. The lesson from this analysis is that, although the cases selected for review in recent years are no more distinct from the pool of all appellate court cases than in the past, the way that the Supreme Court analyzes and discusses the legal issues presented in those cases has grown increasingly idiosyncratic over time.

This finding is quite striking and indicates that, at least according to the measure developed and discussed above, the Court's opinions conform less well to the genre of judicial opinions than in the past, and this change is due to the opinion-drafting process in the Court. Opinions written by the Supreme Court are more characteristic and easily identifiable than in the past; they are, on their face, less-obviously associated with the opinions drafted by the appellate courts.

Concluding Thoughts

In this chapter we explore how several computational text analysis tools can be deployed to better understand the U.S. Supreme Court. Both the stylistic analysis based on sentiment analysis and function words, and the substantive analysis based on topic models, hold substantial potential to contribute to further empirical study of the law. In the past, quantitative analysis of law has traditionally been hampered by the lack of attractive mechanisms for estimating case characteristics or the legal features of opinions. Style analysis and topic modeling provide promising avenues to estimate difficult to capture variables related to the legal content of opinions and judicial temperament. These tools also avoid some of the pitfalls of human readers, including error, bias, and, most important, time and attentional limits. By naïvely characterizing the relevant features of judicial opinions, topic models and stylistic analyses provide a quantitative and computationally tractable method to represent the text of the law. The corpus of the law–the published case law in the state and federal reporters, and other legal texts as well–is an enormous and rich dataset, and computational tools provide an effective means of capturing important characteristics of that data that can be subjected to analysis. With researchers continually introducing new tools and refining existing approaches, there is an ever-expanding frontier in empirical legal scholarship that has substantial potential to improve understanding of the law.

References

Black, Ryan C., and James F. Spriggs II. 2008. "An Empirical Analysis of the Length of U.S. Supreme Court Opinions." *Houston Law Review* 45:622–682.

Black, Ryan C., Ryan J. Owens, Justin Wedeking, and Patrick C. Wohlfarth. 2016. *U.S. Supreme Court Opinions and Their Audiences.* Cambridge University Press.

Blei, David. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84.

Caldeira, Gregory A., and James L. Gibson. 1992. "The Etiology of Public Support for the Supreme Court." *American Journal of Political Science* 36:635–664.

Caldeira, Gregory A., John R. Wright, and Christopher J.W. Zorn. 1999. "Sophisticated Voting and Gate-Keeping in the Supreme Court." *Journal of Law, Economics and Organization* 15:549–572.

Chilton, Adam, Kevin Jiang, and Eric Posner. n.d. "Rappers v. SCOTUS: Who Uses a Bigger Vocabulary, Jay Z or Scalia?" *Slate.* `http://www.slate.com/articles/news%5C_and%20%5C_politics/view%5C_from%5C_chicago/2014/06/supreme%5C_court%5C_and%5C_rappers%5C_who%5C_uses%5C_a%5C_bigger%5C_vocabulary%5C_%20jay%5C_z%5C_o_%CC%8Ascalia.html`.

Choi, Stephen J., and G. Mitu Gulati. 2005. "Which Judges Write Their Opinions (And Should We Care?)" *Florida State University Law Review* 32:1077–1122.

Chow, Gregory C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28 (3): 591–605.

Easton, David. 1965. *A Systems Analysis Of Political Life.* London: John Wiley / Sons Ltd.

Ferguson, Robert A. 1990. “The Judicial Opinion as Literary Genre.” *Yale Journal of Law and Humanities* 2 (1): 201–219.

Garner, Bryan A. 2013. *Legal Writing in Plain English: A Text with Exercises.* University of Chicago Press.

Gibson, James L., and Gregory A. Caldeira. 2009. *Citizens, Courts, And Confirmations.* Princeton, New Jersey: Princeton University Press.

Gibson, James L., Gregory A. Caldeira, and Lester Kenyatta Spence. 2003. “Measuring Attitudes toward the United States Supreme Court.” *American Journal of Political Science* 47 (2): 354–367.

Gibson, James L., Milton Lodge, and Benjamin Woodson. 2014. “Losing, but Accepting: Legitimacy, Positivity Theory, and the Symbols of Judicial Authority.” *Law and Society Review* 48:837–866.

Gibson, James L., and Michael J. Nelson. 2014. “The Legitimacy of the U.S. Supreme Court: Conventional Wisdoms and Recent Challenges Thereto.” *Annual Review of Law and Social Science* 10:201–219.

Hansen, Bruce E. 2001. “The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity.” *Journal of Economic Perspectives* 15 (4): 117–128.

Hughes, James M., David C. Krakauer Nicholas J. Foti, and Daniel N. Rockmore. 2012. “Quantitative Patterns of Stylistic Influence in the Evolution of Literature.” *Proceedings of the National Academy of Sciences* 109:7682–7686.

Johnson, Stephen M. 2014. “The Changing Discourse of the Supreme Court.” *New Hampshire Law Review* 12:29–68.

Little, Laura E. 1998. “Hiding with Words: Obfuscation, Avoidance, and Federal Jurisdiction Opinions.” *UCLA Law Review* 46:75–112.

Long, Lance N., and William F. Christensen. 2013. "When Justices (Subconsciously) Attack: The Theory of Argumentative Threat and the Supreme Court." *Oregon Law Review* 91:933–960.

McArdle, Andrea. 2006. "Teaching Writing in Clinical, Lawyering, and Legal Writing Courses: Negotiating Professional and Personal Voice." *Clinical Law Review* 12:441–499.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58:275–309.

Owens, Ryan J., and Justin P. Wedeking. 2011. "Justices and Legal Clarity: Analyzing the Complexity of U.S. Supreme Court Opinions." *Law and Society Review* 45:1027–1061.

Peppers, Todd C. 2006. *Courtiers Of The Marble Palace: The Rise And Influence Of The Supreme Court Law Clerk.* Stanford Law / Politics.

Peppers, Todd C., and Christopher Zorn. 2008. "Law Clerk Influence on Supreme Court Decision Making: An Empirical Assessment." *Depaul Law Review* 58:51–78.

Perron, Pierre. 2006. "Dealing with Structural Breaks." In *In Palgrave Handbook Of Econometrics: Econometric Theory,* edited by Terence C. Mills and Kerry Patterson, 278–352.

Posner, Richard A. 1995. "Judges' Writing Styles (And Do They Matter?)" *University of Chicago Law Review* 62:1421–1451.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54:209–228.

Riddell, Allen. 2014. “How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models.” In *In: Distant Readings: Topologies of German Culture in the Long Nineteenth Century,* edited by Matt Erlin and Lynne Tatlock, 91–114.

Romantz, David S. 2003. “The Truth About Cats and Dogs: Legal Writing Courses and the Law School Curriculum.” *University of Kansas Law Review* 52:105.

Rosenthal, Jeffrey S., and Albert H. Yoon. 2011. “Detecting Multiple Authorship of United States Supreme Court Legal Decisions Using Function Words.” *Annals Applied Statistics* 5:283–308.

Scheb, John M., and William Lyons. 2000. “The Myth of Legality and Public Evaluation of the Supreme Court.” *Social Science Quarterly* 81:928.

Sulam, Ian. 2014. *Editor in Chief: Opinion Authorship and Clerk Influence on the Supreme Court (unpublished manuscript).* http://icsulam.github.io/pdf/EditorInChief.pdf.

Tanenhaus, Joseph, Marvin Schick, Matthew Muraskin, and Daniel Rosen. 1963. “The Supreme Court’s Certiorari Jurisdiction: Cue Theory.” In *Judicial Decision-Making,* edited by Glendon Schubert, 113–115. Glencoe: Free Press.

Wahlbeck, Paul J., James F. Spriggs II, and Lee Sigelman. 2002. “Ghostwriters on the Court? A Stylistic Analysis of U.S. Supreme Court Opinion Drafts.” *American Political Research* 30:166–192.