– ARTICLE TYPE

An Expert-Sourced Measure of Judicial Ideology

Kevin L. Cope

Associate Professor of Law and Public Policy, Associate Professor of Politics (by courtesy), University of Virginia, 580 Massie Rd., Charlottesville, Virginia 22903. Email: kcope@law.virginia.edu

Abstract

This Article develops the first dynamic method for systematically estimating the ideologies and other traits of nearly the entire federal judiciary. The Jurist-Derived Judicial Ideology Scores (JuDJIS) method derives from computational text analysis of over 20,000 written evaluations by a representative sample of tens of thousands of jurists as part of an ongoing, systematic survey initiative began in 1985. The resulting data constitute not only the first such comprehensive federal-court measure that is dynamic, but also the only such measure that is based on judging, and the only such measure that is potentially multi-dimensional. The results of empirical validity tests reflect these advantages. Validation on a set of several-thousand appellate decisions indicates that the ideology estimates predict outcomes more accurately than the existing appellate measures, such as the Judicial Common Space. In addition to informing theoretical debates about the nature of judicial ideology and decision-making, the JuDJIS initiative might lead courts scholars to revisit some of the lower-court research findings of the last two decades, which are generally based on static, non-judicial models. Perhaps most importantly, this method could foster breakthroughs in courts research that, until now, were impossible due to data limitations.

Keywords: judicial ideology, courts, measurement, natural language processing

For decades courts scholars have sought to explain systematically why judges decide cases as they do, often by estimating quantitatively the *ideologies* of judges (see, e.g., Rohde and Spaeth 1976; Segal and Cover 1989; Bonica and Sen 2021; Bailey 2017). Broadly speaking, the methods for doing so comprise three types: coding and counting judicial votes; using the preferences of related political actors as proxies for the judge's ideology; and drawing on third-party commentaries of the judges. Together, the data produced by these measures over the last few decades have opened the door to a new line of research, which now forms a significant component of judicial politics scholarship.[1] As with any measure, each of these methods has unique strengths and limitations (see Bonica et al. 2017). As to limitations: many metrics are highly attenuated from the concept they purport to measure; most cannot capture more than one dimension; most cannot capture change over time; many fail to distinguish judicial from political ideology; and some capture only a fraction of the judges within a given court. Recognizing these and other limits of the existing methods, in the forthcoming first edition of the *Oxford Handbook of Comparative Judicial Behaviour*, Epstein, Martin, and Quinn (2024) note the lack of

[1] As of 2024, the three leading judicial ideology measures – Segal and Cover (1989), Martin and Quinn (2002), and Epstein et al. (2007) – had collectively been used or cited in nearly 3,000 published courts studies.

any comprehensive metric derived from expert experience, and they call for a new research agenda along those lines.

This Article introduces such an initiative: the *Ju*rist-*D*erived *J*udicial *I*deology *S*cores (JuDJIS, pronounced "judges"), an expert-sourced approach to measuring judicial traits, which attempts to overcome each of the limitations above. Many legal and political scholars have recognized the challenge of capturing judicial ideology, an inherently subjective, multi-dimensional concept, with voting or political behavior data alone. In that vein, my starting point and core assumption is that legal practitioners and other experts have special insight into how judges decide cases, insight not fully captured by either the often-quite-political judicial-appointment process, or judges' own political (as opposed to judicial) behavior.

The JuDJIS scores are estimated using a new text-analysis technique technique that quantifies tens of thousands of written evaluations, systematically solicited from a broad, representative sample of judicial experts. The information is collected by professional survey firms commissioned by the *Almanac of the Federal Judiciary*, a tri-annually published initiative which surveys a stratified sample of qualified experts for each judge. In response to prompts, the experts use their own words to evaluate the judges in five categories, including: ability; demeanor; trial practice/oral argument; settlement/opinion quality; and ideology, with all comments generally published. The evaluations are provided every few years by a set of eight-to-ten lawyers and ex-judicial clerks, each with significant professional experience with the judicial decisions and courtroom behavior of the judge(s) he or she is evaluating. Thus, an established judge would be expected to be evaluated by approximately 16–30 different lawyers over a ten-year period.

The *Almanac*'s 40 years of written evaluations to date comprise approximately 14,500 documents and 11 million words, with updated volumes released every few months. The corpus comprising this complete set of *Almanac* volumes was quantified to produce scaled estimates of ideology and other judicial traits.[2]

The JuDJIS method has several important advantages over existing non-Supreme Court measures of judicial ideology. It is the only such comprehensive measure: (1) that is based on judging (rather than, say, campaign contributions or congressional votes); *or* (2) that allows for change over time; *or* (3) that can produce scores comprising multiple dimensions, covering several non-ideology judicial traits. Its eventual scope – essentially all Article III lower-court judges, over 4,900 in total as of 2024 – is larger than any existing set of scores. Perhaps most importantly, the JuDJIS circuit ideology data predict the outcomes of a representative set of case decisions with significantly greater accuracy than any of the three leading circuit-judge ideology

[2] I call this approach *expert-sourced* because of its resemblance to *crowd-sourcing* techniques (see, e.g., Benoit et al. 2016), with the crowd in this case comprising, not the general public, but selected experts.

measures.

This Article contributes to the field of judicial behavior in three key ways. First, by developing a behavioral model of judicial ideology and showing that the judgements of legal experts can outperform political metrics in predicting case outcomes, it informs ongoing theoretical debates about the nature of judicial ideology and decision-making (see, e.g., Bonica and Sen 2021; Converse 2006; Fischman and Law 2009; Gerring 1997; Lammon 2009), including in the United States, but also in the courts of other countries and international bodies. Second, I hope the JuDJIS method – being judging-based, dynamic, comprehensive, and multi-dimensional – will spark a new line of judicial behavior research, by allowing researchers to raise and analyze important questions in judicial politics that have thus far been intractable due to data limitations. It therefore contributes to a line of path-breaking innovations in measuring judicial ideology, such as methods developed by Segal and Cover (1989), Martin and Quinn (2002), and Epstein et al. (2007). And analogous to Martin and Quinn (2002)'s observation in the context of developing the first dynamic model of the Supreme Court, a dynamic model of the lower courts may also call into question some previous circuit and district court research, which is based almost entirely on static models. Finally, because it applies political-science scaling methods to content derived from doctrinal-legalist perspectives, I hope the JuDJIS method will help to further bridge the theoretical and methodological gulfs that still divide these disciplines.

Section 1 presents the behavioral model and theoretical assumptions motivating the method. Section 2 explains the hierarchical ngram measurement method used to generate the scores. Section 3 empirically validates the method. Section 4 presents the scores for U.S. circuit court judge ideologies, 1990–2017. Section 5 concludes with possible applications for the method.

1. Measuring Judicial Ideology

Since the early-to-mid twentieth century, researchers have been attempting to use quantitative measures to measure judicial behavior (e.g., Gaudet 1933; Schubert 1960; Nagel 1961). In an attempt to predict and explain judges' rulings, social scientists inspired partly by insights from the attitudinal model of judging have developed a variety of quantitative measures that purport to capture judges' ideology.[3]

Those existing methods can be divided into three categories: vote counting, proxy, and third-party (cf. Cope 2024; Fischman and Law 2009). *Vote-counting* entails estimating judicial ideology from judges' preferred case outcomes as expressed with their votes: either *guided* (in which researchers attribute substantive values

[3] For a more detailed discussion of these measures' underlying theoretical assumptions, strengths, and weaknesses, see Cope (2024).

to judges' votes) (e.g., Spaeth et al. 2014) or *agnostic* (recording only whether a judge voted with the majority or minority) (e.g., Martin and Quinn 2002; Windett, Harden, and Hall 2015; Voeten 2007).

Proxy measures draw on observable traits that are theoretically related to judicial ideology but conceptually distinct from it. The proxy is selected because the researcher believes it is empirically correlated with the latent trait of judicial ideology. Early proxy methods included judges' self-identified party affiliations (Schubert 1960; Nagel 1961). Recent proxy methods involve selecting a political actor linked to the judge – often the appointing party or executive – and measuring that actor's political ideology as a stand-in for the judge's judicial ideology.

More complex proxy measures incorporate the ideal points of other political actors. They include the Common Space scores (Giles, Hettinger, and Peppers 2001) and Judicial Common Space (JCS), which rely on the congressional-voting-based NOMINATE ideal points (Poole and Rosenthal 2000) of U.S. senator involved in a judge's nomination (Epstein, Walker, and Dixon 1989), taking advantage of the Senate "blue-slip" custom (McMillion 2017). Specifically, the JCS considers the NOMINATE score of the judicial vacancy state's U.S. senator who shares the appointing president's party, and it attributes the senator's score to that judge. Boyd (2011) extends the JCS method to district courts, creating a data set of district judge ideology. Another proxy measure, the Clerk-Based Ideology (CBI) scores (Bonica et al. 2017), first estimate the political ideology of U.S. federal law clerks based on their campaign contributions. Based on the assumption that judges wish to hire clerks who share their ideologies, CBI uses the mean score of a judge's clerks as a proxy for the judge's own judicial ideology. Similarly, Bonica and Sen (2017) use the campaign contributions that judges themselves make to candidates for political office (typically, before the judges took the bench) as a proxy for the judges' judicial ideology.

Third-party measures consist of observers' judgments or predictions about a judge's ideology. Whereas proxy measures derive from personal political behavior of actors in some way connected to the judge, third-party measures purport to observe and evaluate judicial ideology itself (past or anticipated). There are two main types of third-party measures: editorial-based and expert-based. The leading editorial-based third-party measure, Segal and Cover (1989)'s measure of U.S. Supreme Court justices' ideology relies on human-coded text analysis of media editorials written about judicial nominees after their nomination but before confirmation.

Third-party expert-based measures are an especially promising, but as-of-yet barely explored (Grendstad, Shaffer, and Waltenburg 2012; Wijtvliet and Dyevre 2021) third-party method. Expert-based measures use opinions of legal experts to create quantitative estimates of judicial ideology and other traits. Like other methods, this approach presents several logistical challenges, but it also has many attractive qualities that might overcome some of the shortcomings of existing approaches. Thus far, however, they have seen no known application at large scale, or using computational text analysis, or to any U.S. court.

1.1 An Expert-Sourced Measure of U.S. Federal Judges

The JuDJIS method introduced here incorporates a technique that, to my knowledge, has not been used to measure ideology systematically in any field: computational text analysis of expert evaluations. The method involves a novel text-analysis technique, quantifying to date over three decades of evaluations by tens of thousands of legal experts, eventually covering over 4,900 judges. The evaluations are compiled by academic publisher Wolters Kluwer's *Almanac of the Federal Judiciary* ("the Almanac") and are provided by lawyers and ex-law clerks – in their own words in response to prompts – with professional familiarity with each of the judges, including the judges' written opinions, judging styles, and courtroom/chambers demeanors. In what follows, I set forth the assumptions underlying JuDJIS's behavioral model of measuring judicial behavior.

1.2 Behavioral Model and Theoretical Assumptions

The choice to measure ideology from written expert evaluations makes two primary assumptions. First, ideology is a latent trait and therefore not directly observable (Fischman and Law 2009). But I believe that a judge's "normative beliefs about the appropriate functions of law and courts" directly cause him or her to apply the law in certain ways in written opinions and oral decisions. This judicial behavior – e.g., written orders and opinions, private settlement discussions in chambers, and comments made from the bench – is therefore a direct manifestation of the judge's ideology, and it is the closest thing to ideology itself that can be observed. It is therefore unique among comprehensive lower court measures in capturing *expressed* ideology rather than others' *expectations* of ideology.

Second, judicial "votes" are undeniably a key type of judicial behavior and highly probative of judicial ideology. But a judge's judicial reasoning in her written opinions, orders, and analysis from the bench – which the evaluations capture indirectly via the lawyers' observations of those behaviors – provides clearer insight into her judicial ideology than her decision to simply affirm or reverse.

For these reasons, a well-designed expert-sourced method based on observed judicial behavior should be able to come reasonably close to observing latent judging traits, including judicial ideology. At a minimum, expert-sourced evaluations are likely to relate *more* closely to ideology than the observations underlying most or all current judicial ideology measurement approaches, for instance, pre-confirmation speculation about future behavior (Segal and Cover 1989), the campaign contributions of law clerks (Bonica et al. 2017), the appointing president's party (Schubert 1960; Nagel 1961), the congressional voting records of pre-nomination supporting senators (Epstein et al. 2007), and perhaps even case votes (Martin and Quinn 2002; Spaeth et al. 2014). In light of these assumptions, the JuDJIS method attempts to address potential threats to validity and sensitivity, which can affect any measure of judicial ideology.

2. Hierarchical Ngram Text Analysis

I next describe the process for collecting the underlying source material and generating the JuDJIS data.

2.1 The Almanac

The *Almanac* has been published by Wolters Kluwer Publishing since 1985. It contains detailed biographical data and subjective evaluations entries on all judges in the federal judiciary (including senior judges, as well as bankruptcy, magistrate, and other Article I judges). The judges' entries comprise biographical information, key cases, and, most important for these purposes, the exclusive-to-the-*Almanac lawyers' evaluations*. In response to interviewer prompts, a set of lawyers evaluates different aspects of each judge using the lawyers' own words. The *Almanac* routinely updates the judges' entries, with all judges within a given district or circuit updated in a single batch. The responses form the corpus for the data set.

Wolters Kluwer contracts with professional third-party survey firms to conduct the lawyers' evaluations. For each judge, they seek a stratified representative sample of lawyers who have substantial and recent familiarity with that judge. All evaluators are guaranteed anonymity to promote candor. The surveyors attempt to represent criminal and civil lawyers in approximately equal numbers. The strategy to identify the appropriate sample of lawyers is tailored to the particular jurisdiction in question, as different types of districts (e.g., rural/urban, Northeast/South) have different dynamics between and within the federal courts and bar, but the overarching goal for every court is to achieve a representative sample of those familiar with the judges of the court. Indeed, the business model of the for-profit publication depends on its reputation for accuracy, requiring it to consistently present valid and unbiased information.

The surveyors identify lawyers through a variety of means, including official court records of appearances and third-party publications listing prominent lawyers in the district or circuit. In general, the lawyers interviewed have personally appeared multiple times over the previous few years before the judge in question. For judges who have served for several years, the surveyors interview eight to ten lawyers per survey. In general, all lawyer comments are published, often abridged for the most relevant and substantive language. Over a typical 10-year period, approximately 16 to 30 different lawyers give comments on any established judge. The appendix provides some examples of typical evaluations.

2.2 Text-Analysis: Hierarchical Ngram Analysis

To analyze the corpus, I use a novel text-analysis method that I introduce here: hierarchical ngrams. The method has several advantages – particularly for this type of corpus and research objective – over large-language model machine-learning approaches, most notably its transparency, explainability, and replicability (see, e.g., Albaugh et al. 2014; Albaugh et al. 2013; Grimmer and Stewart 2013). It also overcomes some shortcomings of conventional lexicon-based methods like unigrams or bigrams, such as their inability to draw meaning from syntactic context (c.f. Farah and Kakisim 2023).[4]

The hierarchical ngram method involves the following steps. First every ngram from length 2 (bigram) to length 9 (novagram) in the corpus was identified. For the *Circuit Ideology* dataset, this initial process generated a set of 4,791 unique ngrams. For each of the eight ngram lengths, a threshold frequency was determined for inclusion in the coded dictionary, without regard to content or meaning. The threshold level was determined after examining the relative amount of information contained in each set of ngrams, balancing the conflicting objectives of the greatest possible context and maintaining a dictionary of manageable size.[5] To further improve coverage, some of these ngrams contain wildcards, i.e., a word representing any possible string of words from length 0 to 4, which are selected based on their incidence in the corpus. For instance, <no * leaning> comprises, e.g., <no leaning>, <no apparent leaning>, and <no impression about his leanings>. For the *Circuit Ideology* dataset, the resulting dictionary comprises 2,175 unique ngrams.

Second, a set of trained coders – each upper-level law students with backgrounds in federal courts – assigned a value, ranging from -3 (extremely liberal) to 3 (extremely conservative), or 99 (no ideological salience) to each ngram. For other datasets, appropriate alternative values are used. All ngrams were coded by three coders using the codebook contained in the appendix. I resolved any discrepancies.[6]

[4] In this case, the hierarchical ngram method also marginally outperforms large-language model-based and conventional lexicon-based in predicting case outcomes. See appendix section A2.

[5] For circuit ideology, the thresholds are: bigram: 45; trigram: 25; quadgram: 10; quintgram: 4; sexgram: 4; septgram: 4; octogram: 4; novagram: 4.

[6] The intercoder reliability scores are as follows: for the initial decision on salience/non-salience (i.e., 99 or -3:3), the Krippendorff's alpha intercoder reliability score for the three coders is 0.84. For the ngrams for which the three coders unanimously agreed on salience (41.2% of ngrams), the Krippendorff's alpha intercoder reliability score (Landis and Koch 1977) is 0.95, as to the exact ideological value assigned.

A majority of ngrams were coded as non-salient, with 45.4% of ngrams assigned a value of -3 to 3. To place the scores on the same scale as some existing ideology measures, such as Martin and Quinn (2002) and JCS, the [-3,3] scale was then converted to a [-1,1] scale.

Next, the dictionary values were assigned to the judges in the corpus. As Figure 1 shows, a judge evaluation (*j*) comprises one or more *comments* (each by a single expert reviewer, on a given topic of a given judge in a given year). Each such comment comprises one or more ngrams (*g*) of length 2–9.

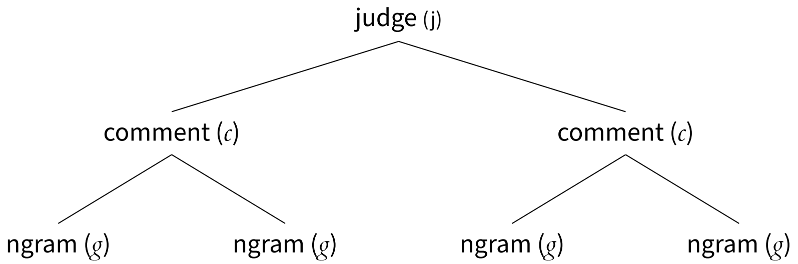


Figure 1. Structure of Document Components (evaluation-comment-ngram)

To assign the dictionary values to the corpus, I use the hierarchical ngram algorithm described below and in the appendix. Intuitively, for each comment, the algorithm identifies all ngrams in the comment that have been assigned a score in the dictionary. There are often multiple non-overlapping ngrams in a given comment that each receive a score. For example, in the following comment, <*In employment cases, she's not really pro-employee, but not really pro-business either; she's usually right down the middle*>, each of the three underlined ngrams would be scored.

Not all ngrams are counted, however, as many overlap with, or are nested within, other ngrams. For instance, the comment <*she's not a particularly harsh sentencer in drug cases*> contains the ngrams <*harsh sentencer*>, <*particularly harsh sentencer*>, and <*not a particularly harsh sentencer*>. Each of these ngrams obviously has different meanings. In order to avoid counting and tallying redundant – or worse, conflicting – scores of overlapping or nested ngrams, the algorithm recognizes a *hierarchy* of ngrams based on length: only the senior ngram, i.e., the longest in any set of nested ngrams is considered and scored.

After the hierarchial process identifies the ngrams to be scored, a judge evaluation is calculated in the following way. First, the ideology of a given comment, i_c, is defined as:

$$i_c = \frac{\sum_{g=1}^{G_c} i_g}{G_c} \tag{1}$$

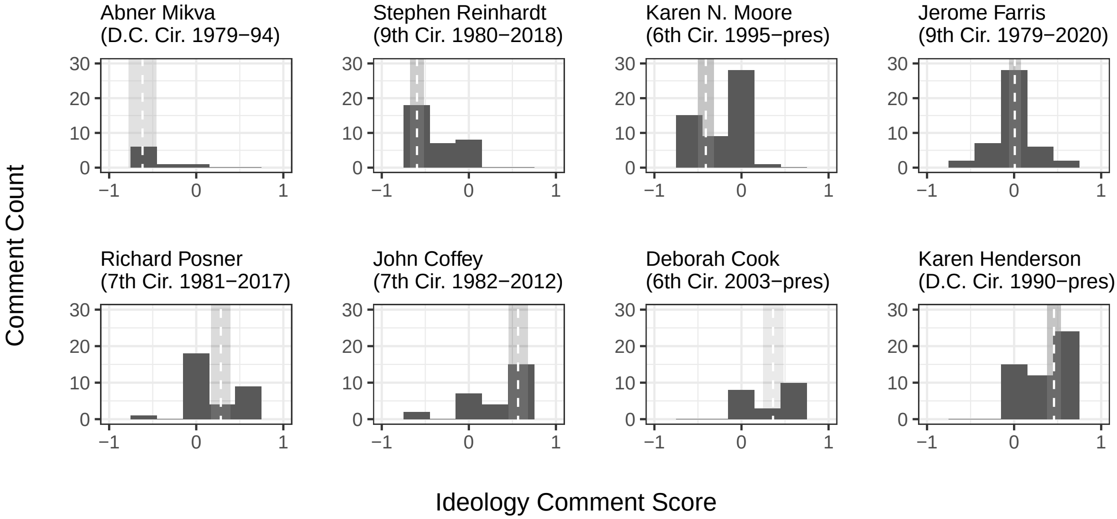
where a comment, c, comprises G_c ngrams, and each ngram is assigned an ideology, i_g. Thus, i_c is the mean ideology score of the G_c ngrams that make up comment c. In turn, each judge, j, for a given year or set of years has an observed ideology, i_j, defined as:

$$i_j = \frac{\sum_{c=1}^{C_j} i_c}{C_j} \tag{2}$$

where the judge's evaluation(s), j, comprises C_j comments. Thus, i_j is the mean ideology score of the C_j comments that make up the judge evaluation(s).[7]

To illustrate how this process can produce scores for particular judges, Figure 2 shows the distribution of comment scores for eight circuit judges: four Democratic-appointed (top row) and four Republican-appointed (bottom row). Each judge differs from the others in both mean ideology score and in distribution and variance. These histograms thus illustrate how the measure can be sensitive to small true variation between individual judges, even those appointed by the same president.

[7] Note that this process ensures that the reviews of evaluators who give lengthier comments are not given more weight than those who give shorter ones. For example, in the comment above, the method might assign each of the scored ngrams a 0, but the comment would contribute only one '0' (the mean of 0, 0, 0) to the evaluation score, not three '0's.



Note: Vertical dashed lines denote the mean score. Gray bands denote 90% confidence intervals around the mean score.

Figure 2. Distribution of JuDJIS Ideology Comment Scores for Sample of Eight Circuit Judges, Aggregated Over Their Tenures

For instance, while comments on Judge Stephen Reinhardt's ideology were not uniform, the majority of comments placed him in the liberal range (–1.5 to –2.5). This relative agreement, combined with the sheer number of comments amassed over his long career on the bench, contribute to fairly high confidence (and therefore, narrow confidence intervals, as indicated by the gray vertical band) about his mean score (indicated by the vertical dashed lines). In contrast, comments on Judge Richard Posner show a more bimodal distribution, with most evaluators split between labeling him moderate (–.5 to .5) and conservative (1.5 to 2.5). Perhaps this division occurs because Judge Posner is often characterized as more libertarian than conservative (Harcourt 2007), meaning his somewhat left-leaning views on some social issues made it difficult to reach a consensus on his placement on a single-dimensional, left-right scale. The resulting uncertainty about his mean score is somewhat larger than that for Judge Reinhardt.

2.3 Advantages and Potential Critiques

Before empirically validating the method, I explore several of the method's *a priori* advantages and potential critiques. Given the theory and method underlying the JuDJIS measures, they have several attractive properties and advantages over existing methods for measuring judicial ideology in the circuit and district courts.

First, the JuDJIS method can include the entire Article III judiciary on one scale, and it tracks changes over time, running from 1990 to the present. Moreover, by drawing on the experiences of legal experts who have studied their opinions and interacted with them in person over time, the JuDJIS scores capture the more subtle nuances of judging, albeit indirectly, beyond simple votes to affirm or reverse. In this one narrow sense, the technique is similar to Segal and Cover (1989)'s analysis of media op-eds, which, for nominees with judicial experience, draw on the judge's previous opinions. But unlike Segal and Cover's source material, which is locked in before the justice is even confirmed, the JuDJIS evaluations are updated continuously based on the judge's conduct in that judicial position.

Finally, the JuDJIS scores are sensitive to small true variation between judges and to ideological evolution over time. Indeed, because the evaluators' scores constitute a sample of a theoretical population of all potential evaluators, it can estimate standard errors and confidence intervals for each point estimate.[8]

The JuDJIS method is also potentially subject to some critiques, which I attempt to anticipate and address here. First, the evaluations necessarily come from lawyers most familiar with the judges, as no single set of experts is substantially familiar with all (or even a meaningful fraction of) the federal judiciary. Thus, different mixes of evaluators evaluate the judges. So although I cannot definitively rule them out, there is no particular reason to expect systematic differences in evaluator characteristics, given the *Almanac*'s objectives and survey methodology. Indeed, it has this trait in common with other leading and established social science indicators.[9] Consider further that federal appellate lawyers are generally cosmopolitan legal actors, with many practicing in several circuits. Appellate lawyers tend to keep abreast of developments in other circuits and the Supreme Court. There is no requirement that the lawyers interviewed for the *Almanac* are geographically located within the circuit they evaluate – only that they have personal experience and expertise with the judge in question.

A related potential issue is the possibility of systemic conscious or unconscious bias against members of certain demographic groups, based on ethnicity, gender, or age, for example (see, e.g., Sen 2014a, 2014b). First, I note that this potential issue exists for other existing measures of ideology such as JCS and CBI, albeit in different ways.[10] Moreover, recent empirical evidence suggests that demographic-driven

[8] In fact, the ability to quantify uncertainty can reduce bias. Where values are estimated with uncertainty, treating the point estimates as precisely determined predictors in a model, rather than points in a distribution, creates measurement error. One method to address this problem is to run simulations in which points are drawn randomly from the distribution.

[9] For instance, the Segal and Cover scores and the Varieties of Democracy (V-DEM) initiative (Coppedge et al. 2021) use different sets of evaluators for different observations. The V-DEM initiative evaluates democracy and related traits in over 200 countries, with each country's scores generated by several of more than 3,700 country-specific experts (Coppedge et al. 2021).

[10] For JCS, for instance, a liberal senator might make stereotypical assumptions about the liberal judicial ideology of a potential nominee (who, often would not have substantial existing judicial

bias against judges is less of a concern than some may fear. In a series of conjoint experiments, Ono and Zilis (2022) find that, when asked to evaluate the degree of bias they expect judges with different profiles to exhibit in immigration and abortion cases, the aggregated subject pool either does not distinguish between either men and women judges, or between Black and non-Black judges, or else expects that women judges and Black judges are more likely to be unbiased. And to the extent people nonetheless make snap judgements when they have access only to a judge's superficial characteristics, studies in social psychology (e.g., Pettigrew and Tropp 2013) have repeatedly shown that bias is mitigated or eliminated as the evaluator interacts with the subject, as with the *Almanac*'s expert evaluators.

Finally, it is likely that a form of ideological drift occurs, in which what is considered, say, "moderate," is different in 1990 and 2010. This challenge exists for essentially all attempts to measure ideology over time.[11] For instance, the NOMINATE-dervied JCS scores are determined relative to the issues facing Congress, such that the meaning of a moderate judge depends on what constitutes centrist behavior in Congress during the given period. Likewise, the JuDJIS scores must be interpreted as estimating a judge's position relative to the ideology norms of that period. In using these scores – as with other ideology measures – researchers should be cautious in interpreting changes over long periods. Over short periods, as within a given administration, change can be interpreted with a relatively high degree of confidence.

3. Method Validation

To determine how well the qualitative evaluations capture the essence of judicial ideology, I compare the JuDJIS *Circuit Ideology* with scores from established and recently developed datasets of ideology produced using different methods. Based on three different validation metrics using four different data sets, I find that, in predicting case outcomes, the JuDJIS *Circuit Ideology* scores outperform by significant margins the three existing circuit ideology measures.

3.1 Data Comparison

First, Figure 3 is a matrix of pairwise scatterplots, comparing the JuDJIS Circuit Ideology scores (over each judge's full tenure) with the scores for those same judges

experience) based on the candidate's (Black, Latino, or Indigenous, e.g.) race or (female) sex. Somewhat similarly for CBI, in hiring clerks, judges might make comparable assumptions about their potential clerks' ideology based partly on the clerks' race, gender, or expressed sexual orientation (cf., e.g., Vick and Cunningham 2018). (While this last phenomenon would involve bias *by* judges instead of bias *toward* judges, it could nonetheless bias the scores similarly.)

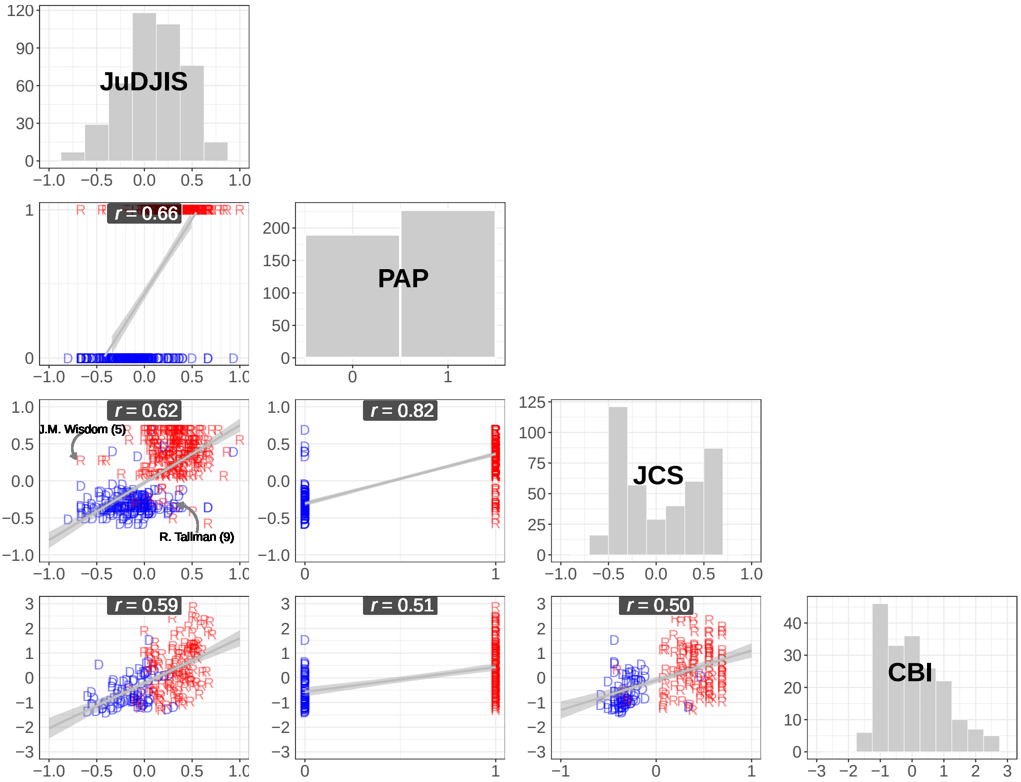
[11] Bailey (2007) develops a bridging solution that connects the Supreme Court, Congress, and the President over time, but that method is not available here.

(where available) based on: the Party-of-the-Appointing-President (PAP); the Judicial Common Space (JCS); and the Clerkship-Based Ideology (CBI) scores. The matrix also compares each of the scores with every other one. A least-squares regression of the *y*-axis scores on the *x*-axis scores is indicated by a gray line, with 90% confidence intervals indicated by the lighter gray band. The *r* values denote the correlation coefficients. The matrix diagonal (from top-left to bottom-right) of Figure 3 is a set of histograms indicating the respective distributions of each of the dataset values.

Reviewing the matrix of scatterplots and histograms, a few things are apparent. First, the datasets have notably different distributions. The JCS histogram shows that JCS is bimodal on a -1 to 1 scale, with the majority of scores falling either in the center-left range (-.6 to -.4) or center-right range (0.4 to 0.6). This distribution is not surprising given JCS's design, which assigns judges' ideological values based primarily on Senate voting records; for all of the covered period, Congress has been highly polarized by party, albeit to different degrees (Poole and Rosenthal 1985, 2000; Barber et al. 2015). In contrast, both JuDJIS and CBI are unimodal, with CBI skewed right (many more liberal judges than conservative, and a long right tail). Given that lawyers as a group, and especially recent law grads (who are most like to serve as clerks), are more liberal than society generally (Bonica, Chilton, and Sen 2016), this distribution is also unsurprising given CBI's methodology. Finally, the distribution of JuDJIS data appears to be quite close to normal.

Turning to the scatterplots, JuDJIS is positively correlated with each of the three other ideology measures at statistically significant levels. The substantive correlations are each moderately high, with Pearson correlation coefficients of $r = 0.662$, $\sigma = 0.037$ (vs. PAP), $r = 0.624$, $\sigma = 0.038$ (vs. JCS), and $r = 0.589$, $\sigma = 0.040$ (vs. CBI). Again, these moderate levels of correlation are unsurprising, given (as discussed in the Appendix) the measures' different implicit assumptions about the nature of ideology and the strategies for measuring it. Such moderate correlation levels imply that, while the four measures may all be attempting to measure the same general underlying concept, it is plausible that a study's choice of ideology measure might sometimes affect the results (cf. Cope, Crabtree, and Fariss 2020).

Another trait the scatterplots reveal is the degree to which the different measures can distinguish between individual judges. To different degrees, both PAP and JCS place judges in noticeable silos, in which many judges share the same ideology score. That this would occur for PAP is self-evident, as there are only two parties that have nominated judges. JCS's silos are far more numerous, and they group fewer judges together. But they result from a similar phenomenon: a given political actor's ideology, sometimes that of the president, is attributed to all judges whose appointment for which the actor is responsible. Both CBI and JuDJIS feature very few judges with the same scores, in part because each is based on at least several-



Note: Observations denote judges (averaged over each judge's Court of Appeals tenure) 1990–2017; red 'R' = Republican-appointed; blue 'D' = Democratic-appointed. Top-left *r* values indicate correlation coefficient of two score sets

Figure 3. Matrix of pairwise scatterplots: JuDJIS vs. PAP vs. JCS vs. CBI measures

dozen individual campaign contributions (for CBI) or evaluations (for JuDJIS). As a result, these scores are more sensitive to small differences between any two judges' latent ideologies.

Examining a few of the judges lying outside the diagonal also illustrates some key differences between JuDJIS and other methods. For instance, consider the two labeled judges in the JuDJIS-JCS scatterplot in Figure 3, the Fifth Circuit's John Minor Wisdom (1957–99) and the Ninth Circuit's Richard Tallman (2000-present). Judge Wisdom was an Eisenhower appointee; a liberal Southern Republican from New Orleans, Judge Wisdom was one of the "Fifth Circuit Four," a group of judges who significantly expanded civil rights for African-Americans during the 1950s and 60s in the face of strong, sometimes violent, local White opposition (Grinstein 2020). Taking senior status in 1977, Wisdom was considered among the most progressive judges in the country until his death in 1999. Drawing on the evaluations of the practicing bar, JuDJIS rates Wisdom among the 5% most liberal judges in the data set. JCS, considering the politics of his appointers, rates his as a moderate conservative. Conversely, 1999 Clinton-appointee Judge Richard Tallman was a Republican lawyer who was personally recommended by a prominent conservative Washington state supreme court justice/former Ninth Circuit nominee. Clinton agreed to nominate Tallman as part of a political deal in which Washington's Republican senator agreed to unblock three of Clinton's preferred nominees (Slotnick 2006). JuDJIS, in part reflecting Tallman's 69% conservative voting record in en banc cases, rates him a moderate conservative. JCS, based on the congressional voting record of Washington's *Democratic* (Clinton's party) senator, considers him a moderate liberal.

3.2 Predicting Case Outcomes

Although they produce several interesting insights, gauging validity by observing how well measures correlate takes us only so far. A better test of a measure's value is how well it predicts behavior (Cope 2024), that is, its *predictive validity*. I therefore proceed to determine how well the JuDJIS *Circuit Ideology* data predict case outcomes relative to how well the three major existing circuit-judge ideology measures, respectively, predict those outcomes. To do so, I draw on a new dataset of en banc decisions, comprising all such decisions during the relevant period (1990-2017), covering all numbered circuits and the D.C. Circuit. By way of background, a federal circuit court hears a small number of cases *en banc*, in which the whole court typically reviews a decision of a circuit panel. The cases tend to be contentious and are more likely to feature issues implicating traditional ideological cleavages. I validate the data using these cases because they span all circuits and relevant years, and they contain a greater proportion of "harder" cases, i.e, those in which a judge's ideology is more likely to be salient. Note, however, that "hard" is not equivalent to "ideological" in the traditional, political sense. Not all of these cases involve legal issues that traditionally divide liberals and conservatives; many of the en banc courts split over issues with less traditional ideological salience, such as technical or procedural legal questions. On average, however, we should expect stronger relationships between ideology scores and case outcomes for these hard cases as a whole.

Each of the 414 decisions, including dissents and any concurrences, was read, and each judge's vote was classified as either conservative or liberal. Using these data, I conduct three tests: (1) goodness of fit; (2) a logit regression, comparing the respective normalized correlation coefficients; and (3) a receiver operating characteristic (ROC) curve, comparing the respective measures' areas under the curve.

Predictive Validation Using En Banc Votes: Goodness of Fit

I first explore the goodness-of-fit between votes and judge ideology score for each of the four measures in turn. The Pearson's product-moment correlation coefficients are: JuDJIS: r = 0.403, 95% CI [0.371, 0.434]; PAP: r = 0.351, 95% CI [0.319, 0.382]; JCS: r = 0.323, 95% CI [0.290, 0.354]; and CBI: r = 0.264, 95% CI [0.215, 0.311]. Thus, a judge's JuDJIS score explains significantly more variation in these data's judge votes than the equivalent scores of the three existing measures do.

Predictive Validation Using En Banc Votes: Logit Regression

I next estimate a logit model, regressing votes (whether the judge voted for a conservative outcome) on the predictor (the ideology score of the judge, as respectively estimated by JuDJIS, Party-of-the-Appointing President, JCS, and CBI). Table 1 and Figure 4 present the results. The coefficients indicate the marginal effects of a two-standard-deviation increase in conservativeness in each respective ideology measure on the probability of a conservative vote. (The overall incidence of conservative votes in the full data set is approximately 55.9%) Thus, for example, a judge with a JuDJIS ideology score of 0.41 is about 45.4% percentage points more likely to lodge a conservative vote than a judge with a JuDJIS ideology score of -0.14. Each of the four scores are associated with votes at highly significant levels. But change in JuDJIS score is associated with a greater change in probability of a conservative vote than the equivalent change for the other three scores.

Table 1. Logit predictions: Marginal effects of judge score on probability of casting a conservative en banc vote

	Probability of a judge's casting a conservative en banc vote			
JuDJIS Score	0.454 (5.80×10^{-87})			
PAP Score		0.349 (5.58×10^{-93})		
JCS Score			0.335 (2.63×10^{-66})	
CBI Score				0.284 (2.28×10^{-22})
Num. obs.	2,745	3,016	2,999	1,445

Note: Coefficients are normalized to indicate the change in probability associated with a two standard-deviation change in the given score. p scores are in parentheses.

Figure 4 graphically illustrates this difference. The four graphs plot, for each vote, the judge's ideology score on the x-axis. The judge's votes are plotted on the y-axis, with conservative votes (1) at the top and liberal ones (0) at the bottom. (They are vertically jittered to show density.) For each graph, a logit regression curve shows the relationship between the two variables. Though the correlations are positive and significant in all four cases, the difference in the relationships' magnitude is evident from the shapes of the respective s-curves.

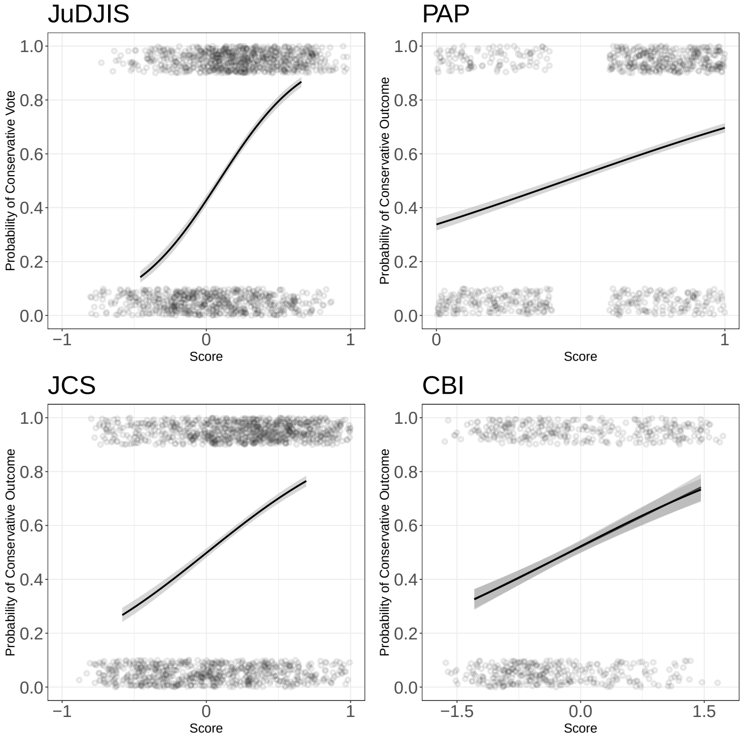


Figure 4. Probability of a conservative vote as a function of judge ideology score, by measure: en banc cases

Predictive Validation Using En Banc Votes: ROC Analysis

To further gauge predictive validity, I next estimate a logistic regression and plot a receiver operating characteristic (ROC) curve for each of the four measures. Long common in the diagnostic medicine literature and now often used in political research (Mueller and Rauh 2018; Imai and Khanna 2016), an ROC curve is an arguably more-intuitive method for assessing how well a metric accurately classifies observations into binary outcomes. In this case, the predicted variable is whether the judge casts a conservative vote, as described above. The predicting variable is the ideology score of the judge, as respectively estimated by JuDJIS, Party-of-the-Appointing President, JCS, and CBI.[12] For every judge ideology threshold, the ROC curve plots the true positive rate against the false positive rate (Wang 2019; Fischman 2011). The Area Under the Curve (AUC) therefore represents

[12] For dynamic scores, a judge's score is averaged over their tenure.

each measure's relative success at predicting votes (see Hanley and McNeil 1982). Specifically, it denotes the probability that the given measure will rank a randomly chosen conservative vote as conservative instead of liberal.

The top panel of Figure 5 displays the ROC curves. The bottom panel gives the AUC values for each of the four measures. JuDJIS achieves an AUC value of 0.734; JCS's AUC value is 0.675; PAP's AUC value is 0.672; and CBI's AUC value is 0.645. A DeLong test indicates that PAP is statistically indistinguishable from JCS ($p = 0.793$) and marginally significantly higher than CBI ($p = 0.115$). But JuDJIS performs significantly better than all three: since JCS is about 17.5 percentage points above a random classification and JuDJIS is 23.4 percentage points above random, JuDJIS's performance represents an improvement of 33.7% over JCS's performance. A DeLong test indicates that the difference between the two is highly significant ($p = 2.06 \times 10^{-5}$).

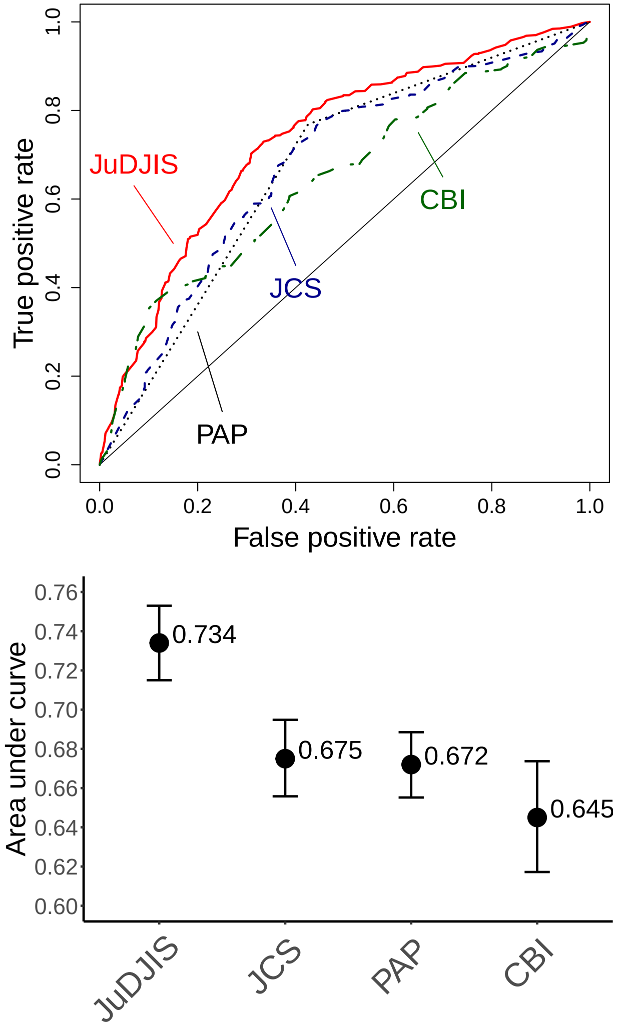


Figure 5. Top: ROC curves comparing four measures' success at predicting en banc votes; Bottom: Comparative areas under the ROC curve

Robustness Analysis

Finally, to show these results' robustness to different data and model specifications, I perform comparable analyses of a larger data set comprising three-judge panel decisions ($n = 4, 482$). While these cases disproportionately constitute "easy" cases, in which the members of the panel are unanimous over 95% of the time, I include them because they better reflect the run-of-the-mill decision-making of circuit judges. The results are substantially similar to those produced by the en banc cases, JuDJIS performing better – albeit not quite as decisively – than the other three measures in each case. (See Appendix Figure A6.1 for analysis and results).

Thus, for 21 head-to-head predictive comparisons across different data forms and model specifications, the JuDJIS *Circuit Ideology* scores predict outcomes more accurately than the other measure in all 21 of them, with the difference statistically signficant in 20 of the 21. Together, this set of validations indicates that the JuDJIS *Circuit Ideology* data outperform all existing measures of circuit ideology – using a variety of metrics – in predicting how judges decide cases. And although JuDJIS's accuracy in reflecting change over time cannot be tested against other measures (because no others are currently dynamic), we would expect that this dynamism confers JuDJIS with additional advantage in predictive power and general validity.

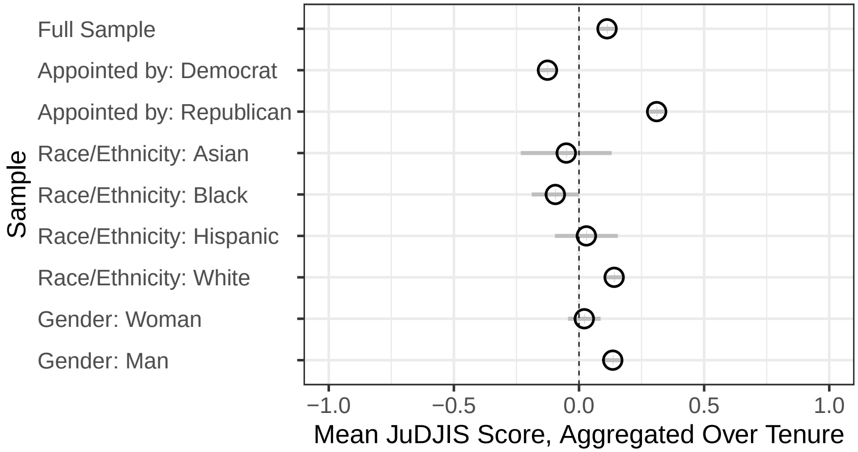
4. Analyses of Federal Circuit Judge Ideology

4.1 Descriptive Statistics

In what follows, I present the *Circuit Ideology* scores in more detail.[13] To produce the *Circuit Ideology* scores, I applied the process described above for the ideology category for all active judges on the U.S. courts of appeals who served in any year between 1990–2017. Figure 6 displays summary statistics for the data set, aggregated, and disaggregated, by party of appointing president and by the judge's gender.

At the judge level (i.e., a judge's scores aggregated over full tenure), the mean ideology score is 0.11, the median score is 0.10, and the standard deviation is 0.33. Thus, the dataset exhibits a clear conservative slant. Figure 6 shows that this appears to stem from two phenomena: (1) Republican presidents have appointed most (54.7%) of the judges in the dataset; and (2) Democratic-appointed judges are more moderate, i.e., the average ideology score of Republican-appointed judges (.31) is more conservative than the average ideology score of Democrat-nominated judges (-.13) is liberal.

[13] Section 6.4 of the appendix provides descriptive statistics of a sample of the JuDJIS *District Ideology* data.



Note: Circles denote means for the given sample; gray bars denote 95% confidence intervals.

Figure 6. JuDJIS Circuit Ideology Descriptive Statistics

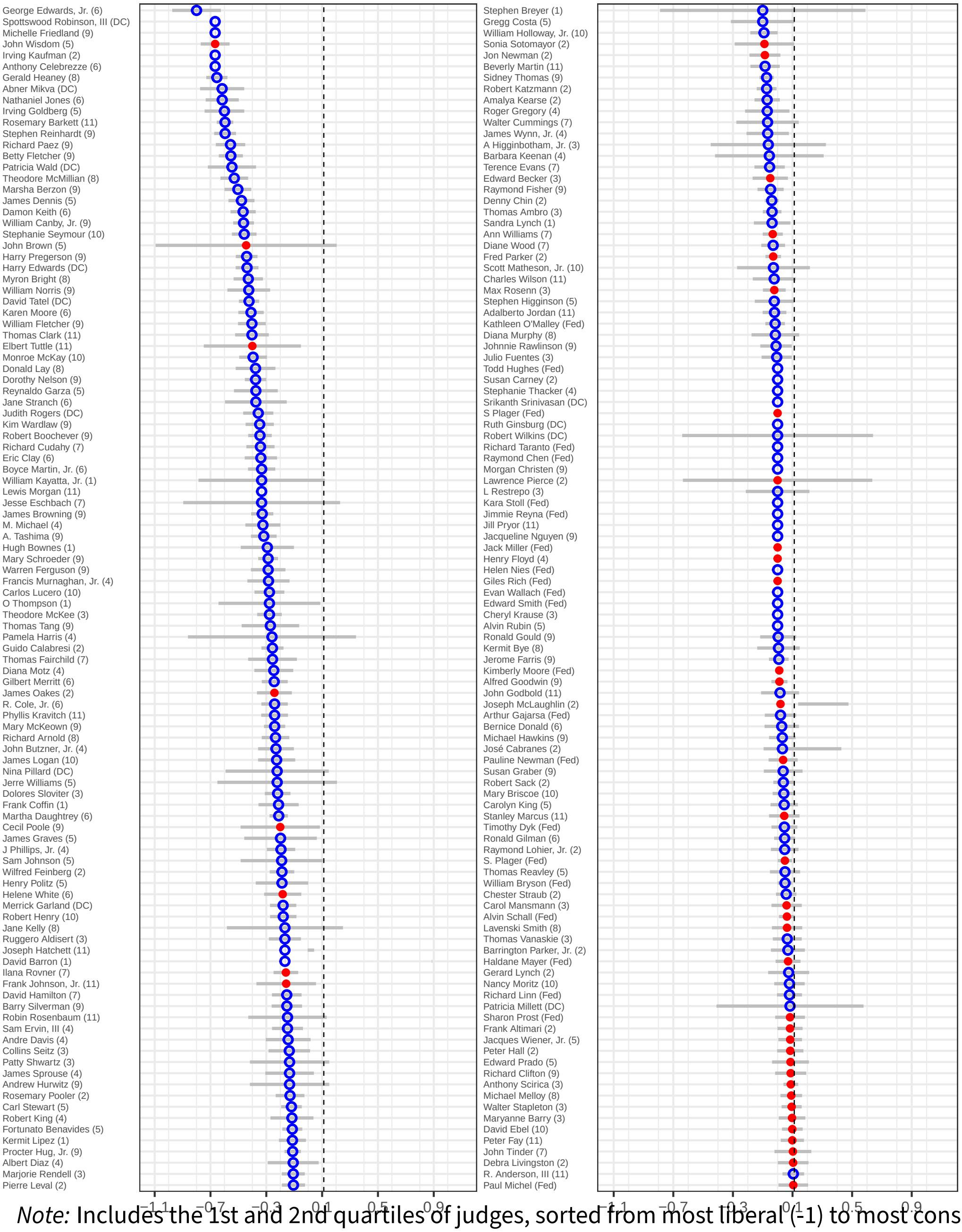
To further illustrate the variation between judges, Figures 7 and 8 provide the judicial ideology point estimate and 90% confidence intervals for all 418 judges in the JuDJIS *Circuit Ideology* data set, averaged over each judge's Court of Appeals tenure. The judges are ordered from most liberal (top-left of Figure 7) to most conservative (bottom-right of Figure 8). As the figures show, the confidence intervals vary considerably between judges. As explained in the methods section above, confidence in the ideology estimate is a function of two factors: the number of total comments the judge received over his or her tenure and the uniformity of those comments.[14] Thus, the judges with particularly large confidence intervals are almost uniformly those with just one evaluation in the data set because they had a very short circuit tenure, left the bench shortly after 1990, or joined the bench shortly before 2017. (For this last group, uncertainty about the estimate will likely decrease as evaluations from 2018-on are incorporated into the data set.) As the figures show, beginning with Judge George Edwards's (6th Cir., 1963-1995) score of -2.40, the 419 ideology scores rise incrementally, ending with Judge George MacKinnon's (D.C. Cir., 1969–95) score of 3.00.

Seven eventual Supreme Court nominees are included in the data set based on their circuit court tenures.[15] One particularly notable score is the '0' assigned to Ruth Bader Ginsburg, who, before her 1993 elevation to the Supreme Court, served thirteen years on the D.C. Circuit on appointment by President Carter. Though such a moderate score may surprise some who know the late judge/justice as a liberal

[14] In the rare cases where all comment scores are uniform (usually, for judges with very short tenures over the covered period), the standard error and confidence intervals are undefined and therefore missing.

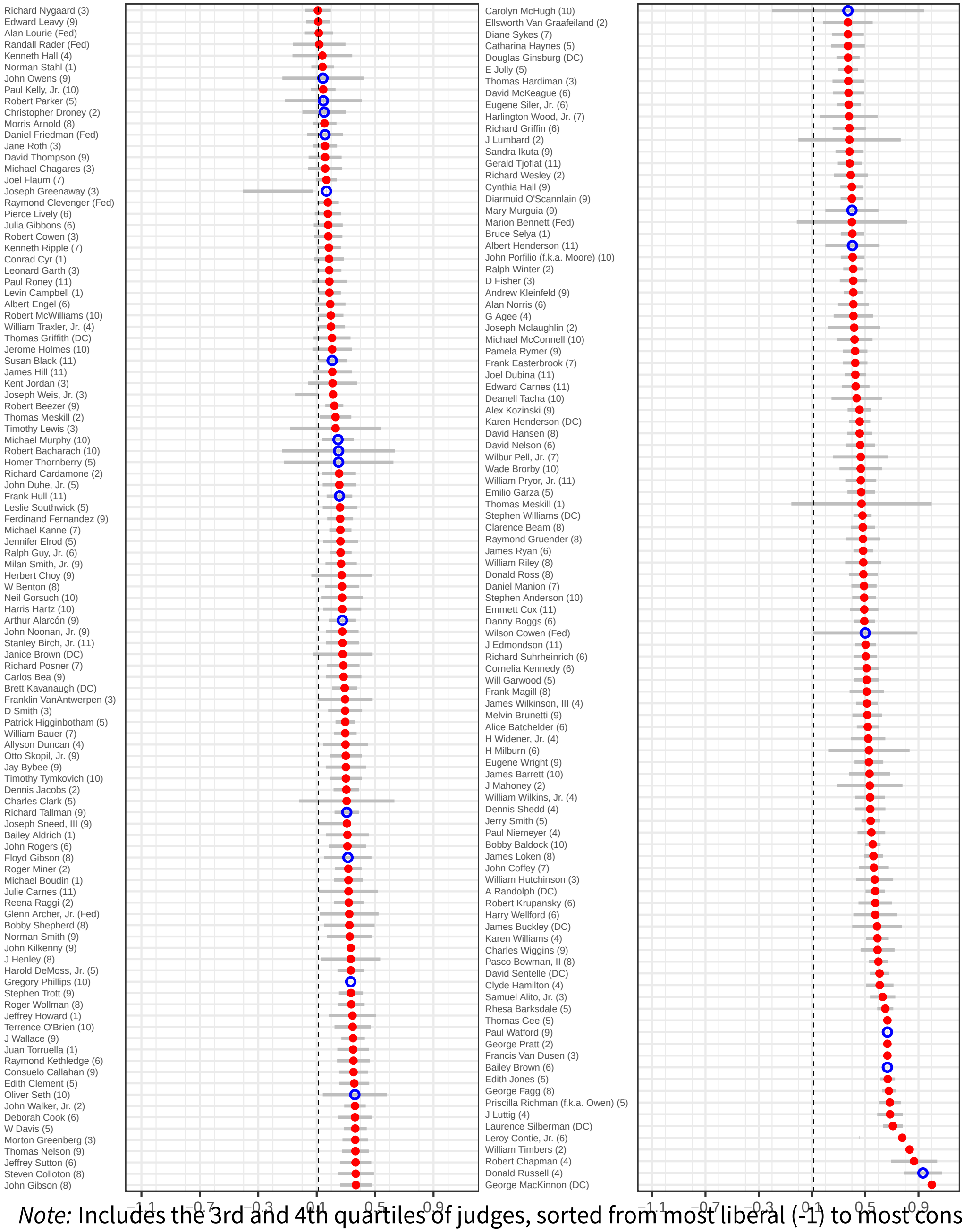
[15] See Figure A6.1 in the appendix for a comparison of those seven judges.

champion of women's and other civil rights, her tenure as a circuit judge was, in fact, considered positively centrist. Indeed, according to a 1987 empirical study of the court, *Judge* Ginsburg was more likely to vote with Republican than Democratic appointees and generally opposed expanding corporate regulation (Lepore 2018). According to a 2018 biography in *The New Yorker*, UC Santa Barbara history professor Sherron De Hart described Ginsburg's D.C. Circuit tenure as "something like a decontamination chamber," in which Ginsburg was "rinsed and scrubbed of the hazard of her thirteen years as an advocate for women's rights." By 1993, the article observed, Ginsburg had been "sufficiently depolarized" for nomination to the high court (Lepore 2018).



Note: Includes the 1st and 2nd quartiles of judges, sorted from most liberal (-1) to most conservative (+1) by average ideology score, averaged over the judge's Court of Appeals tenure. Blue open circles denote Democratic president appointees; red closed circles denote Republican president appointees.

Figure 7. *JuDJIS* U.S. Circuit Judge Ideologies, 1990-2017 (Tiers 1 and 2: #1-211)



Note: Includes the 3rd and 4th quartiles of judges, sorted from most liberal (-1) to most conservative (+1) by average ideology score, averaged over the judge's Court of Appeals tenure. Blue open circles denote Democratic president appointees; red closed circles denote Republican president appointees.

Figure 8. *JuDJIS* U.S. Circuit Judge Ideologies, 1990-2017 (Tiers 3 and 4: #212-419)

5. Conclusion

This Article developed and introduced the Jurist-Derived Judicial Ideology Scores, the first dynamic method for systematically estimating the ideologies and other traits of nearly the entire federal judiciary. Derived from tens of thousands of qualitative evaluations, it can potentially locate on a multi-dimensional scale nearly every Article III U.S. federal judge serving since 1990. Not surprisingly given the quality of the content underlying the scores, JuDJIS ideology data predict case outcomes with significantly greater accuracy than any of the three leading circuit-judge ideology measures.

The analysis above suggests that expert crowds' observation of judging is a valid method for measuring ideology. It validates the assumption that legal practitioners and other experts have special insight into how judges decide cases, insight that cannot be captured as successfully by political phenomena such as the judicial-appointment process and judges' own *political* behavior.

I hope that JuDJIS's four non-ideology measures, to be introduced in future work, will further demonstrate empirically the multi-dimensional character of the judging process. In addition to shedding light on important questions themselves, I hope that other findings like these will help to further close the theoretical and methodological gaps that still divide scholars studying how judges make decisions.

Finally, the hierarchical ngram method introduced here might be applied to estimate ideology or judgment in other contexts, such as country human rights reports and public-officials' statements. Thus, the method might eventually aid measurement and text-analysis research in several other political research domains.

References

Albaugh, Quinn, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. The automated coding of policy agendas: a dictionary-based approach. In *6th annual comparative agendas conference, atnwerp, beligum.*

Albaugh, Quinn, Stuart Soroka, Jeroen Joly, Peter Loewen, Julie Sevenans, and Stefaan Walgrave. 2014. Comparing and combining machine learning and dictionary-based approaches to topic coding. In *Th annual comparative agendas project (cap) conference, konstanz, germany.*

Bailey, Michael A. 2007. Comparable preference estimates across time and institutions for the court, congress, and presidency. *American Journal of Political Science* 51 (3): 433–448.

———. 2017. Measuring ideology on the courts. In *Routledge handbook of judicial behavior,* 62–83. Routledge.

Barber, Michael, Nolan McCarty, Jane Mansbridge, and Cathie Jo Martin. 2015. Causes and consequences of polarization. *Political negotiation: A handbook* 37:39–43.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review* 110 (2): 278–295.

Bonica, Adam, Adam S Chilton, Jacob Goldin, Kyle Rozema, and Maya Sen. 2017. Measuring judicial ideology using law clerk hiring. *American Law and Economics Review* 19 (1): 129–161.

Bonica, Adam, Adam S Chilton, and Maya Sen. 2016. The political ideologies of american lawyers. *Journal of Legal Analysis* 8 (2): 277–335.

Bonica, Adam, and Maya Sen. 2017. A common-space scaling of the american judiciary and legal profession. *Political Analysis* 25 (1): 114–121.

———. 2021. Estimating judicial ideology. *Journal of Economic Perspectives* 35 (1): 97–118.

Boyd, Christina L. 2011. Federal district court judge ideology data. *University of Georgia.*

Converse, Philip E. 2006. The nature of belief systems in mass publics (1964). *Critical review* 18 (1-3): 1–74.

Cope, Kevin L. 2024. The oxford handbook of comparative judicial behaviour. Chap. The Conceptual Challenge to Measuring Ideology, edited by Lee Epstein et al. Oxford University Press.

Cope, Kevin L, Charles Crabtree, and Christopher J Fariss. 2020. Patterns of disagreement in indicators of state repression. *Political Science Research and Methods* 8 (1): 178–187.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M Steven Fish, Lisa Gastaldi, et al. 2021. V-dem codebook v11.

Epstein, Lee, Andrew D Martin, and Kevin Quinn. 2024. Measuring political preferences. *The Oxford Handbook of Comparative Judicial Behaviour,* XXX–XX.

Epstein, Lee, Andrew D Martin, Jeffrey A Segal, and Chad Westerland. 2007. The judicial common space. *The Journal of Law, Economics, & Organization* 23 (2): 303–325.

Epstein, Lee, Thomas G Walker, and William J Dixon. 1989. The supreme court and criminal justice disputes: a neo-institutional perspective. *American Journal of Political Science,* 825–841.

Farah, Hassan Abdirahman, and Arzu Gorgulu Kakisim. 2023. Enhancing lexicon based sentiment analysis using n-gram approach. In *Smart applications with advanced machine learning and human-centred problem design,* 213–221. Springer.

Fischman, Joshua B. 2011. Estimating preferences of circuit judges: a model of consensus voting. *The Journal of Law and Economics* 54 (4): 781–809.

Fischman, Joshua B, and David S Law. 2009. What is judicial ideology, and how should we measure it. *Wash. UJL & Pol'y* 29:133.

Gaudet, Harris. 1933. St. john, individual differences in the sentencing tendencies of judges, 23 j. *AM. INST. CRIM. L. & CRIMINOLOGY* 811.

Gerring, John. 1997. Ideology: a definitional analysis. *Political Research Quarterly* 50 (4): 957–994.

Giles, Micheal W, Virginia A Hettinger, and Todd Peppers. 2001. Picking federal judges: a note on policy and partisan selection agendas. *Political Research Quarterly* 54 (3): 623–641.

Grendstad, Gunnar, William R Shaffer, and Eric N Waltenburg. 2012. Ideologi og grunnholdninger hos dommerne i norges høyesterett. *Lov og rett* 51 (4): 240–253.

Grimmer, Justin, and Brandon M Stewart. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21 (3): 267–297.

Grinstein, Max. 2020. The fifth circuit four. *The History Teacher* 54 (1): 155–179.

Hanley, James A, and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143 (1): 29–36.

Harcourt, Bernard E. 2007. Judge richard posner on civil liberties: pragmatic authoritarian libertarian. *U. Chi. L. Rev.* 74:1723.

Imai, Kosuke, and Kabir Khanna. 2016. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis* 24 (2): 263–272.

Lammon, Bryan D. 2009. What we talk about when we talk about ideology: judicial politics scholarship and naive legal realism. *. John's L. Rev.* 83:231.

Landis, J Richard, and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics,* 159–174.

Lepore, Jill. 2018. Ruth bader ginsburg's unlikely path to the supreme court. *The New Yorker* 1.

Martin, Andrew D, and Kevin M Quinn. 2002. Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political Analysis* 10 (2): 134–153.

McMillion, Barry J. 2017. *The blue slip process for us circuit and district court nominations: frequently asked questions.* Congressional Research Service Washington, DC.

Mueller, Hannes, and Christopher Rauh. 2018. Reading between the lines: prediction of political violence using newspaper text. *American Political Science Review* 112 (2): 358–375.

Nagel, Stuart S. 1961. Political party affiliation and judges' decisions. *American Political Science Review* 55 (4): 843–850.

Ono, Yoshikuni, and Michael A Zilis. 2022. Ascriptive characteristics and perceptions of impropriety in the rule of law: race, gender, and public assessments of whether judges can be impartial. *American journal of political science* 66 (1): 43–58.

Pettigrew, Thomas F, and Linda R Tropp. 2013. Does intergroup contact reduce prejudice? recent meta-analytic findings. In *Reducing prejudice and discrimination,* 93–114. Psychology Press.

Poole, Keith T, and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American journal of political science,* 357–384.

———. 2000. *Congress: a political-economic history of roll call voting.* Oxford University Press on Demand.

Rohde, David W, and Harold J Spaeth. 1976. *Supreme court decision making.* WH Freeman.

Schubert, Glendon A. 1960. *Quantitative analysis of judicial behavior.* Free Press.

Segal, Jeffrey A, and Albert D Cover. 1989. Ideological values and the votes of us supreme court justices. *American Political Science Review* 83 (2): 557–565.

Sen, Maya. 2014a. How judicial qualification ratings may disadvantage minority and female candidates. *Journal of Law and Courts* 2 (1): 33–65.

———. 2014b. Minority judicial candidates have changed: the aba ratings gap has not. *Judicature* 98:46.

Slotnick, Elliot E. 2006. Appellate judicial selection during the bush administration: business as usual or a nuclear winter. *Ariz. L. Rev.* 48:225.

Spaeth, Harold, Lee Epstein, Ted Ruger, Keith Whittington, Jeffrey Segal, and Andrew D Martin. 2014. Supreme court database code book. *URL: http://scdb. wustl. edu.*

Vick, Astin D, and George Cunningham. 2018. Bias against latina and african american women job applicants: a field experiment. *Sport, Business and Management: An International Journal* 8 (4): 410–430.

Voeten, Erik. 2007. The politics of international judicial appointments: evidence from the european court of human rights. *International Organization* 61 (4): 669–701.

Wang, Yu. 2019. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data: a comment. *Political Analysis* 27 (1): 107–110.

Wijtvliet, Wessel, and Arthur Dyevre. 2021. Judicial ideology in economic cases: evidence from the general court of the european union. *European Union Politics* 22 (1): 25–45.

Windett, Jason H, Jeffrey J Harden, and Matthew EK Hall. 2015. Estimating dynamic ideal points for state supreme courts. *Political Analysis* 23 (3): 461–469.

– ARTICLE TYPE

An Expert-Sourced Measure of Judicial Ideology

Kevin L. Cope

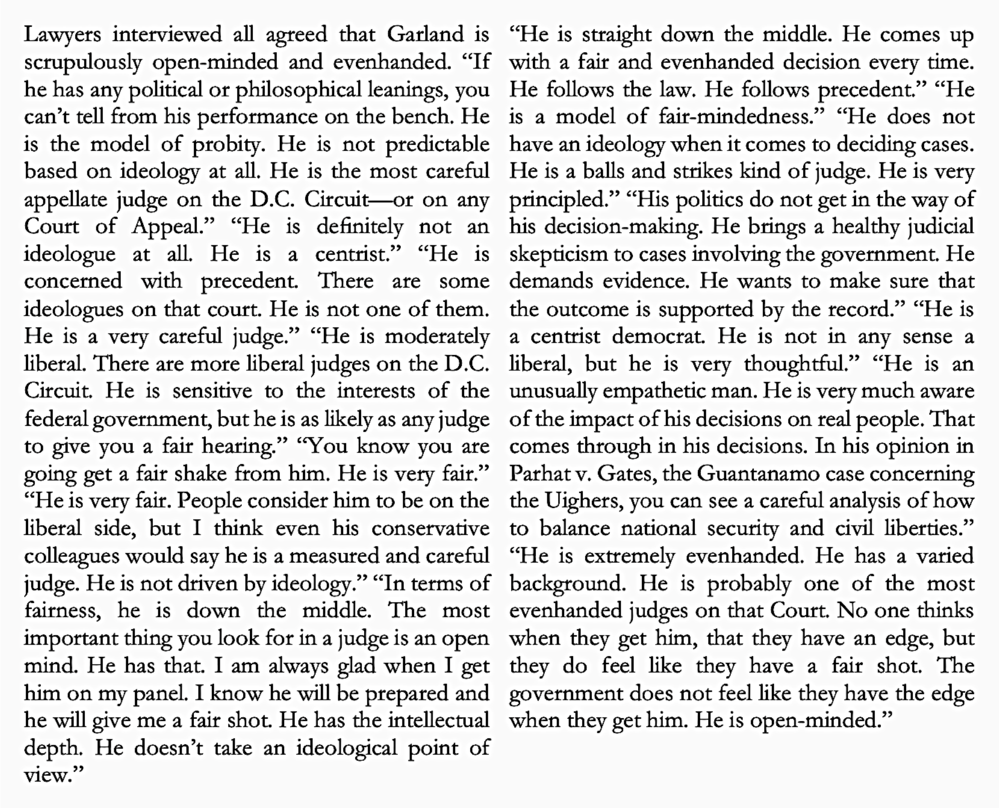
Associate Professor of Law and Public Policy, University of Virginia, 580 Massie Rd., Charlottesville, Virginia 22903. Email: kcope@law.virginia.edu

Appendix

Contents

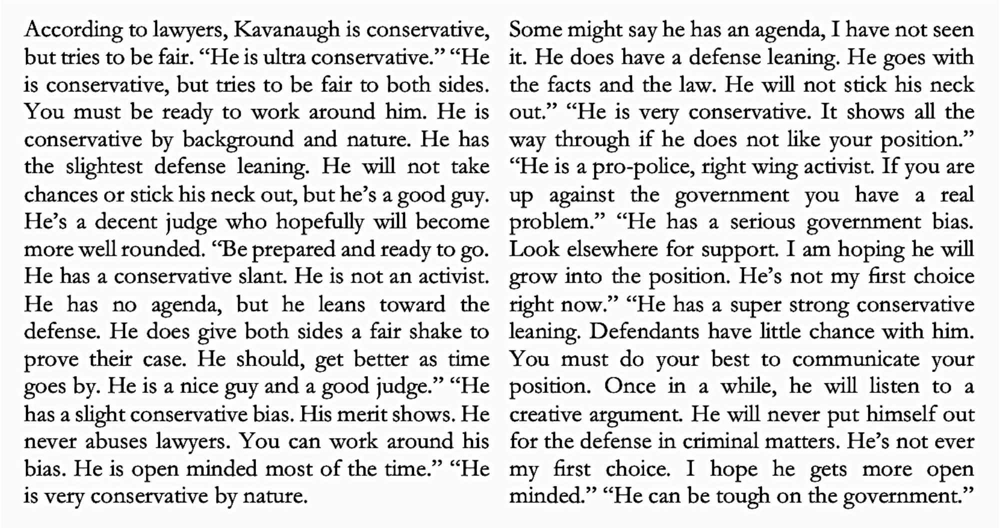
1. Almanac example evaluations

To illustrate some typical evaluations, Figures A1.1 and A1.2 are excerpts from two ideology reviews for 2009, Merrick Garland and Brett Kavanaugh, both then judges on the D.C. Circuit. Individual evaluators are delineated by quotation marks. As shown, each evaluator uses his or her own words to describe the judge ideology, meaning that some similar words and phrases are used frequently, and other words and phrases appear rarely. For Garland, despite being a Democratic (Clinton) appointee, most evaluators characterize him as open-minded, middle-of-the-road, or moderate, with just a few characterizing him as left-of-center. Indeed, though he became somewhat of a liberal cause célèbre after Senate Republicans ignored Garland's 2016 nomination to the Supreme Court, some at the time considered Garland so moderate that Republican Senate Majority Leader Mitch McConnell might feel compelled to act on the nomination (Bellin 2016). For Kavanaugh, on the other hand, there is somewhat less consensus, but a common theme of reliable conservatism emerges.



Lawyers interviewed all agreed that Garland is scrupulously open-minded and evenhanded. "If he has any political or philosophical leanings, you can't tell from his performance on the bench. He is the model of probity. He is not predictable based on ideology at all. He is the most careful appellate judge on the D.C. Circuit—or on any Court of Appeal." "He is definitely not an ideologue at all. He is a centrist." "He is concerned with precedent. There are some ideologues on that court. He is not one of them. He is a very careful judge." "He is moderately liberal. There are more liberal judges on the D.C. Circuit. He is sensitive to the interests of the federal government, but he is as likely as any judge to give you a fair hearing." "You know you are going get a fair shake from him. He is very fair." "He is very fair. People consider him to be on the liberal side, but I think even his conservative colleagues would say he is a measured and careful judge. He is not driven by ideology." "In terms of fairness, he is down the middle. The most important thing you look for in a judge is an open mind. He has that. I am always glad when I get him on my panel. I know he will be prepared and he will give me a fair shot. He has the intellectual depth. He doesn't take an ideological point of view." "He is straight down the middle. He comes up with a fair and evenhanded decision every time. He follows the law. He follows precedent." "He is a model of fair-mindedness." "He does not have an ideology when it comes to deciding cases. He is a balls and strikes kind of judge. He is very principled." "His politics do not get in the way of his decision-making. He brings a healthy judicial skepticism to cases involving the government. He demands evidence. He wants to make sure that the outcome is supported by the record." "He is a centrist democrat. He is not in any sense a liberal, but he is very thoughtful." "He is an unusually empathetic man. He is very much aware of the impact of his decisions on real people. That comes through in his decisions. In his opinion in Parhat v. Gates, the Guantanamo case concerning the Uighers, you can see a careful analysis of how to balance national security and civil liberties." "He is extremely evenhanded. He has a varied background. He is probably one of the most evenhanded judges on that Court. No one thinks when they get him, that they have an edge, but they do feel like they have a fair shot. The government does not feel like they have the edge when they get him. He is open-minded."

Figure A1.1. Sample Ideology Evaluation: Judge Merrick Garland (2009)



According to lawyers, Kavanaugh is conservative, but tries to be fair. “He is ultra conservative.” “He is conservative, but tries to be fair to both sides. You must be ready to work around him. He is conservative by background and nature. He has the slightest defense leaning. He will not take chances or stick his neck out, but he’s a good guy. He’s a decent judge who hopefully will become more well rounded. “Be prepared and ready to go. He has a conservative slant. He is not an activist. He has no agenda, but he leans toward the defense. He does give both sides a fair shake to prove their case. He should, get better as time goes by. He is a nice guy and a good judge.” “He has a slight conservative bias. His merit shows. He never abuses lawyers. You can work around his bias. He is open minded most of the time.” “He is very conservative by nature. Some might say he has an agenda, I have not seen it. He does have a defense leaning. He goes with the facts and the law. He will not stick his neck out.” “He is very conservative. It shows all the way through if he does not like your position.” “He is a pro-police, right wing activist. If you are up against the government you have a real problem.” “He has a serious government bias. Look elsewhere for support. I am hoping he will grow into the position. He’s not my first choice right now.” “He has a super strong conservative leaning. Defendants have little chance with him. You must do your best to communicate your position. Once in a while, he will listen to a creative argument. He will never put himself out for the defense in criminal matters. He’s not ever my first choice. I hope he gets more open minded.” “He can be tough on the government.”

Figure A1.2. Sample Ideology Evaluation: Judge Brett Kavanaugh (2009)

2. The optimal text-analysis method

To derive meaning from a corpus, several methods are available, especially given recent advances in machine learning models and computational power generally. In this section, I conduct a pilot analysis of several potential methods to determine how each performs in quantifying the *Almanac* corpus along several criteria including case predictive validity, transparency, and reproducability.

The existing text-analysis literature generally recognizes two broad computational approaches to quantifying meaning from text: lexicon (dictionary)-based, and machine-learning (Xhymshiti 2020). Three possible specific application of these methods appeared most plausible for this research objective: (1) a large-language model (LLM) method using Open AI's GPT-4o; (2) a traditional bigrams method; and (3) a refinement of conventional customized dictionary ngrams that I call hierarchical ngram analysis.

Large-language model

For the LLM method, Open AI's GPT-4o was used to generate ideology scores at the evaluation level (rather than parsing its component phrases, as with hierarchical ngrams). Each call gave the prompt:

> "Below is a lawyer's evaluation of judge's ideology. Please code the statement's ideology on a scale of -3 (extremely liberal) to 3 (extremely conservative), with 0 being moderate. Your output should start with the integer rating, followed by a brief rationale of one or two sentence, starting with 'Rationale:'. The integer rating and rationale should be separated by the character $, Here is the evaluation:"

followed by a complete judge evaluation. The temperature was set at 0, minimizing inter-call variation. The LLM complied by providing the requested ratings.

Conventional Bigrams In considering lexicon-based methods, I note that there exists no off-the-shelf dictionary suitable for the *Almanac* corpus. As such, it would be necessary to develop a customized dictionary based on words and phrases in the *Almanac*. Indeed, Osnabrügge, Ash, and Morelli (2023) note that customized dictionaries have high annotation efficiency, specificity, and interpretability, relative to some other methods. That leaves open the question of how much complexity is optimal for any given corpus and research question. Although some recent studies have experimented with ngrams of up to five words (Dai et al. 2020; Dey, Jenamani, and Thakkar 2018; Farah and Kakisim 2023; Presannakumar and Mohamed 2021), unigrams and bigrams remain the standard approach for most lexicon-based tasks. While they suffer from a narrow feature space, they are more straightforward to develop. apply, and explain. I created a customized bigram dictionary to apply to the *Almanac* corpus.

Hierarchical Ngram Analysis As I explain in the Article's main text, the hierarchical ngram method attempts to combine some of the strengths of lexicon and machine-learning approaches. It features a reasonably large feature space, capturing meaning from word order and context, while retaining the transparency, explainability, and reproducability of lexicon-based models. More specific information about the method is provided in the Article's main text. I used the hierarchical ngram method to develop and apply a customized dictionary.

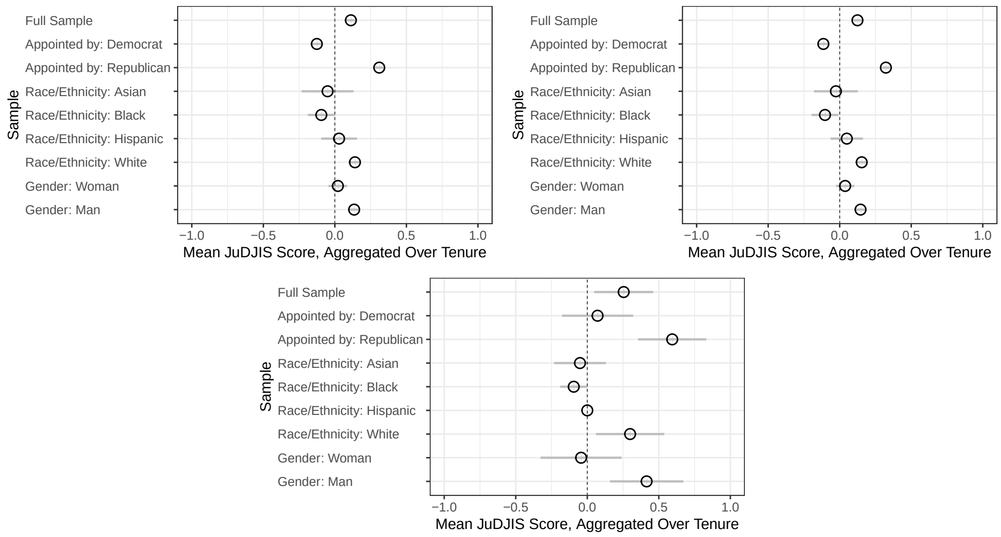


Figure A2.3. Comparison of Scores Produced by Hierarchical Ngrams (top-left); LLM (top-right); and Bigrams (bottom)

Comparison

Figure A2.3 shows descriptive statistics using the three methods. Notably, the scores produced by hierarchical ngrams and the LLM are remarkably similar. In contrast, the bigram-produced scores are different in both mean and variance. This difference certainly stems from the fact that the bigrams produce significantly fewer data points, leading to greater uncertainty.

I next assigned scores to each ideology evaluation (for a given judge in a given year) using each of the three methods. I then conducted a validation exercise using one of the same data sets – the 414 en banc cases, with judge votes as the unit of analysis – as I use to compare JuDJIS to the three existing ideology scores (party of the appointing president, JCS, and CBI).

I find that the hierarchical ngram method moderately outperforms the other two in predicting case outcomes, although the differences between the hierarchical ngram method, on one hand, and both the large-language model method and the bigram method, on the other, are not significant. See the figure below, which provides the Area Under the Curve for an ROC curve analysis of the three methods.

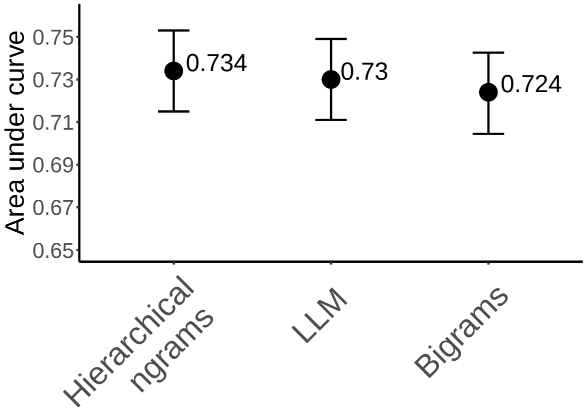


Figure A2.4. Comparison of ROC areas under the curve for three text-analysis methods

Of course, in choosing a text-analysis method, validity is one of several pertinent considerations. Resource demands, transparency, and reproducibility are also important. On those points, LLMs are generally more efficient, producing results quickly, without demanding much human capital or expertise. However, they are also often non-transparent, drawing on "black box" computations that cannot be explained to readers or other researchers (see, e.g., Albaugh et al. 2014; Albaugh et al. 2013; Grimmer and Stewart 2013). (GPT-4o, though it possesses several other advantages over other AI foundation models, received a transparency rating of just 49% in The Foundation Model Transparency Index v.1.1, placing it 11 out of 14 such models (Bommasani et al. 2024).) In contrast, lexicon-based methods can be quite transparent; each step of the construction process can be described. Partly for that reason, LLMs are less conducive to reproducibility. Even when setting "temperature" (variance) to minimum levels, model output often varies between iterations, for reasons a researcher can neither easily explain nor replicate. This is particularly concerning where, as with an ongoing dynamic ideology project, the process must be repeated over time to include new scores. It is unclear how the model's algorithm may have changed, and therefore, if new results can be validly compared with existing ones. In contrast, a researcher using a dictionary method can provide a detailed description of the process, allowing future researchers to replicate it, including with new data.

The conventional (non-hierarchical) ngram method – using only unigrams or bigrams (here, bigrams) – offers the transparency and reproducibility benefits of other dictionary-based methods, and it is more straightforward to produce. But because this conventional method analyzes less text (in this case, producing 73% as many ngrams), it is expected to omit documents that contain none of the dictionary's uni-/bigrams (c.f. Farah and Kakisim 2023). With a larger, more complex, more nuanced set of ngrams, that risk is substantially lower.

Based on these considerations, I conclude that the hierarchical ngram method is preferable to the other two for this, and perhaps other, political coding initiatives. As to its predictive validity, the hierarchical dictionary method produces scores that are probably at least as valid as – and perhaps slightly more so than – the two alternatives. And as discussed, it is plainly superior to LLMs on transparency and reproducibility. Though it is generally more resource intensive, that drawback is not nearly large enough to trump its other clear advantages. I therefore proceed to describe and use the hierarchical ngram method to produce the JuDJIS scores that I introduce here.

3. Judicial ideology coding guide

Judicial Ideology Coding Guide
Kevin Cope, Principal Investigator
Instructions:

You will be shown a very long set of phrases (several thousand phrases in total), each comprising 1 to 9 words. Each phrase has been used to describe the judicial ideology of one or more U.S. federal judges.

Some phrases contain wildcard letters or words, denoted with an asterisk ('*'). The wildcard means that one or more letters or words could take its place. If the asterisk is connected to a word, it denotes a wildcard *letter* (e.g., "leans towards the prosecut*" could include, e.g., "leans towards the prosecution" or "leans towards the prosecutor," etc.). If the asterisk is separate from any word, it denotes a wildcard *word* (e.g., "known for * conservatism" could include e.g., "known for *her* conservatism" or known for *their* conservatism," etc.). For phrases containing wildcards, we have provided one or more examples of ways the phrase could manifest. But you should code the wildcard-containing phrase (in all its possible manifestations), not the example(s) given.

For each phrase of any type, you will use your knowledge of courts and judicial ideology, as well as the definitions below, to determine the judicial ideology associated with the phrase. You will assign a score on a 7-point scale, ranging from -3 (an extremely liberal ideology) to +3 (an extremely conservative ideology).

In assigning the scores, use the definitions below of judicial *conservatism*, *liberalism*, and *centrism*. Also use the definition associated with each of the 7 possible scores. Your score should apply generally to judges at all levels of the federal judiciary (i.e., it should not apply only to judges of trial courts, circuit courts, or the Supreme Court in particular). In scoring, use only integers; do not use half scores.

Definitions:

- *Judicial Conservatism* – the belief that the primary functions of the law and the judiciary are to effectuate the intent of democratically elected authorities and constitutional drafters, to settle private disputes, and to allow people and organizations the freedom to pursue their goals and interests with minimal constraints.

- *Judicial conservatism's emphasis on respect for democratic decisions and government authority leads conservative judges to defer to government civil/criminal prosecutorial actions, including government decisions to restrict procedural protections, to impose punishments, and to prosecute in the first place. This is true even if those decisions violate contemporary notions of fairness or disproportionately harm political minorities. This deference to political branches, often coupled with an originalist – and therefore, often limited – view of constitutional protections, also makes conservative judges wary of overturning statutes and regulations enacted by legislatures or executive agencies. Likewise, in civil rights actions, judicial conservatism's skepticism toward government regulation of private transactions leads conservative judges to interpret broadly the scope of legally permissible/reasonable conduct by government and corporate defendants.

- *Judicial Liberalism* – the belief that the primary functions of the law and the judiciary are to ensure enjoyment of fundamental rights, to promote equity and legal equality among people, and to protect people from abuse by politically and economically powerful actors, both public and private.

- *Judicial liberalism's emphasis on protecting rights makes liberal judges open to second-guessing government civil/criminal prosecutorial actions, including government decisions to restrict procedural protections, to impose punishments, and to prosecute in the first place. This is particularly true where those decisions run counter to contemporary notions of fairness or disproportionately harm political minorities. This rights-oriented view, coupled with the notion that constitutional protections can evolve over time with social norms, also makes liberal judges relatively open to overturning democratically enacted statutes and regulations. Likewise, in civil rights actions, judicial liberalism's emphasis on the law as a tool for promoting equality and preventing abuse by powerful actors leads liberal judges to interpret narrowly the scope of legally permissible/reasonable conduct by government and corporate defendants.

- *Judicial Centrism* – a belief about the primary functions of the law and the judiciary that lies between conservatism and liberalism or which combines roughly equal elements of both.

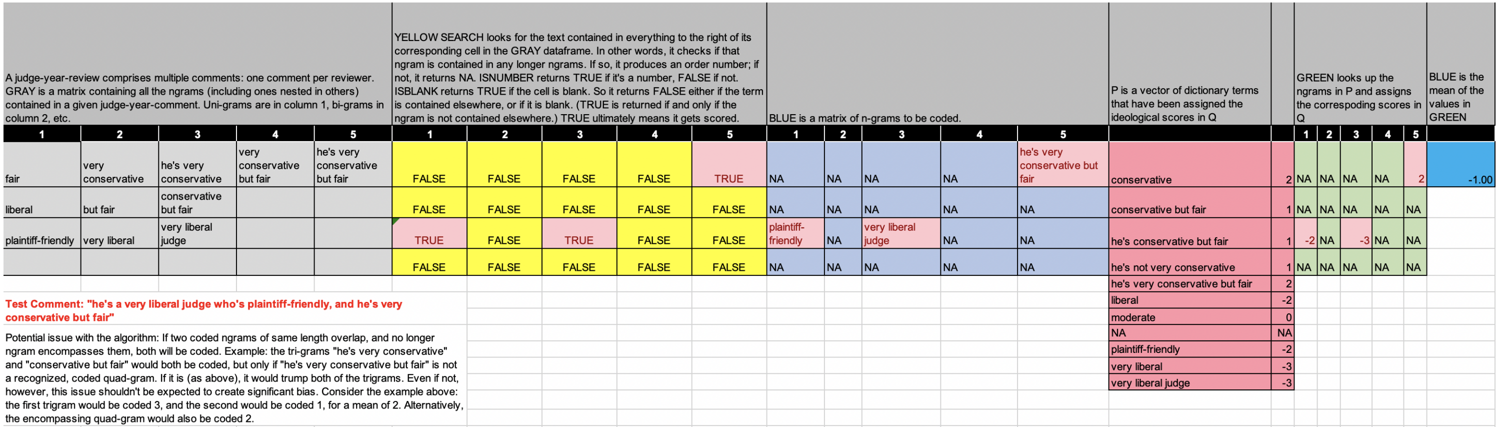
Assign scores based on the following:

-3 – an *extremely liberal* ideology
-2 – a *liberal* ideology (not at all centrist, but also not extremely liberal)
-1 – a *mildly liberal* ideology (roughly equal parts centrist and liberal)
0 – a completely *centrist* judicial ideology, without ideological leaning
1 – a *mildly conservative* ideology (roughly equal parts centrist and conservative)
2 – a *conservative ideology* (not at all centrist, but also not extremely conservative)
3 – an *extremely conservative* ideology

Note that references to “the government” indicate the *criminal* prosecutor.
For example, “always sides with government” might get a score of 3. “Consistent bleeding heart” might receive a -3. “Pretty harsh sentences” might get a 2. “Middle of the road” and “not ideological” would receive a 0.

Many phrases (indeed, the majority) do not have any particular ideological relevance. That is, they are not linked directly to judicial conservatism, liberalism, or centrism, or they otherwise give no useful information about ideology. For those, assign a score of ’99.’ (Do **not** give a score of ‘0,’ which would denote centrism.) This rule applies to statements that some (including yourself) may associate empirically with ideology but are not *conceptually* linked to ideology itself as defined above. That is, suppose that you believe that liberal judges tend to be more “activist.” Because those traits are not associated with the definitions of ideology, they are irrelevant and should not be given an ideology score. For example, “a great legal mind” and “opinions short and succinct,” are ideologically irrelevant and would receive a 99. Likewise, although the phrase “very biased” would implicate ideology, it is uninformative (i.e., it could be either -3 or 3) and should also receive a 99.

4. Visual illustration of hierarchical ngram algorithm



A judge-year-review comprises multiple comments: one comment per reviewer. GRAY is a matrix containing all the ngrams (including ones nested in others) contained in a given judge-year-comment. Uni-grams are in column 1, bi-grams in column 2, etc.					YELLOW SEARCH looks for the text contained in everything to the right of its corresponding cell in the GRAY dataframe. In other words, it checks if that ngram is contained in any longer ngrams. If so, it produces an order number; if not, it returns NA. ISNUMBER returns TRUE if it's a number, FALSE if not. ISBLANK returns TRUE if the cell is blank. So it returns FALSE either if the term is contained elsewhere, or if it is blank. (TRUE is returned if and only if the ngram is not contained elsewhere.) TRUE ultimately means it gets scored.					BLUE is a matrix of n-grams to be coded.					P is a vector of dictionary terms that have been assigned the ideological scores in Q		GREEN looks up the ngrams in P and assigns the correspoding scores in Q					BLUE is the mean of the values in GREEN
1	2	3	4	5	1	2	3	4	5	1	2	3	4	5			1	2	3	4	5	
fair	very conservative	he's very conservative	very conservative but fair	he's very conservative but fair	FALSE	FALSE	FALSE	FALSE	TRUE	NA	NA	NA	NA	he's very conservative but fair	conservative	2	NA	NA	NA	NA	2	-1.00
liberal	but fair	conservative but fair			FALSE	FALSE	FALSE	FALSE	FALSE	NA	NA	NA	NA	NA	conservative but fair	1	NA	NA	NA	NA	NA	
plaintiff-friendly	very liberal	very liberal judge			TRUE	FALSE	TRUE	FALSE	FALSE	plaintiff-friendly	NA	very liberal judge	NA	NA	he's conservative but fair	1	-2	NA	-3	NA	NA	
					FALSE	FALSE	FALSE	FALSE	FALSE	NA	NA	NA	NA	NA	he's not very conservative	1	NA	NA	NA	NA	NA	
															he's very conservative but fair	2						
															liberal	-2						
															moderate	0						
															NA	NA						
															plaintiff-friendly	-2						
															very liberal	-3						
															very liberal judge	-3						

Test Comment: "he's a very liberal judge who's plaintiff-friendly, and he's very conservative but fair"

Potential issue with the algorithm: If two coded ngrams of same length overlap, and no longer ngram encompasses them, both will be coded. Example: the tri-grams "he's very conservative" and "conservative but fair" would both be coded, but only if "he's very conservative but fair" is not a recognized, coded quad-gram. If it is (as above), it would trump both of the trigrams. Even if not, however, this issue shouldn't be expected to create significant bias. Consider the example above: the first trigram would be coded 3, and the second would be coded 1, for a mean of 2. Alternatively, the encompassing quad-gram would also be coded 2.

Figure A4.1

5. Additional predictive validity tests

To further gauge the predictive validity of the JuDJIS *Circuit Ideology* measure relative to existing measures of circuit-judge ideology, I draw on a data set comprising three sets of federal three-judge panel circuit cases, assembled by Cope and Fischman (2017), Cope and Fischman (2020), and Law (2004). The dataset comprises 4,482 decisions on several subject areas for the Seventh Circuit (1,835 cases, 2017–20), Ninth Circuit (1,693 cases, 1995–2000), and Tenth Circuit (954 cases, 2006–16). While these cases disproportionately constitute "easy" cases, in which the members of the panel are unanimous over 95% of the time, I include them here as robustness analysis, because they better reflect the run-of-the-mill decision-making of circuit judges. As with the en banc data, each decision is coded as a liberal or conservative result, using a method similar to that of Spaeth et al. (2014).

As Table A5.1 shows, I conduct two tests, one at the panel level (using the ideology of the panel's median judge as the predictor, and the panel decision as the outcome) and one at the judge level (using an individual judge's ideology as the predictor, and that judge's vote as the outcome). For each, I (1) estimate a logit model, comparing the respective normalized correlation coefficients, and (2) plot an ROC curve, comparing the areas under the curve.

Table A5.1. Predictive Validation Types and Locations

Case Type Unit of Analysis	Three-Judge Panel	En Banc
Panel	Appendix	–
Judge	Appendix	Text

Note: Validations marked as "App'x" are conducted in the appendix but summarized or referenced in the main text.

5.1 Predictive validation using panel outcomes

I first analyze the data using the median ideology score – for each of the four scores – as the predictor variable. In this analysis, where one or two of the three scores are unavailable (typically, if the judge is missing from that ideology data set or if the judge (often a district judge) is sitting by designation from outside the circuit), the median score is calculated nonetheless. As a robustness check, I also conduct this analysis using an alternate specification, in which the case is coded NA if any of the judge scores are unavailable. Both results are reported below.

5.1.1 Panel outcome logit regression

I first estimate a logit model. For dynamic scores, a judge's score is averaged over their tenure. That is, I regress the outcome (whether the panel decided the case in a conservative way) on the predictor (the mean ideology of the three-judge panel, as respectively estimated by JuDJIS, Party-of-the-Appointing President, JCS, and CBI). As with the other validity tests, the data come from data sets assembled by Cope and Fischman (2017), Cope and Fischman (2020), and (Law 2004). Cases are retained as long as at least one judge's score is available.

Table A5.2 presents the results from the logit models using those data. The coefficients indicate the marginal effects of a two-standard-deviation increase in conservativeness in each respective ideology measure on the probability of a conservative case outcome. (The overall incidence of conservative outcomes in the data set is approximately 0.85.) Thus, for example, a panel with a median judge JuDJIS ideology score of 1.08 is about 7.7 percentage points more likely to produce a conservative outcome than a panel with a median JuDJIS ideology score of -0.26. All four scores are associated with case outcomes at highly significant levels, but the strength of the other three measures' relationships between median-judge score and outcome is significantly smaller than JuDJIS's: 5.4 percentage points (CBI); 4.7 percentage points (JCS); and 4.3 percentage points (PAP).

Table A5.2. Logit predictions: Marginal effects of median panel score on probability of a conservative case outcome

	Probability of a conservative case outcome			
Median JuDJIS Score	0.077			
	(5.12×10^{-15})			
Median PAP Score		0.043		
		(2.58×10^{-5})		
Median JCS Score			0.047	
			(4.71×10^{-6})	
Median CBI Score				0.054
				(1.66×10^{-6})
Num. obs.	4, 441	4, 481	4, 481	4, 205

Note: Coefficients are normalized to indicate the change in probability associated with a two standard-deviation change in the given measure's panel median. p scores are in parentheses.

Figure A5.1 further illustrates these relationships by plotting, for each panel outcome, the median judge's ideology score on the x-axis. The outcomes are plotted on the y-axis, with conservative outcomes (1) at the top and liberal ones (0) at the bottom. (They are vertically jittered to show density.) For each graph, a logit regression curve shows the relationship between the two variables.

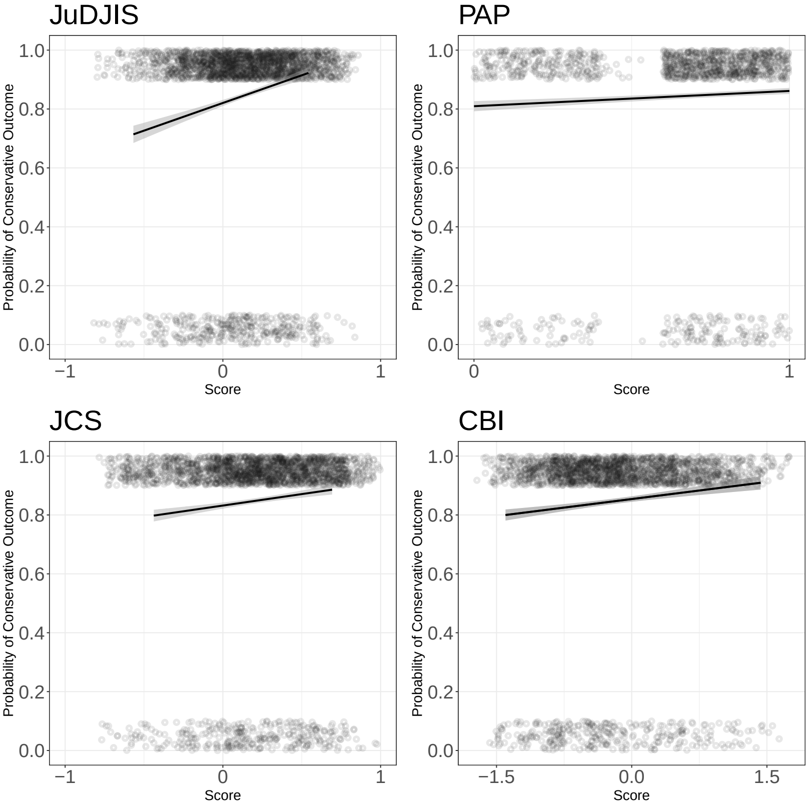


Figure A5.1. Probability of a conservative outcome as a function of median panel ideology score, by measure: three-judge panel cases

5.1.2 Panel outcome ROC analysis

I next plot ROC curves, indicating the accuracy of each measure in predicting panel outcomes from panel ideology scores. The top panel of Figure A5.4 displays the curves. The bottom panel gives the AUC values for each of the four measures. JuDJIS achieves an AUC value of 0.600; CBI's AUC value is 0.565; JCS's AUC value is 0.553; and PAP's AUC value is 0.540. The predictive performance of CBI, JCS, and PAP are similar, but JuDJIS performs significantly better. A DeLong test indicates that the difference between JuDJIS and each of the others is significant or borderline significant at conventional levels (vs. CBI: p = 0.051; vs. JCS: p = 0.005; vs. PAP: p = 0.0001).[1]

[1] Under an alternate specification, cases are dropped for that ideology data set if an ideological score is unavailable for any of the three judges. This results in the following number of dropped cases for each measure: JuDJIS: 1,099 (24.5%); PAP: 400 (8.9%); JCS: 400 (8.9%); CBI: 3,286 (76.7%). The ROC/AUC results are as follows: JuDJIS AUC: 0.606; PAP AUC: 0.5334; JCS AUC: 0.5469; CBI AUC: 0.5297. JuDJIS's AUC is statistically distinguishable from each of the other three ($p \leq 0.013$).

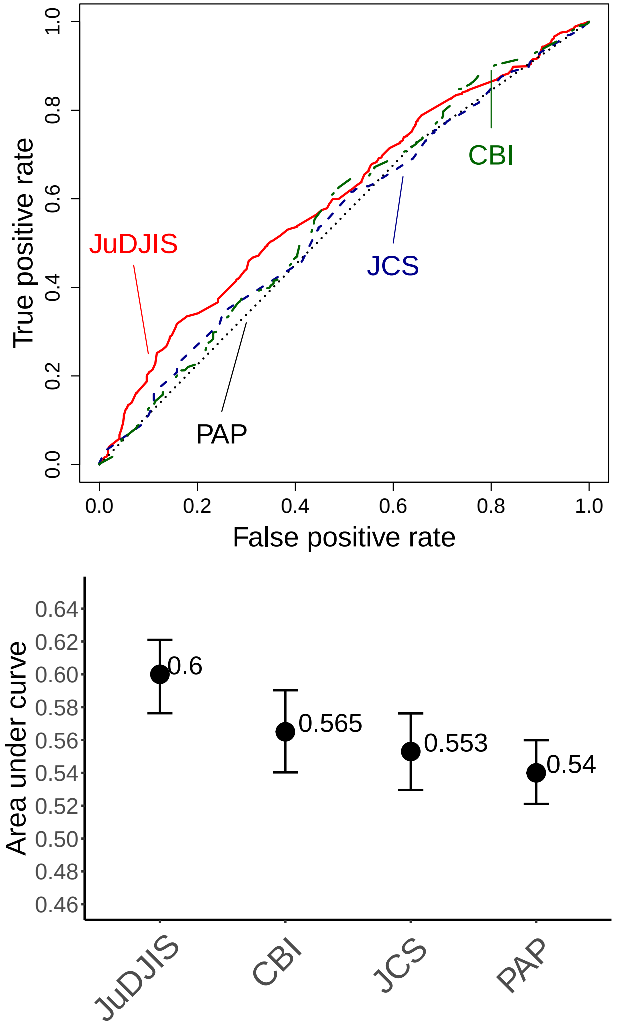


Figure A5.2. Top: ROC curves comparing four measures' success at predicting case outcomes; Bottom: Comparative areas under the ROC curve

5.2 Predictive validation using panel judge votes

I next gauge the respective measures' predictability at the judge level, by using an individual judge's ideology as the predictor, and that judge's vote as the outcome. As with the three-judge-panel validity test, the data come from data sets assembled by Cope and Fischman (2017), Cope and Fischman (2020), and (Law 2004).

5.2.1 Panel judge votes logit regression

Again, I first estimate a logit model. Table A5.3 presents the results. As above, the coefficients indicate the marginal effects of a two-standard-deviation increase in conservativeness in each respective ideology measure on the probability of a conservative vote. (The overall incidence of conservative votes in the data set is approximately 0.845.) Thus, for example, a judge with a JuDJIS ideology score of 0.39 is about 6.5 percentage points more likely to vote conservatively than a judge with a JuDJIS ideology score of -0.16. All four scores are associated with case votes at highly significant levels, but the relationship between a two-standard-deviation change in median-judge score is significantly smaller: 4.7 percentage points (PAP); 4.3 percentage points (JCS); and 4.4 percentage points (CBI).

Table A5.3. Logit predictions: Marginal effects of judge score on probability of casting a conservative panel vote

	Probability of a judge's casting a conservative panel vote			
JuDJIS Score	0.065			
	(2.20×10^{-16})			
PAP Score		0.047		
		(1.43×10^{-12})		
JCS Score			0.043	
			(2.16×10^{-12})	
CBI Score				0.044
				(8.15×10^{-8})
Num. obs.	13,043	13,043	13,043	7,967

Note: Coefficients are normalized to indicate the change in probability associated with a two standard-deviation change in the given measure's sample mean. p scores are in parentheses.

Figure A5.3 further illustrates these relationships by again plotting, for each vote, the judge's ideology score on the x-axis. The judge's votes are plotted on the y-axis, with conservative votes (1) at the top and liberal ones (0) at the bottom. (They are vertically jittered to show density.) For each graph, a logit regression curve shows the relationship between the two variables.

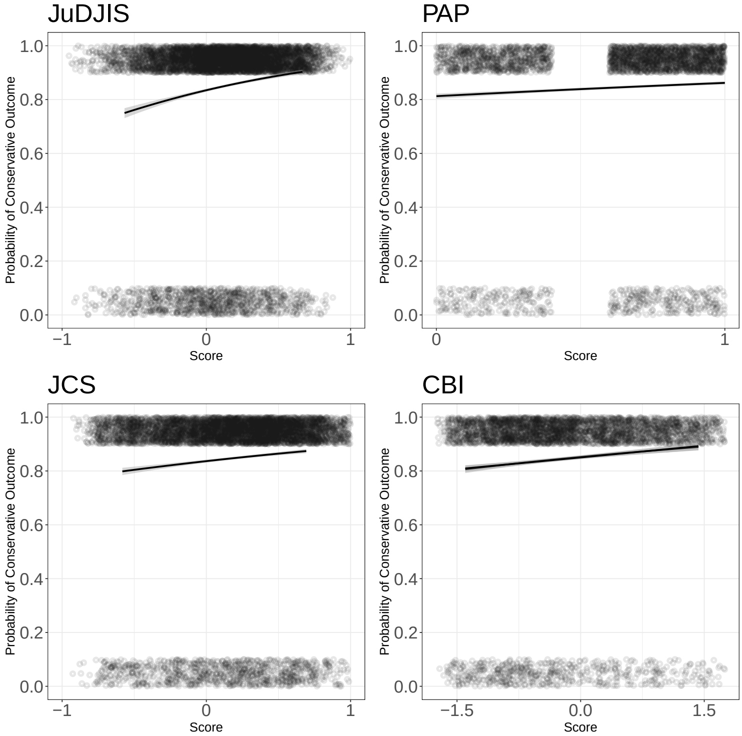


Figure A5.3. Probability of a conservative vote as a function of judge ideology score, by measure: three-judge panel cases

5.2.2 Panel judge votes ROC analysis

I once again plot ROC curves, indicating the accuracy of each measure in predicting judge votes from judge ideology scores. The top panel of Figure A5.4 displays the ROC curves for the judge-level analysis. The bottom panel gives the AUC values for each of the four measures. JuDJIS achieves an AUC value of 0.573; CBI's AUC value is 0.558; JCS's AUC value is 0.544; and PAP's AUC value is 0.542. The predictive performance of JCS and PAP are quite close; a DeLong test indicates that JCS's and PAP's areas are statistically indistinguishable from each other. ($p = 0.687$). JuDJIS performs significantly better than both JCS ($p = 1.01 \times 10^{-5}$) and PAP ($p = 2.31 \times 10^{-9}$), but is not distinguishable from CBI at conventional levels ($p = 0.172$).

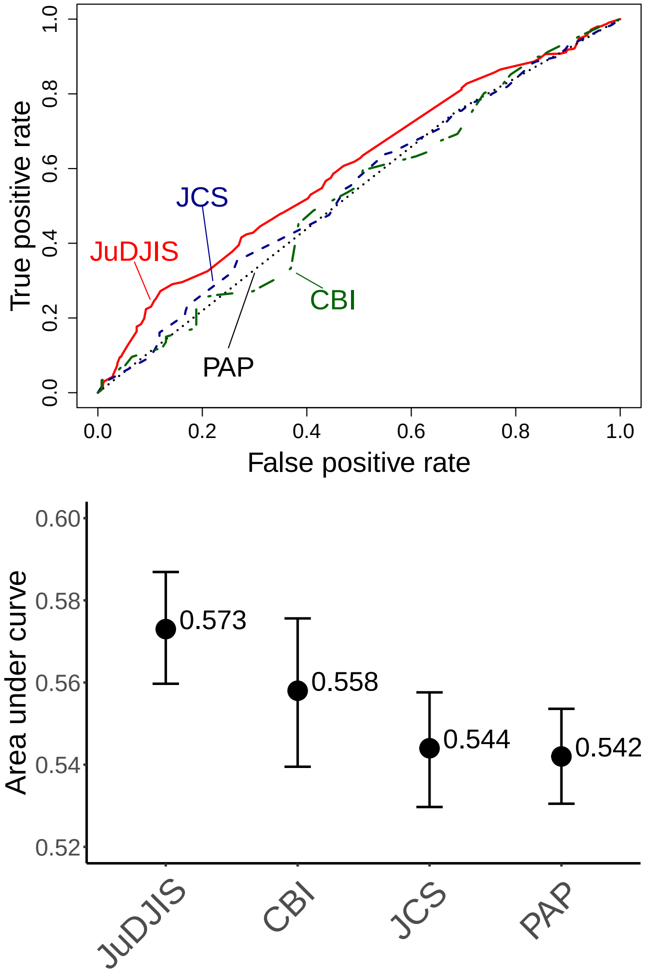


Figure A5.4. Top: ROC curves comparing four measures' success at predicting case outcomes; Bottom: Comparative areas under the ROC curve

6. JuDJIS ideology data

This section provides additional statistics from the JuDJIS ideology data. It includes a table of descriptive statistics; a comparison of ideology scores of future Supreme Court nominees; a discussion of judge-level ideology changes; and a pilot analysis of a sample of district court chief judges.

6.1 Descriptive statistics

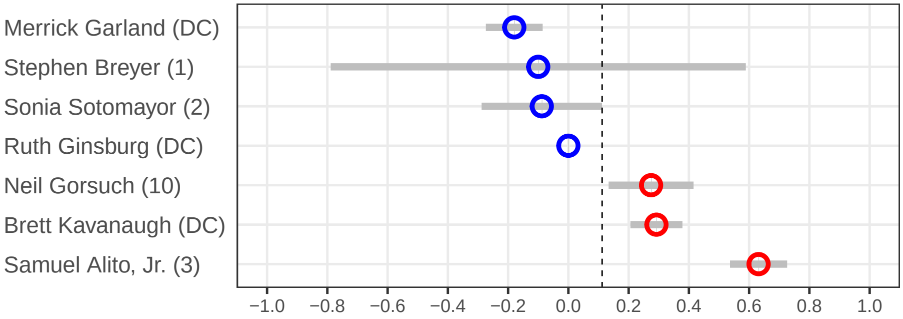
Table A6.4 provides summary statistics for the JuDJIS *Circuit Ideology* data set, aggregated, and disaggregated by party of appointing president and by the judge's gender. At the judge level (i.e., a judge's scores aggregated over full tenure), the mean ideology score is 0.11, the median score is 0.10, and the standard deviation is 0.33. Thus, the dataset exhibits a clear conservative slant. Table A6.4 shows that this appears to stem from two phenomena: (1) Republican presidents have appointed most (55%) of the judges in the dataset; and (2) Democratic-appointed judges are more moderate, i.e., the mean ideology score of Republican-appointed judges (.31) is more conservative than the mean ideology score of Democrat-nominated judges (-.13) is liberal.

Table A6.4. Summary statistics: JuDJIS circuit-judge ideology dataset

		Appt. Pres.		Race/Ethnicity				Gender	
	Aggreg.	**Dem.**	**Rep.**	**Asian**	**Black**	**Hispanic**	**White**	**Woman**	**Man**
					Judge Level				
Mean	0.11	-0.13	0.31	-0.05	-0.09	0.03	0.14	0.02	0.13
Median	0.10	-0.11	0.33	0.00	-0.07	0.00	0.16	0.00	0.16
Min.	-0.80	-0.80	-0.67	-0.32	-0.67	-0.56	-0.80	-0.67	-0.80
Max.	1.00	0.93	1.00	0.27	0.67	0.47	1.00	0.69	1.00
SD	0.33	0.26	0.23	0.20	0.27	0.28	0.33	0.30	0.33
N	418	190	228	7	35	21	355	84	334
Percent	1.00	0.45	0.55	0.02	0.08	0.05	0.85	0.20	0.80
					Judge-Year Level				
Mean	0.13	-0.16	0.32	-0.09	-0.15	0.05	0.16	0.03	0.15
Median	0.12	-0.13	0.33	0.00	-0.12	0.03	0.17	0.00	0.17
Min.	-0.94	-0.94	-0.93	-0.61	-0.76	-0.62	-0.94	-0.80	-0.94
Max.	1.00	1.00	1.00	0.40	0.79	0.69	1.00	0.87	1.00
SD	0.36	0.30	0.27	0.28	0.30	0.32	0.36	0.36	0.36
N	1857	742	1115	15	125	86	1631	338	1519
Percent	1.00	0.40	0.60	0.01	0.07	0.05	0.88	0.18	0.82

6.2 Supreme court nominees

Figure A6.1 gives the JuDJIS ideology scores, aggregated over their tenures, for the seven former judges in the data set who were nominated to the Supreme Court. Not surprisingly, the four Democratic nominees have scores that are negative or zero. (Ruth Bader Ginsburg's score is discussed in the main text of this Article.) In contrast, the three Republican nominees have positive scores. Note that Stephen Breyer's score indicates a moderate liberal, though it is estimated imprecisely, given his short stint on the First Circuit and relatively limited record for review.



Note: Circles denote means; horizontal gray bars denote 90% confidence intervals.

Figure A6.1. JuDJIS Ideology Scores: Supreme Court Nominees

6.3 Change analyses

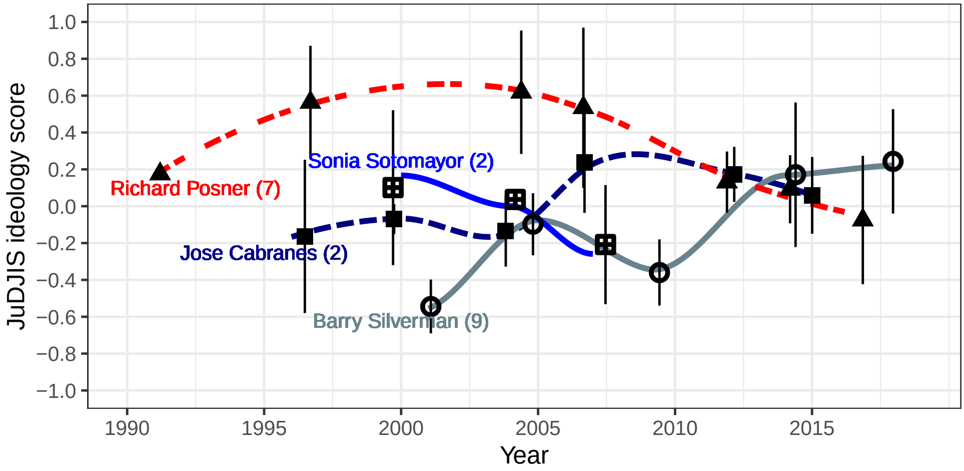
Below I provide a brief analysis of how the ideologies of individual judges have changed over time. Table A6.5 provides summary statistics of judge-level change in JuDJIS ideology score. As it shows, the mean change is a small shift toward liberalism, though the change is not statistically significant. Notably, Democrat-appointed judges show a small (but insignificant) conservative shift, and Republican-appointed judges show a small (but, again, insignificant) liberal shift. Though we cannot conclude that these figures represent real change, the possibility that judges tend to moderate slightly over their career would represent a significant and important finding, meriting further study.

Table A6.5. Summary statistics: Change in judge-level JuDJIS circuit ideology

	Min.	Median	Mean	Max.
Full Sample	-0.88	0.00	-0.018	0.83
Appointed by: Democrat	-0.76	0.00	0.02	0.83
Appointed by: Republican	-0.88	-0.01	-0.05	0.58

To illustrate particular cases driving these trends, Figure A6.2 gives examples of judges with notable observed changes over their tenures. They include high-profile judges like Richard Posner and Sonia Sotomayor, both of whom finished their circuit tenures more liberal than earlier points in their careers. Both were initially appointed to the federal bench by Republican presidents: Posner to the Seventh Circuit by Ronald Reagan in 1981, and Sotomayor to the U.S. District Court for the Southern District of New York by George H. W. Bush in 1992. Sotomayor was later appointed to the Second Circuit by Bill Clinton in 1998. Conversely, Jose

Cabranes and Barry Silverman appeared to become somewhat more conservative toward the end of the period. (Both were still serving as senior judges as of 2024.) Cabranes is a 1994 Clinton appointee, and Silverman is a 1998 Clinton appointee. Thus, Posner, Cabranes, and Silverman would seem to represent examples of the moderating trend that the full statistics suggest.

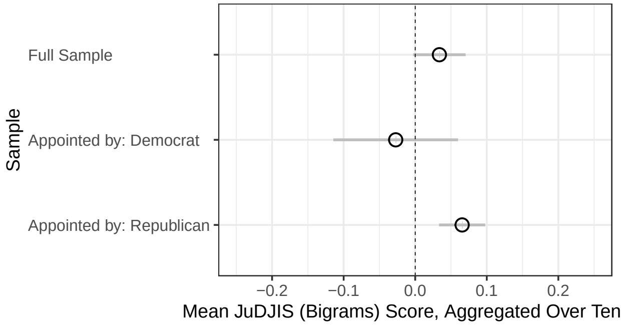


Note: Circles denote means; vertical gray bars denote 90% confidence intervals.

Figure A6.2. Notable Ideological Shifts

6.4 District court ideology data

This section provides a brief pilot descriptive analysis of a sample of the *District Ideology* data, namely, the 149 *chief* judges of the 98 districts who served at any point between 1990 and 2001. I focus on the chief judges here because, by definition, chief judges have achieved a high level of seniority and thus have typically assembled substantial records for evaluation. The data below aggregate the judges' scores over their tenure, and there is no particular reason to think that the ideologies of those who have served as chief district judges differ meaningfully from those of district judges generally.



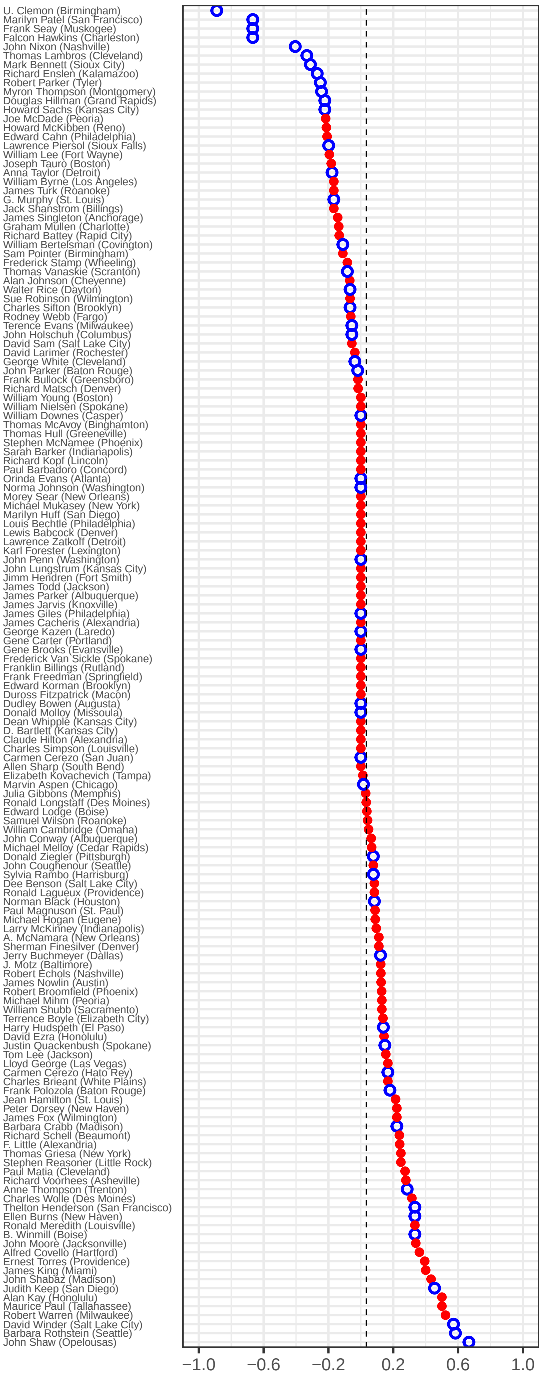
Note: Circles denote means for the given sample; gray bars denote 95% confidence intervals.

Figure A6.1. JuDJIS Descriptive Ideology Descriptive Statistics

As Figure A6.1 shows, the mean judge-level score for this data sample is 0.036. The mean score for Democratic-president appointees is -0.033; the mean score for Republican-president appointees is 0.070.

Thus, compared with the circuit judges, there is less variation between judges, and less variation between judges by appointing president party. There are several possible reasons for this difference. The most plausible is that the job of district judge simply involves less opportunity to express one's ideology. This pushes judges' scores (with many notable exceptions) closer to the center than those of appellate judges.

Figure A6.2 gives the ideology scores for the 149 judges in this sample, ranked from most liberal to most conservative. Again, it is notable that, while there is notable variation, the majority of judges lie within the [-0.5,0.5] range.



Note: Sorted from most liberal (-1) to most conservative (+1) by average ideology score, averaged over the judge's district court tenure. Blue open circles denote Democratic president appointees; red closed circles denote Republican president appointees.

Figure A6.2. *JuDJIS* U.S. Chief District Judge Ideologies, 1990-2001

References

Albaugh, Quinn, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. The automated coding of policy agendas: a dictionary-based approach. In *6th annual comparative agendas conference, atnwerp, beligum.*

Albaugh, Quinn, Stuart Soroka, Jeroen Joly, Peter Loewen, Julie Sevenans, and Stefaan Walgrave. 2014. Comparing and combining machine learning and dictionary-based approaches to topic coding. In *Th annual comparative agendas project (cap) conference, konstanz, germany.*

Bellin, Jeffrey. 2016. How merrick garland could help heal america.

Bommasani, Rishi, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. 2024. The foundation model transparency index v1. 1: may 2024. *arXiv preprint arXiv:2407.12929.*

Cope, Kevin L, and Joshua B Fischman. 2017. For a trump nominee, neil gorsuch's record is surprisingly moderate on immigration and employment discrimination. *FiveThirtyEight, https://fivethirtyeight.com/fea a-trump-nominee-neil-gorsuchs-record-is-surprisingly-moderate-on-immigration/.*

———. 2020. An empirical analysis of judge amy coney barrett's record on the seventh circuit. *Available at SSRN 3710951.*

Dai, Hetong, Heng Li, Che-Shao Chen, Weiyi Shang, and Tse-Hsun Chen. 2020. Logram: efficient log parsing using *n* n-gram dictionaries. *IEEE Transactions on Software Engineering* 48 (3): 879–892.

Dey, Atanu, Mamata Jenamani, and Jitesh J Thakkar. 2018. Senti-n-gram: an n-gram lexicon for sentiment analysis. *Expert Systems with Applications* 103:92–105.

Farah, Hassan Abdirahman, and Arzu Gorgulu Kakisim. 2023. Enhancing lexicon based sentiment analysis using n-gram approach. In *Smart applications with advanced machine learning and human-centred problem design,* 213–221. Springer.

Grimmer, Justin, and Brandon M Stewart. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21 (3): 267–297.

Law, David S. 2004. Strategic judicial lawmaking: ideology, publication, and asylum law in the ninth circuit. *U. Cin. L. Rev.* 73:817.

Osnabrügge, Moritz, Elliott Ash, and Massimo Morelli. 2023. Cross-domain topic classification for political texts. *Political Analysis* 31 (1): 59–80.

Presannakumar, Krishna, and Anuj Mohamed. 2021. An enhanced method for review mining using n-gram approaches. In *Innovative data communication technologies and application: proceedings of icidca 2020,* 615–626. Springer.

Spaeth, Harold, Lee Epstein, Ted Ruger, Keith Whittington, Jeffrey Segal, and Andrew D Martin. 2014. Supreme court database code book. *URL: http://scdb. wustl. edu.*

Xhymshiti, Meriton. 2020. Domain independence of machine learning and lexicon based methods in sentiment analysis. B.S. thesis, University of Twente.