How Judicial Identity Changes The Text Of Legal Rulings

Michael Z. Gill and Andrew B. Hall*

September 20, 2013

Abstract

In the common-law tradition, word-usage in legal documents matters because of the principle of *stare decisis*, the doctrine that requires judges to read and interpret previous rulings relevant to the case at hand. Yet we understand very little of how legal authorities actually write legal texts. We analyze 22,773 cases from the United States Courts of Appeals, and we show how judicial identity affects the text of the written case rulings. We demonstrate that the random assignment of a female judge or of a non-white judge to U.S. Appellate Court panels causes systematic changes in the frequencies with which specific, legally-important words appear in the final ruling, along with the rates at which constitutional amendments and landmark Supreme Court cases are cited. Panels present different arguments—and thus leave a different legacy for future jurists to interpret—depending on the identity of the judges chosen by lot to preside over the case at hand.

*Both authors contributed equally. Michael Z. Gill (http://scholar.harvard.edu/gill, mzgill@fas.harvard.edu) and Andrew B. Hall (http://www.andrewbenjaminhall.com, hall@fas.harvard.edu) are Ph.D. Candidates in the Harvard University Department of Government. Earlier versions of this project were presented at the 2012 Annual National Midwest Political Science Association Conference, the poster session for the 2011 Annual Meeting for the Society for Political Methodology, the interdepartmental Gender & Sexuality Studies Workshop at Harvard University, and as a guest lecture for API-208 at the Harvard Kennedy School of Government. The authors thank the participants of these workshops for helpful feedback and criticism. In addition, the authors especially thank Alexis Diamond, Anthony Fowler, Adam Glynn, Gary King, Luke Miratrix, and Arthur Spirling. All remaining errors are the sole responsibility of the authors.

"These words cannot be meaningless, else they would not have been used."
United States v. Butler, 297 U.S. 1, 65 (1936) (per Roberts, J.).

Introduction

Judges, lawyers, legal scholars, and citizens interpret law by reading its precepts in written word. In the common-law tradition, word-usage in legal documents is especially important because of the principle of *stare decisis*, the doctrine "requiring that judges apply the same reasoning to lawsuits as has been used in prior similar cases."[1] Yet we understand very little of how legal authorities actually write the legal texts that future jurists read. Our theories and empirical studies of legal behavior focus on the final disposition of cases, rather than the content of written rulings. In this paper, we study written rulings themselves. We analyze a new dataset of 22,773 written rulings from the United States Courts of Appeals, 1970-2010, and we show how the composition of the three-judge panel that writes the ruling profoundly changes the legal vocabulary employed. In particular, we demonstrate that the random assignment of a female judge or of a non-white judge to U.S. Appellate Court panels causes large changes in the frequencies with which legally-important words appear in the final ruling, along with the rates at which the final ruling cites constitutional amendments and landmark Supreme Court decisions. Panels use different words and stress different concepts—and thus leave a different legacy for future jurists to interpret—depending on the identity of the judges chosen by lot to preside over the case at hand.

We contribute to a body of work concerning the effects of judicial identity on case outcomes (e.g., Boyd, Epstein and Martin 2010; Kastellec 2013; Sunstein et al. 2006) by offering new text-based measures of judicial outcomes. We also offer a new dataset that represents, to our knowledge, the largest collection of appellate court cases—and judicial text—analyzed

[1] `http://www.law.cornell.edu/wex/stare_decisis`

to date. We take advantage of the dataset's scale in investigating more rulings over a longer time period than has previously been possible. Finally, we also contribute to the burgeoning literature on "text as data" by offering several simple techniques for combining statistical methods for causal inference with methods from text analysis and information retrieval.

The paper is organized as follows. First, we describe the new dataset we use and we motivate our study of judicial text. Next, we describe our empirical strategy and present our basic statistical results on the overall differences in vocabulary between differently-composed panels. Following that, we explore the specific, legally-meaningful words that drive the overall statistical findings. Finally, we conclude.

Data and Background

In the United States Courts of Appeals, like in many legal bodies, a circuit's primary medium of communication is its written opinions (e.g., Schauer 1987). These texts, which explain how rulings are reached and why they are fair, inform future cases by precedent (Cross 2007). Whether this deference to precedent occurs because of a legal norm (e.g., Knight and Epstein 1996), other forms of cue-taking (e.g., George and Epstein 1992) or for strategic reasons (e.g., Bueno De Mesquita and Stephenson 2002; Kritzer and Richards 2003; Richards and Kritzer 2002), the result is that the content and style of opinions must be carefully written.[2] Indeed, "...appellate court judges do not simply decide which party will prevail. Their reasoned decisions are expected to address the issues raised by litigants and subsequently shape legal policy for the circuit" (Haire and Moyer 2007).

[2]Much of the literature debating the rationality of deference to precedent focuses on the Supreme Court, but the intuition often extends to lower courts as well. For a recent review of the "law vs. ideology" debate see Moyer (2012).

Case Selection

We collected a database of 22,773 legal rulings. Using LexisNexis Academic's database on "US Federal & State Cases," we retrieved the opinions published for all relevant cases heard by the US Courts of Appeals over more than four decades, excluding cases heard in the Federal Circuit and in the Temporary Emergency Court of Appeals.[3] With the search terms "Abortion," "ADA," "Affirmative Action," "Campaign Finance," "Capital Punishment," "Contract Clause," "EPA," "Federalism," "Piercing the Corporate Veil," "Sex Discrimination," "Sex Harassment," "Takings Clause," "Title VII Race," "Domestic Abuse," plus the word "Opinion," we obtained the raw text opinions of every relevant case heard under the jurisdiction of "All US Courts of Appeals" over the timeframe January 1st, 1970, to December 31st, 2012.[4]

Details On Dataset

Along with the text of the published rulings, the dataset contains information on the panel membership, including the gender and race of the judges. In this paper, we confine our study to two "treatments": the random assignment of a female judge and the random assignment

[3]By "relevant" we refer to a subset of legal issue areas commonly used in this literature, using the list of issue areas described in the Sunstein et al. (2006) dataset. Furthermore, analysis is limited to cases heard the US Courts of Appeals, due to both the mechanism of treatment assignment, and for reasons of across-issue sample size—e.g., the US Supreme Court hears on average 80 cases per session (Scalia and Garner 2012).

[4]Once all relevant documents were obtained, we assembled a unique corpus of documents for each of the legal issue areas listed above. To do this we wrote a script in R to aggregate each of the batched text files and separate the pooled texts into lists of texts for each individual case. Within each case, we used our script to extract pertinent information about each case in this sample: e.g., the name of the case, the circuit in which the case was heard, the names of the judges who sat on the panel, whether the opinion was published, whether the case was heard en banc, whether the case was a class-action suit, whether a case contained a dissenting opinion, and the year the case was argued. The list of judges names, circuits, and case years was then matched with a dataset of judicial biographical information to determine each judge's gender. The dataset on judicial biographies was taken from the "Biographical Directory of Federal Judges, 1789–present," available online at: http://www.fjc.gov/history/home.nsf/page/judges.html. Lastly, to clean our texts for analysis, we removed all LexisNexis legal headnotes, irrelevant non-opinion text (e.g., case titles, names of counsel), and the texts of dissenting opinions, if applicable. Term-document matrices were then generated for each of the legal issue areas listed above.

of a minority judge to appellate court panels. We leverage the random assignment of judges to appellate court cases (e.g., Ashenfelter, Eisenberg and Schwab 1995; Boyd, Epstein and Martin 2010; Kastellec 2013; Sen 2012; Sunstein et al. 2006). Because assignment is random, the presence of a woman or a minority on a case is randomly determined and, in expectation, is uncorrelated with confounding variables. We can therefore evaluate the causal effects of the presence of female or non-white judges on written rulings.[5]

Figures 1 and 2 show the incidence of treated and control observations for the gender and non-white treatments, respectively, over time across the major issue areas we analyze. Title VII race, federalism, and sex discrimination are among the most common case issue areas, while domestic abuse and campaign finance are among the least common. In the first figure, the treatment sizes trend upwards somewhat over time, representing the increased number of female judges on appellate courts.[6]

Vocabulary Effects

Most cases in the U.S. Courts of Appeals are heard by three-judge panels, with formal randomization used to assign judges to panels.[7] This randomization occurs at the circuit level. Pooling across circuits induces bias if the frequency of treatment across circuits is correlated with other variables (e.g., female judges might sit disproportionately in more

[5]We cannot—nor, indeed, can anyone—evaluate the causal role of "sex" or "race" itself. Such questions are incoherent in the causal inference framework, due the immutable nature of characteristics (e.g., Greiner and Rubin 2011; Holland 1986). It is not the judge's race or sex which is assigned, but rather his or her presence on the panel. For more discussion in the context of appellate courts, see Boyd, Epstein and Martin (2010). For general discussion, see for example Angrist and Pischke (2009).

[6]Not all legal issue areas are represented over the full interval from 1970 to 2012. For example, the Americans with Disabilities Act (ADA, which "prohibits discrimination on the basis of disability in employment, and other areas such as access to public places," was only passed by Congress in 1990 (as quoted from the U.S. Equal Employment Opportunity Commission official website, accessible online at: http://www.eeoc.gov/eeoc/history/45th/ada20/).

[7]We remove *en banc* decisions, which are heard by the entire Circuit, from all analyses.

Figure 1 – Total Gender-Treated and Control Observations, by Issue Area and Year. Each graph indicates the number of all-male panels (solid line) and panels with at least one female judge (dotted lines).

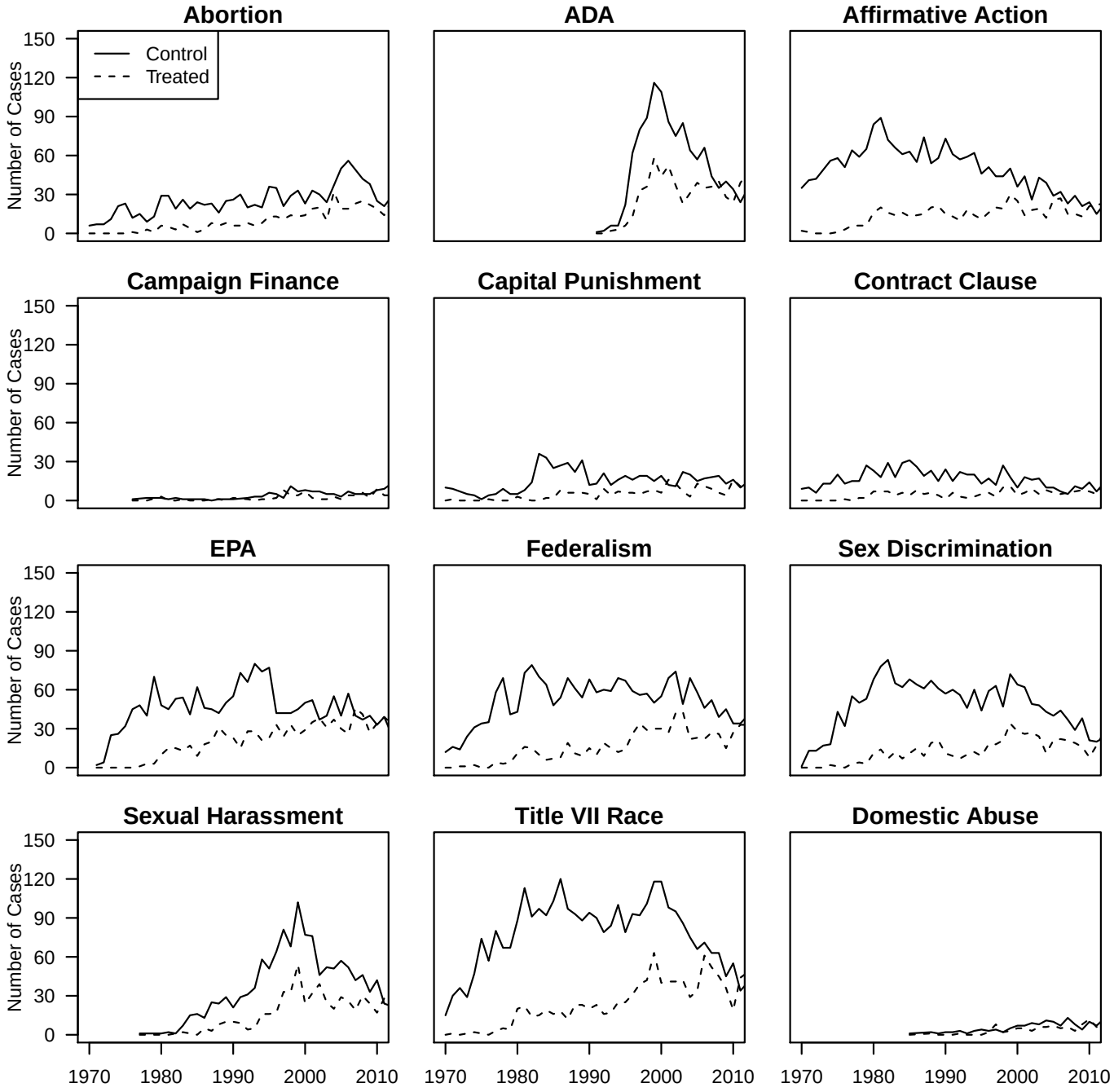
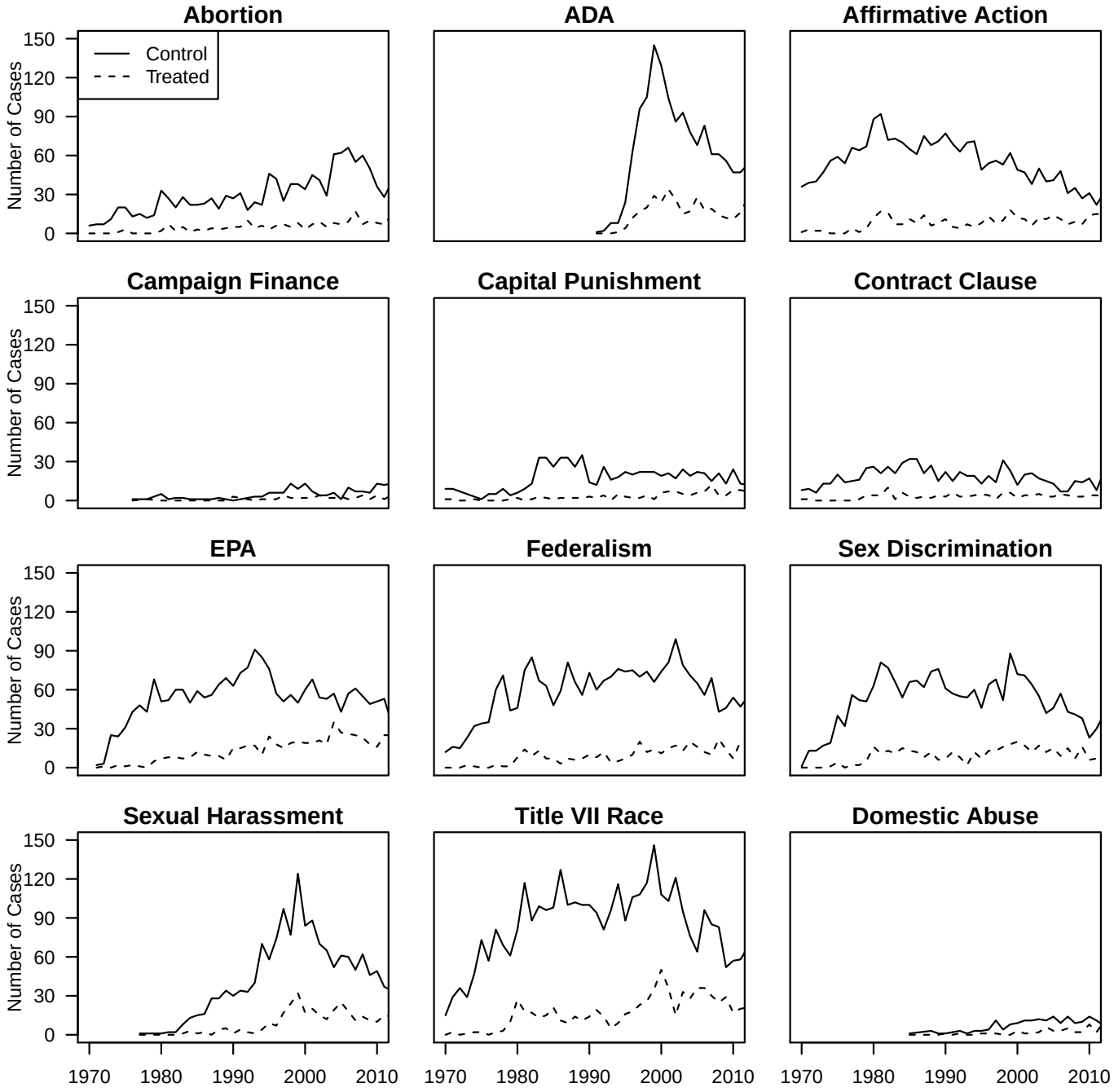


Figure 2 – Total Race-Treated and Control Observations, by Issue Area and Year. Each graph indicates the number of all-white panels (solid line) and panels with at least one non-white judge (dotted lines).



liberal circuits). All analysis therefore proceeds at the circuit level, calculating within-circuit treatment effects and creating a sample-size-weighted overall average effect.

To measure the disparities between the written rulings of panels with at least one female or at least one minority judge, respectively, we first use the standard "bag of words" approach to reduce text to data.[8] For each written ruling, we remove punctuation and stop words, along with words that appear in less than 5% or more than 95% of all texts. In order to focus only on meaningful words, we also remove all non-legally relevant words by filtering based on the legal words dictionary from FindLaw (`http://dictionary.findlaw.com/`).[9] Finally, we stem the words and condense all treated and control rulings within each issue area by creating a single word-frequency vector for each. To do so, we take the simple average of the document-specific word frequencies for all treated and control texts within an issue area, respectively.[10]

We then apply a common measure of document similarity,[11] the cosine similarity, defined for two document vectors A and B to be

$$Similarity(A, B) = \frac{A \cdot B}{||A||\ ||B||}.$$

Using this measure, we calculate the within-circuit sample-size-weighted average treatment effect $\hat{\tau}$ as

$$\hat{\tau}_k^{vocab} = \sum_{j=1}^{J} \frac{n_{jk}}{N_k} Similarity(Treat_{jk}, Control_{jk}), \tag{1}$$

[8] This approach is common in text analysis. For discussion see, for example, Hopkins and King (2010) and Grimmer and King (2011).

[9] We have also run the analyses using all words. Results are substantively the same, but word analysis is more difficult due to the large volume of words, most of which have no legal bearing.

[10] This method treats all documents as equally important, rather than weighting longer documents or documents from more salient cases more. We have no *a priori* technique for providing a weighting scheme, so we do not do so. Future research could explore other ways to weight documents in constructing a condensed word-frequency vector for applied purposes like ours.

[11] We have replicated the relevant results using Kullback-Leibler Divergence as an alternative distance measure and found little difference in substantive conclusions.

where circuits are indexed by j, issue areas are indexed by k, and N_k represents the overall sample size in issue area k (with n_{jk} therefore indicating the sample size in circuit j for issue area k). The variables $Treat_{jk}$ and $Control_{jk}$ are the issue-specific condensed document vectors for the treated and control groups in circuit j and issue area k.[12]

Inference

Within each each issue area, we estimate exact p-values through Monte Carlo permutation (e.g., Good 2005; Lehmann 2006). This technique is standard in the analysis of experimental and quasi-experimental data (e.g., Imbens and Rubin 2010; Rosenbaum 2010), and has been applied to political science research with increasing frequency in recent years (e.g., Gerber and Green 2012; Keele, McConnaughy and White 2012). In classical permutation framework (Fisher 1935; Pitman 1937), exact p-values denote the proportion of all possible permutations of treatment and control assignment that produce an estimated effect at least as extreme as the observed value of $\hat{\tau}$.[13]

Permutation tests are fully non-parametric, requiring no assumptions over statistical distributions.[14] Another way to state this is that permutation tests are "statistic-inclusive"

[12]The circuit-level randomization ensures that our treatment variables are uncorrelated, in expectation, with confounding variables, so long as we calculate the treatment effect within each circuit and weight the estimates together, rather than pool over circuits. Boyd, Epstein and Martin (2010), on the other hand, pools over circuits and uses matching on observed covariates as an alternative way to address bias. We do not pursue this course because of the increased sample size now available to us. In such situations, the inclusion of pre-treatment covariates via, e.g., regression or matching, can provide efficiency gains but is not necessary for addressing concerns of bias. We choose to avoid the use of covariates for a fairly simple reason. Given the large sample sizes a lack of statistical power is unlikely to be a problem.

[13]For example, consider a generic two-sample study where n_T is the number of units randomly assigned to treatment and n_C is the number randomly assigned to control. Given a total sample size of $n_T + n_C = N$, the total number of of unique assignments of size n to treatment and m to control is equal to $\binom{n_T}{n_C} = M$. The exact p-value corresponds to associated quantile of the observed value of $\hat{\tau}$ when compared against M re-calculations of the test statistic.

[14]Permutation tests differ in an important philosophical way from most parametric tests, in that they test the "sharp" null hypothesis that the treatment effect is zero for all units, not simply that the average across all units is zero. In some cases this null hypothesis may be of less interest (for example, if the social planner has quadratic loss over outcomes, then testing should focus on the Average Treatment Effect. However, rejecting the sharp null hypothesis also implies rejecting the more standard weak null, so our findings are

(e.g., Boos and Levanski 2013), meaning that any conceivable comparison statistic is valid for analysis since permutation occurs over treatment assignment profiles, not over outcomes. The method is thus ideally suited for the present context because our estimator—the cosine similarity between two groups of texts—has no analytic sampling distribution. Because sample sizes are large, we estimate p-values using Monte Carlo methods, repeatedly randomly permuting texts into treatment and control categories and recalculating the cosine similarity.[15] Following the construction of our test statistic, which calculates within-circuit effects to avoid bias from pooling, we permute observations at the circuit level.

Basic Results

We begin with a simple overarching question: does the presence of a female or non-white judge change the vocabulary of published rulings in a systematic way? This question is distinct from asking simply whether different judges use different words, which would manifest itself as pure noise in our estimates. Instead, we are looking for a systematic link between the presence of a particular group of judges and the content of written rulings. Below, we present the results for, in order, the gender treatment and the non-white treatment. In both cases, the answer is a clear "yes."

Gender Treatment

Figure 3 presents the results comparing all-male panels to panels with at least one female judge. Each figure presents the within-circuit sample-size-weighted cosine similarity (from Equation 1) between the treated and control panels—indicated by the dotted vertical line—

informative regardless (Ding 2013). In the present context, sample sizes are large, so the use of the sharp null hypothesis is unlikely to be of consequence.

[15] Since the estimation error of the simulated p-value (versus the exact p-value) comes purely from simulation error, we can estimate this quantity with arbitrarily high precision by increasing our simulation size.

as compared to the permutation distribution (the histogram). Statistically-significant effects are those for which the dotted-line appears extreme relative to the permutation distribution.

As the figure shows, the presence of a female judge causes statistically-discernible changes in the vocabulary of written rulings across many issue areas. Like Boyd, Epstein and Martin (2010), we find strong results in rulings over sex discrimination; however, we also find marked effects for abortion cases, affirmative action, capital punishment, EPA, and Title VII race cases. We find less evidence for an effect in ADA, campaign finance, contract clause, federalism, and sexual harassment cases.

Racial Treatment

Figure 4 presents the same results but comparing all-white panels to panels with at least one minority judge. Again, we find effects for many issue areas. We find statistically-discernible evidence that the presence of a minority judge changes the vocabulary of rulings on abortion, capital punishment, sex discrimination, and especially Title VII race cases.

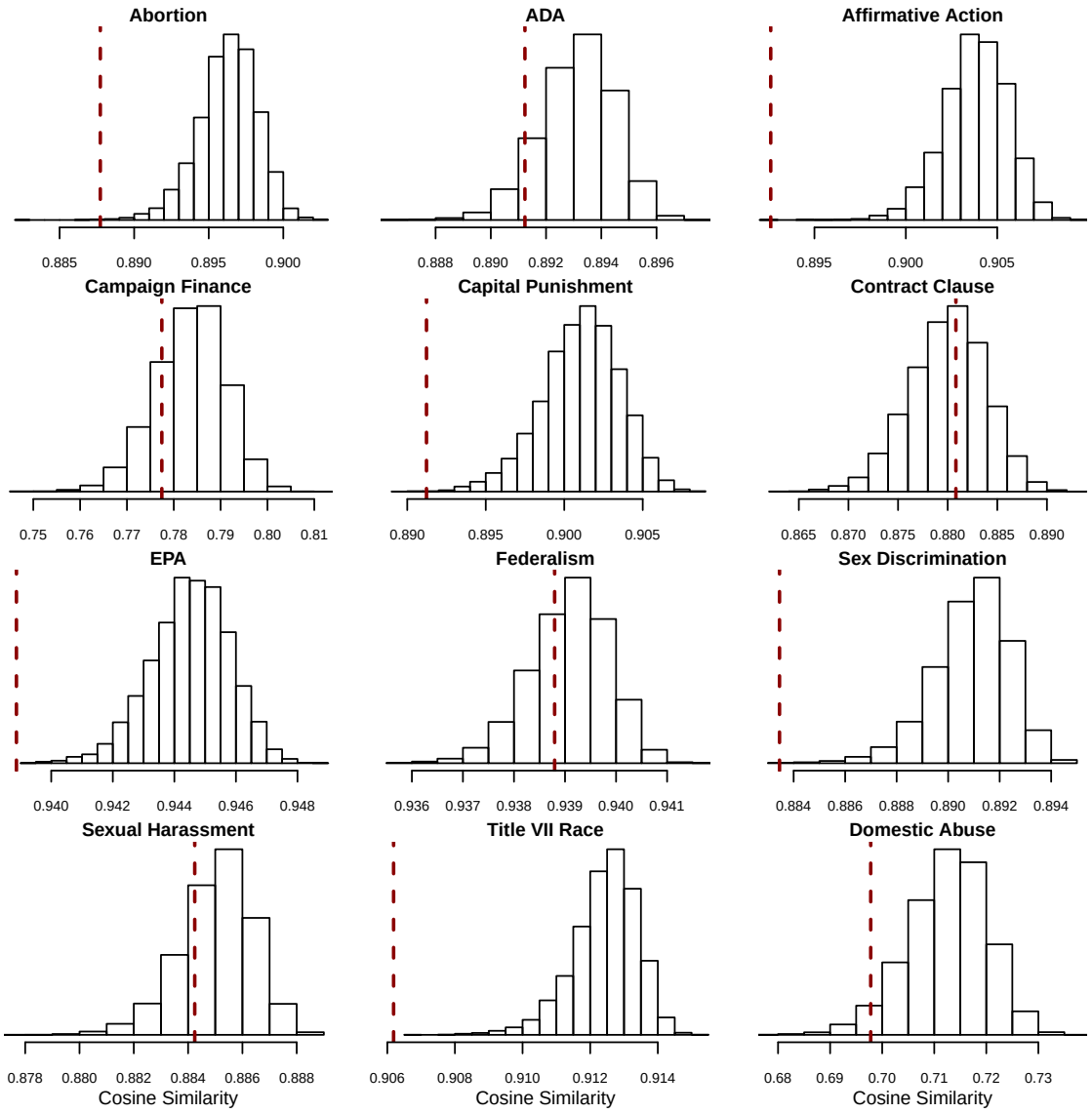
Discussion of Basic Results

A clear conclusion from these results is that panel composition, as influenced through judicial identity, affects the vocabulary of written rulings. In roughly half of all issue areas tested, extremely strong statistical differences appear.[16]

In addition, a tentative conclusion, consistent with contemporary research on judicial identity and case outcomes (e.g., Boyd, Epstein and Martin 2010; Kastellec 2013), is that judicial identity matters most in cases especially salient to judges' backgrounds, perhaps due to informational effects. In line with Boyd, Epstein and Martin (2010) in their study of case outcomes, we find that female judges have a strong effect on vocabulary in cases about

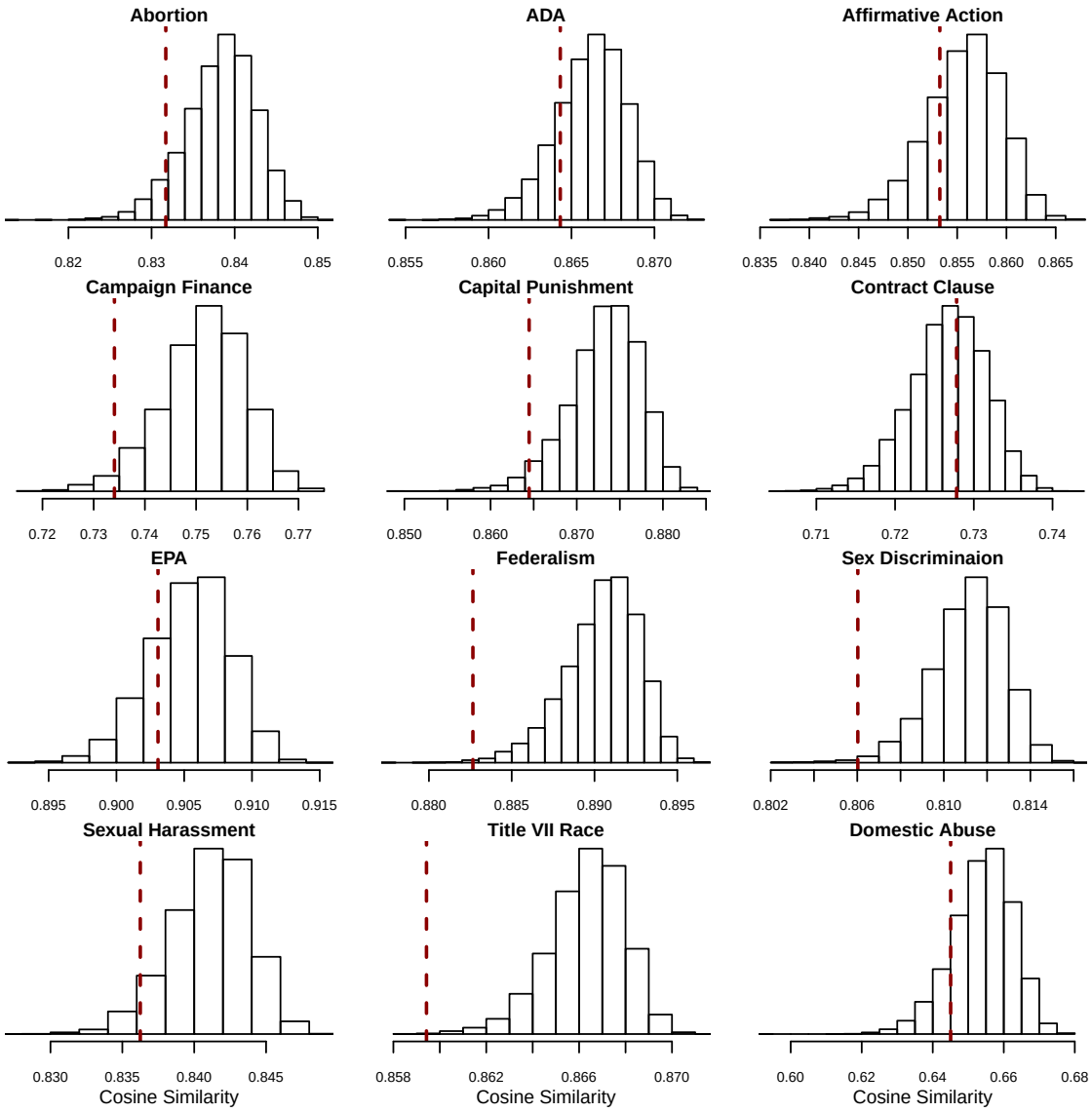
[16]This is far more than the 5% we would expect to reject by chance. Below, when we employ a large number of tests, we correct directly for multiple testing using the Bonferroni correction (e.g., Demsar 2006; Dunn 1961).

Figure 3 – Female Judges: Permutation Tests Across Issue Areas.



For each issue area, the within-circuit sample-size-weighted cosine similarity (from Equation 1) between all-male panels and panels with at least one female judge is represented by the dotted vertical line. These effects are compared to the permutation distributions within each issue area. Statistically-significant effects are those for which the dotted-line appears extreme relative to the permutation distribution.

Figure 4 – Minority Judges: Permutation Tests Across Issue Areas.



For each issue area, the within-circuit sample-size-weighted cosine similarity (from Equation 1) between all-white panels and panels with at least one minority judge is represented by the dotted vertical line. These effects are compared to the permutation distributions within each issue area.

sexual discrimination. And in a similar vein, we find strong evidence that minority judges influence the vocabulary of Title VII race cases.

At the same time, we also find strong effects in other issue areas one might not think of as salient to these groups. For example, we find effects for both female and minority judges on EPA cases. While we can only speculate as to the possible reasons for this, one possibility worth mentioning is that female and minority judges are also likely to be younger and more liberal.[17] Through this channel as well as others, we might expect these judges to exert influence on many kinds of cases and not just on those of identity-based salience. Such effects do not manifest themselves in the outcome of cases in these issue areas (Boyd, Epstein and Martin 2010), but they are present in the writing style. Though the outcome of the case is of primary import to the parties involved, it is the reasoning of the case, expressed through its written ruling, that will influence future cases. As such, the presence of vocabulary-based effects across so many issue areas is remarkable.

In the next section, we examine these results in more detail by looking at the actual words that drive the observed differences in vocabulary. By doing so we are able to present a fuller picture of how—and how much—judicial identity matters for the writing of legal rulings.

Analysis of Word Usage

The previous section established that the random assignment of a female or non-white judge systematically causes an overall change in the vocabulary used in published rulings. This result provides a first "litmus test" of the hypothesis that judicial identity changes written rulings, but it tells us only little about how *much* judicial identity matters for written

[17]There is increasing evidence that any judge's ideology is likely to vary across legal issue areas (e.g., Lauderdale and Clark 2012). However, to test the the effect of "judicial ideology" (a latent variable) as the assigned treatment requires unbiased estimates of judges' within-issue area ideologies.

rulings, or in what ways. In this section and those following it, we present several analyses to identify the specific legally-important words and citations that treated and control panels use at different rates.

To delve deeper, we estimate the causal effect of the presence of a female or non-white judge on the use of each of the 2,463 legally-important words in FindLaw's dictionary. To do so, we continue to use sample-size-weighted within-circuit estimates of the average treatment effect, using the equation

$$\hat{\tau}_{ik}^{words} = \sum_{j=1}^{J} \frac{n_{jk}}{N_k} (\bar{X}_{ijk}^1 - \bar{X}_{ijk}^0), \tag{2}$$

where $\bar{X}_{ijk}^1$ and $\bar{X}_{ijk}^0$ are the average frequencies for word i for cases in issue area k in circuit j in the treated (1) and control (0) panels. To adjust for any possible heteroskedasticity, standard errors are estimated with a non-parametric bootstrap. The typical issue area employs somewhere between 600 and 700 of the legally-important words. Over the 12 issues, we carry out 8,008 tests. The Bonferroni correction thus requires $p \approx 0.000006$ in order to reject at the .05 significance level (calculated as $0.05/8008 = 0.000006$).

In Figures 5, 6, 7, and 8 we present the estimates for $\hat{\tau}_{ij}^{words}$ for 8 of the 12 issue areas. For presentational purposes, we omit four categories for which sample sizes are smaller and differences therefore more variable. Graphs for these other issue areas are available in the Appendix.

In each figure, all legally-important word stems that appear in any rulings in the given issue area, arrayed vertically in alphabetical order. The horizontal axis represents the estimated treatment effect for that word stem in that issue area, with words towards the right used more in treated panels, and those to the left used more in control panels. Word stems that appear in red are those for which we can reject the null hypothesis of no treatment effect after the Bonferroni correction is applied.

Across the figures, systematic differences in the use of many legally-important words appear. To choose a few examples: panels with at least one female judge use the word stem "harass" 2.5 times more per Sex Discrimination case than do all-male panels; panels with at least one non-white judge use the word stem "state" 8 more times per affirmative action case than do all-white panels; and panels with at least one non-white judge use the word stem "discrimin" (as in, e.g., "discriminate") 4 times more per Title VII race case. These are just a few examples; a vast array of words are used differently across the different groups. Many of these words exhibit not merely statistically significant results but substantively meaningful differences in frequencies.

Differences in Citation Behavior

The analysis in the previous section revealed the substantial differences in vocabulary that panels employ depending on the identity of the judges assigned to the case. Along with the words rulings use, future jurists also scrutinize the connections authors draw to other legal texts. These connections are a crucial component of the legal arguments put forward in the ruling. In the next two subsections, we examine how treated and control panels differ in their legal reasoning in two important ways: first, in the way they cite constitutional amendments, and second, in the way they cite landmark Supreme Court cases.

Constitutional Amendments

To examine one aspect of citation behavior, we searched the content of written rulings for references to any of the 27 amendments to the U.S. Constitution. Figure 9 shows the relative frequency of these references across issue areas. Predictable patterns emerge. Abortion and campaign finance cases cite the first amendment with high frequency; Title VII race cases cite the 14th amendment with high frequency, and so forth.

Figure 5 – Female Treatment: Differential Word Usage. Words are plotted in alphabetical order based on the estimated effect from Equation 2. Words to the right are used more often in treated panels; those to the left more often in control panels. Words in red are those for which the null hypothesis of no effect is rejected at the .05 level after Bonferroni correction.

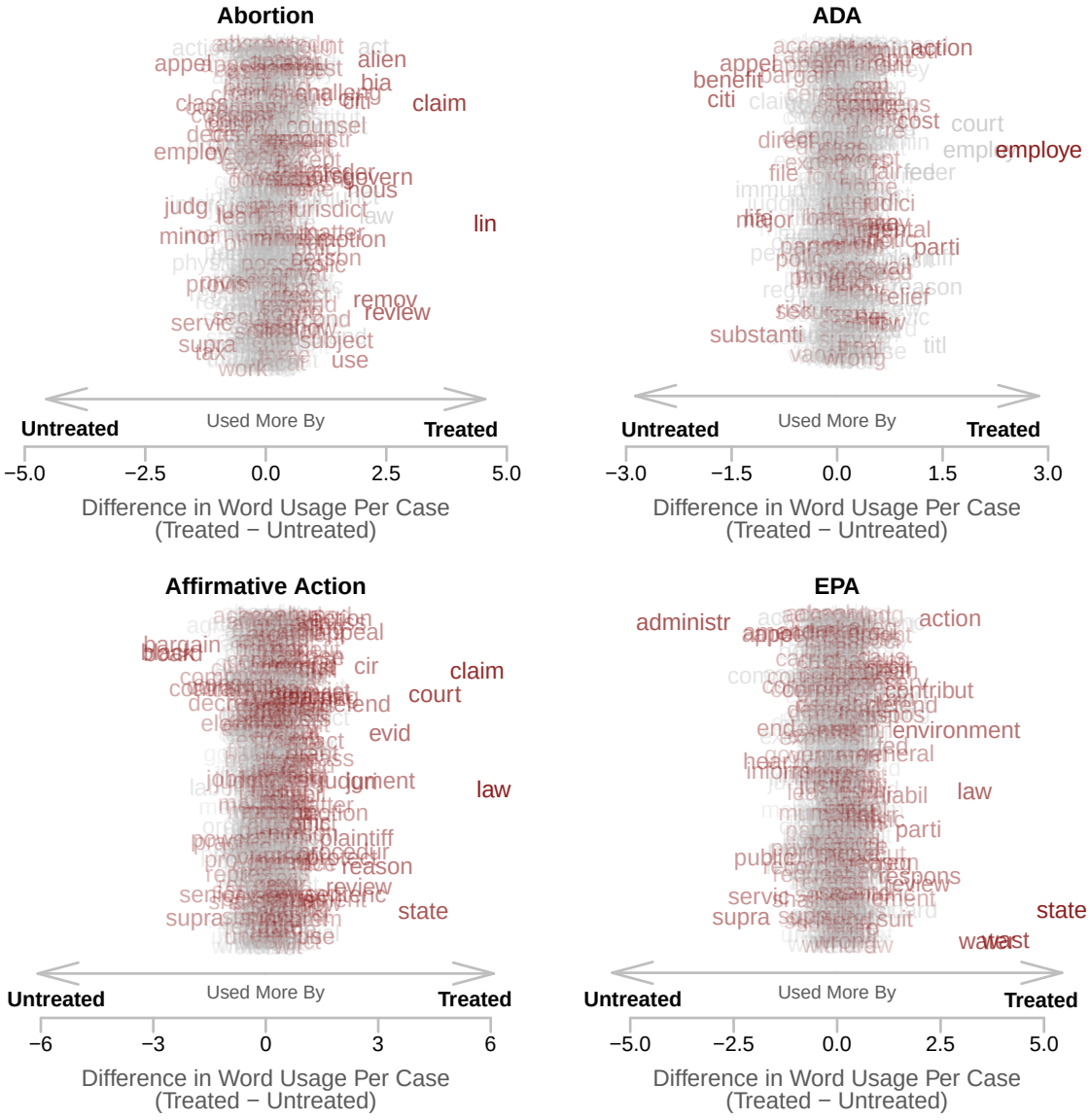


Figure 6 – Female Treatment: Differential Word Usage. Words are plotted in alphabetical order based on the estimated effect from Equation 2. Words to the right are used more often in treated panels; those to the left more often in control panels. Words in red are those for which the null hypothesis of no effect is rejected at the .05 level after Bonferroni correction.

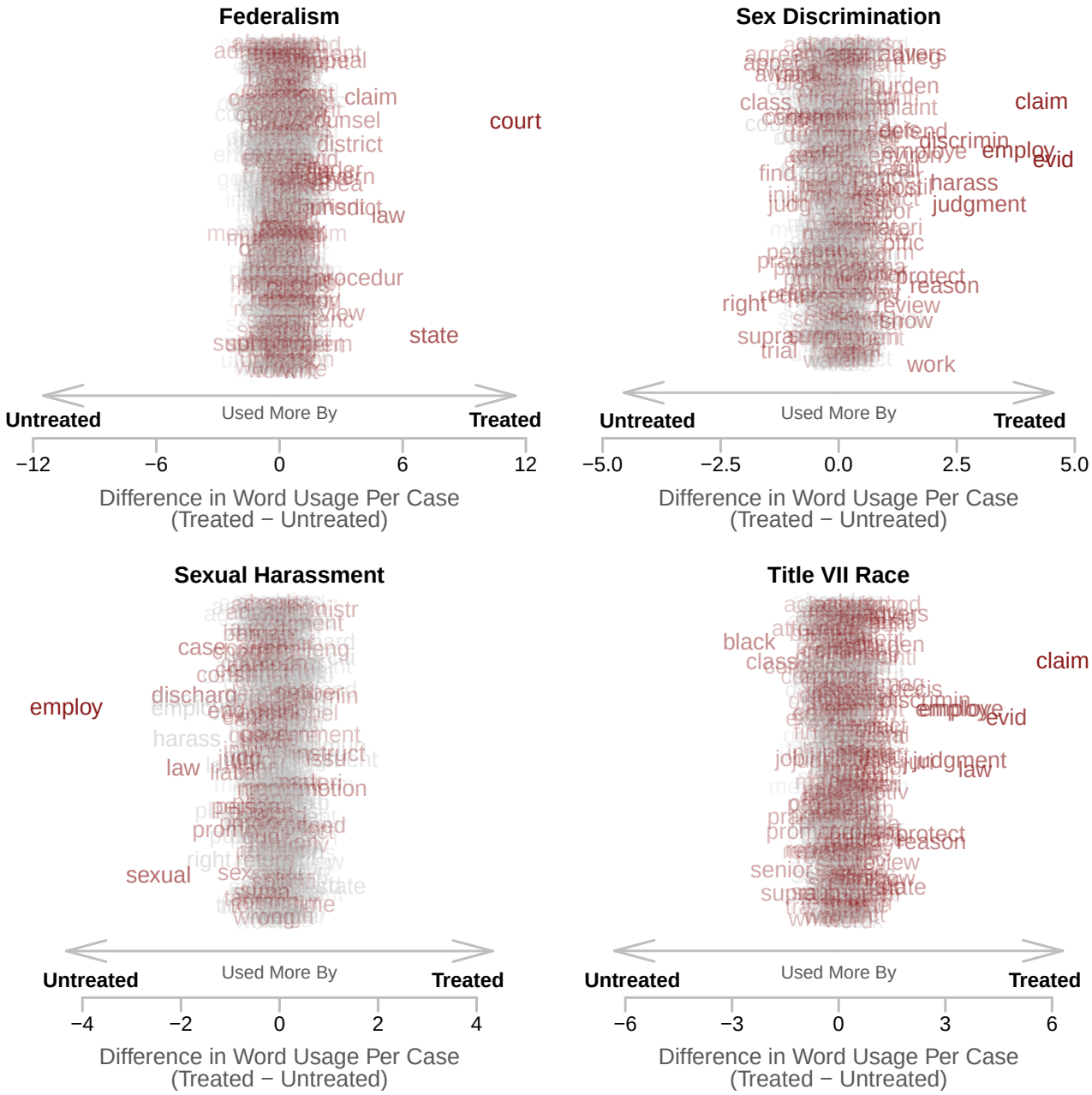


Figure 7 – Non-White Treatment: Differential Word Usage. Words are plotted in alphabetical order based on the estimated effect from Equation 2. Words to the right are used more often in treated panels; those to the left more often in control panels. Words in red are those for which the null hypothesis of no effect is rejected at the .05 level after Bonferroni correction.

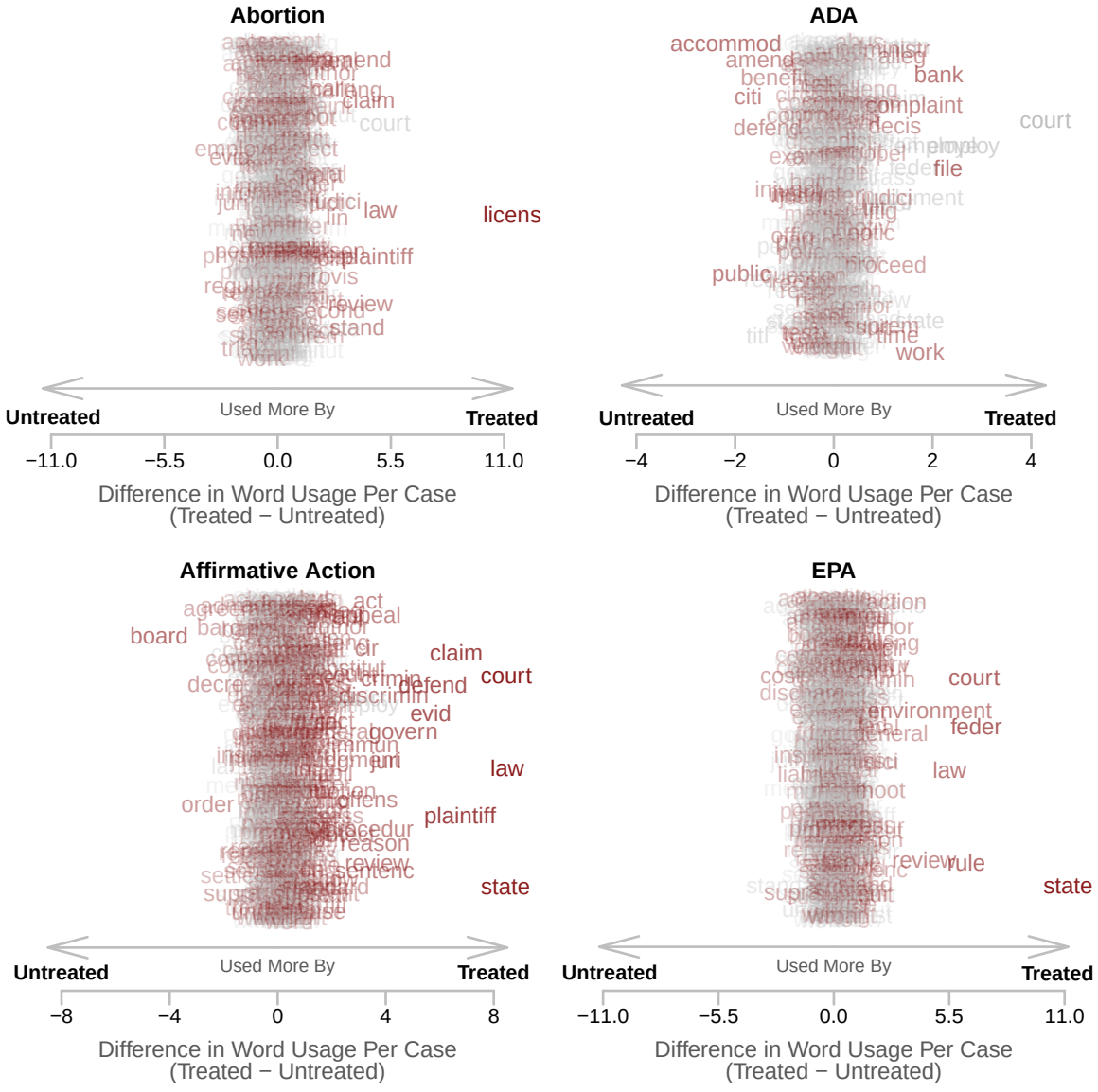
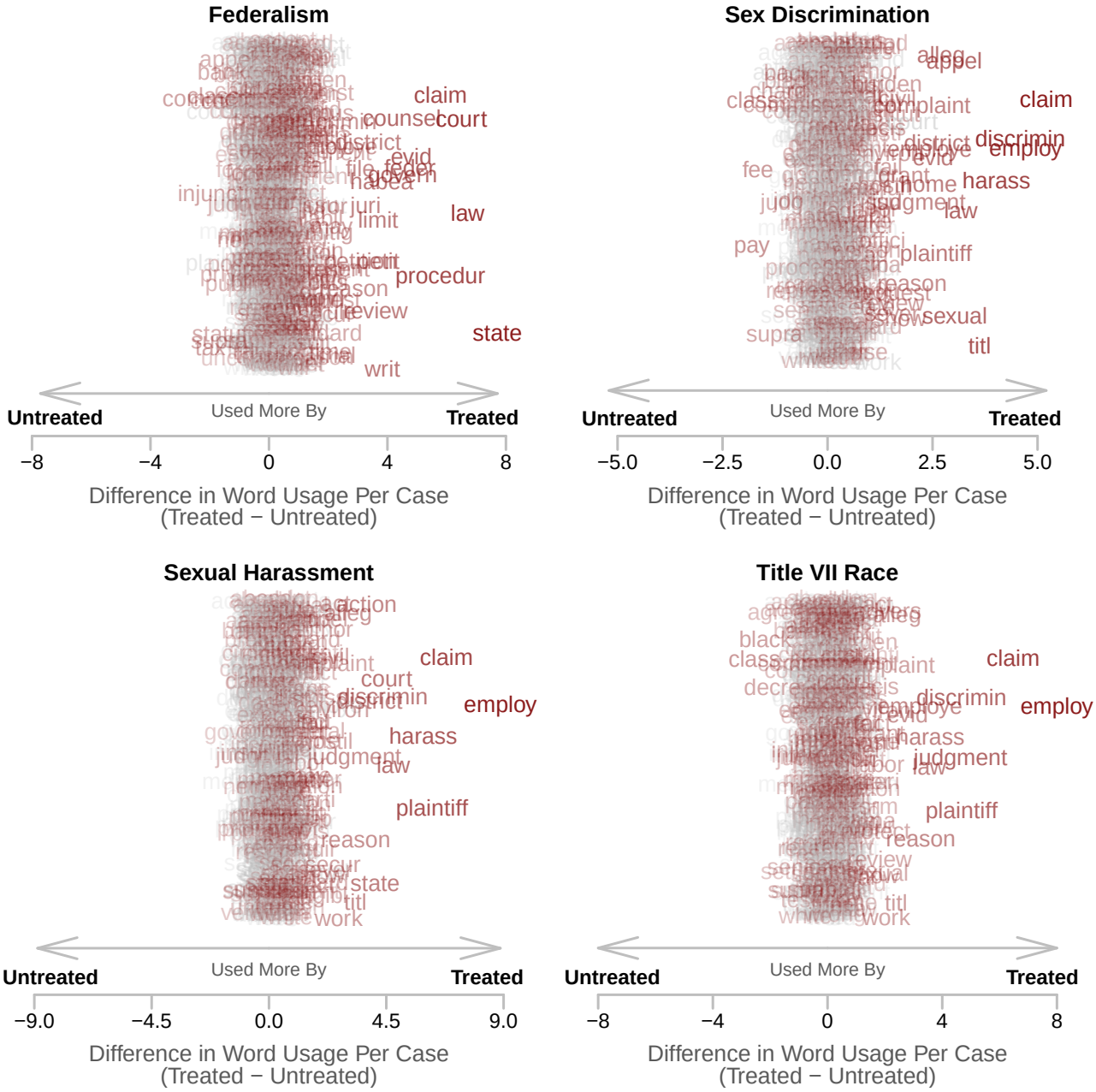


Figure 8 – Non-White Treatment: Differential Word Usage. Words are plotted in alphabetical order based on the estimated effect from Equation 2. Words to the right are used more often in treated panels; those to the left more often in control panels. Words in red are those for which the null hypothesis of no effect is rejected at the .05 level after Bonferroni correction.



For each issue area and each amendment, we record the presence of any citation to each amendment (rather than the count),[18] and we calculate the simple sample-size-weighted within-circuit average treatment effect as

$$\hat{\tau}_{ik}^{cite} = \sum_{j=1}^{J} \frac{n_{jk}}{N_k} (\bar{X}_{ijk}^1 - \bar{X}_{ijk}^0), \tag{3}$$

where $\bar{X}_{ijk}^1$ and $\bar{X}_{ijk}^0$ are the average appearance rates for amendment i for cases in issue area k in circuit j in the treated (1) and control (0) panels. To adjust for any possible heteroskedasticity, standard errors are estimated with a non-parametric bootstrap.

There are 27 amendments and 12 issue areas, which means that we estimate 324 separate treatment effects in this analysis. In the present context, the Bonferroni correction requires a p-value of approximately 0.00015 in order to reject the null hypothesis for a test at 0.05 significance. In addition, since there are so many tests and so much data, it is important to focus on the substantive size of effects, and not just their significance. Many effects, as will be seen, are discernible yet negligible in size. Others, however, are large and meaningful.

Figures 10 and 11 plot the estimated treatment effects. The numbers in the figure indicate the amendment, and the horizontal axis indicates the difference in probability of a citation between treated and control panels. Numbers and confidence intervals colored in red are those which are statistically significant after the Bonferroni correction is applied. The figures show some remarkable differences. To pick a few examples: panels with at least one female judge are more than 4 percentage points more likely to cite the first amendment in affirmative action cases; panels with at least one non-white judge are 2 percentage points less likely to cite the 14th amendment in Title VII race cases; and panels with at least one

[18] We have also analyzed the count data. Results are similar but the disadvantage is that the estimates are less interpretable (the difference between not citing and citing is clear, but the difference between citing twice and citing three times is less clear).

non-white judge are more than 4 percentage points more likely to cite the first amendment in abortion cases. Many other such differences appear in the results.

Results like these, by themselves, do not support broad conclusions about the direction of case dispositions. Panels might cite these amendments as support for a positive argument or during an explanation of why they are not applicable. Nevertheless, the random assignment of judges ensures that, on average, the case facts are the same across treated and control cases. The very fact that treated and control panels are seen to cite different amendments at differing rates, then, is strong evidence that the content of legal arguments depends, systematically, on the identities of the judges hearing the case.

Citations to Landmark Decisions

Using the same method from the previous analysis, we can also investigate the rates at which treated and control panels cite "landmark decisions." We define landmark decisions to be those listed as such by the Cornell Legal Information Institute. Figures 12 and 13 present the estimated results. Here, the numbers in the plot indicate the landmark category number, which range from 1-149. The Appendix provides a dictionary that explains what categories these numbers pertain to. Again, effects in red are those where the null of no effect is rejected at the .05 level after Bonferroni correction.

The figures show that, consistent with the previous analyses, the random assignment of a female or non-white judge causes major changes in landmark cases that judges cite. These citations are an important part of the legal arguments the panel presents, and as a consequence shape the impressions that future readers of the ruling receive.

Figure 9 – Amendment References by Issue Area. Each graph plots the raw count of references to each Constitutional Amendment (numbered 1-27) within each issue area.

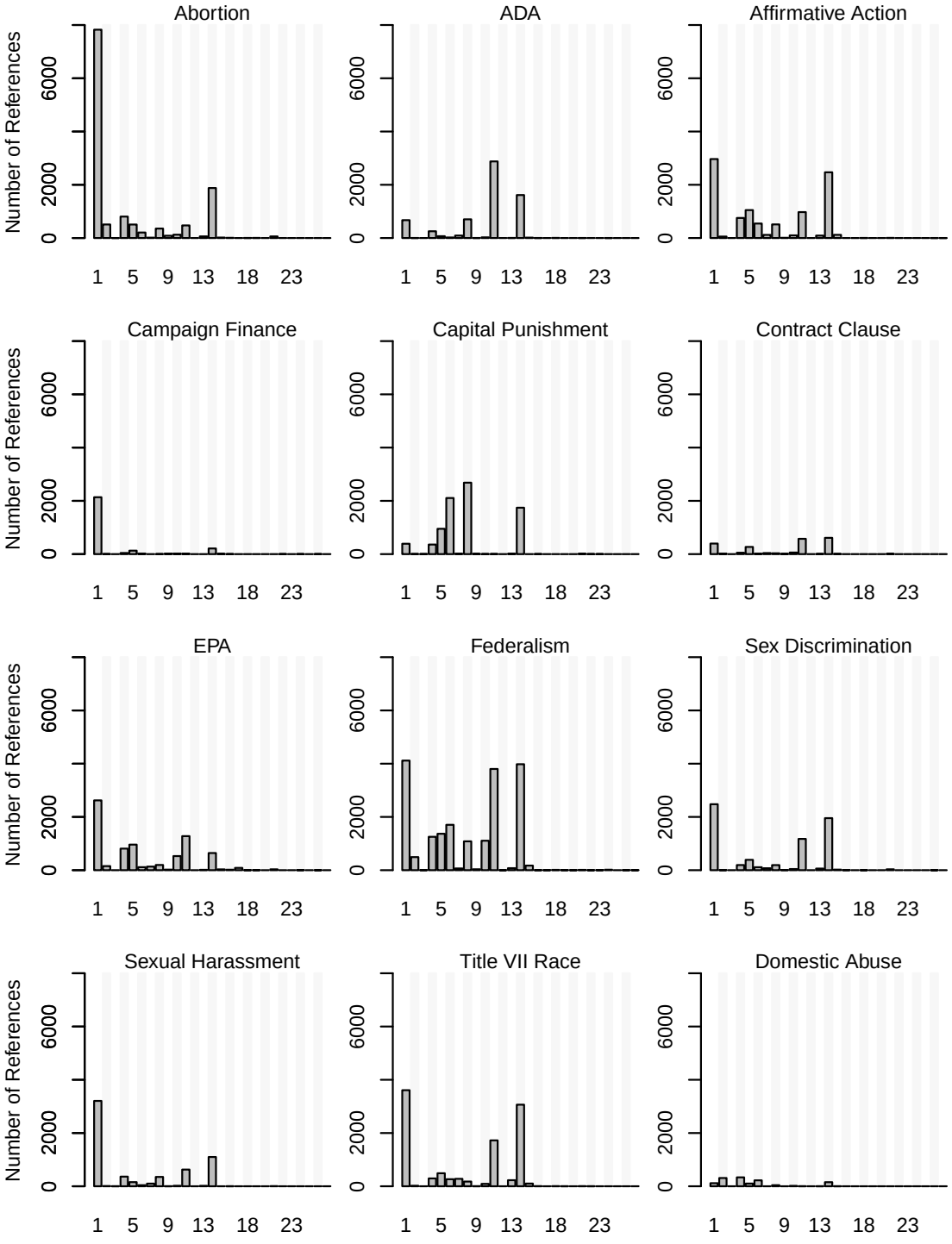


Figure 10 – Female Judges: Differential Reference to Constitutional Amendments. Plots the estimated within-circuit sample-size-weighted average treatment effects for each issue area and each constitutional amendment (1-27). Bars represent 95% confidence intervals from non-parametric bootstrap. Estimates in red are those where the null is rejected at the 0.05 level after Bonferroni correction for multiple testing.

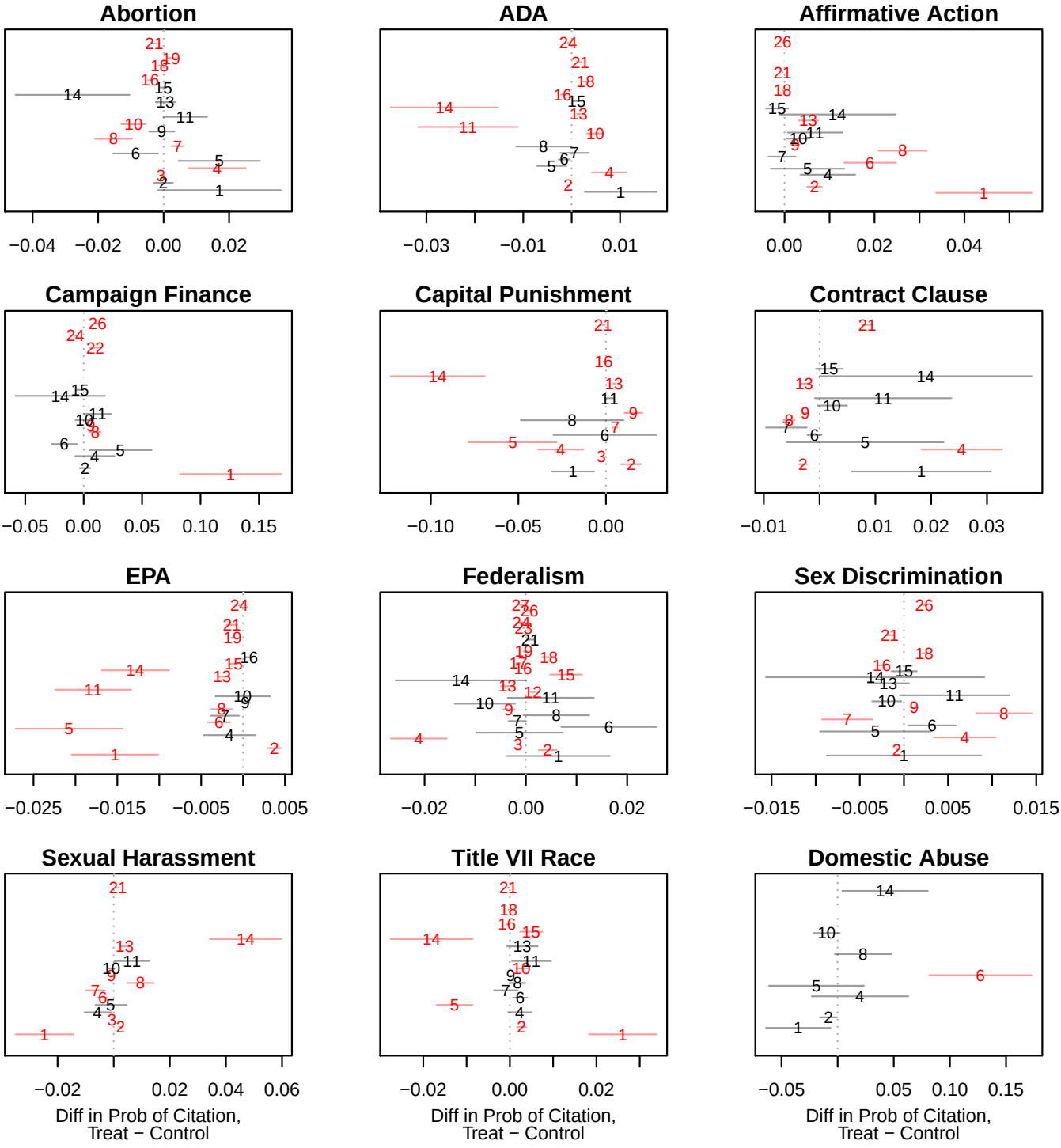


Figure 11 – Non-White Judges: Differential Reference to Constitutional Amendments. Plots the estimated within-circuit sample-size-weighted average treatment effects for each issue area and each constitutional amendment (1-27). Bars represent 95% confidence intervals from non-parametric bootstrap. Estimates in red are those where the null is rejected at the 0.05 level after Bonferroni correction for multiple testing.

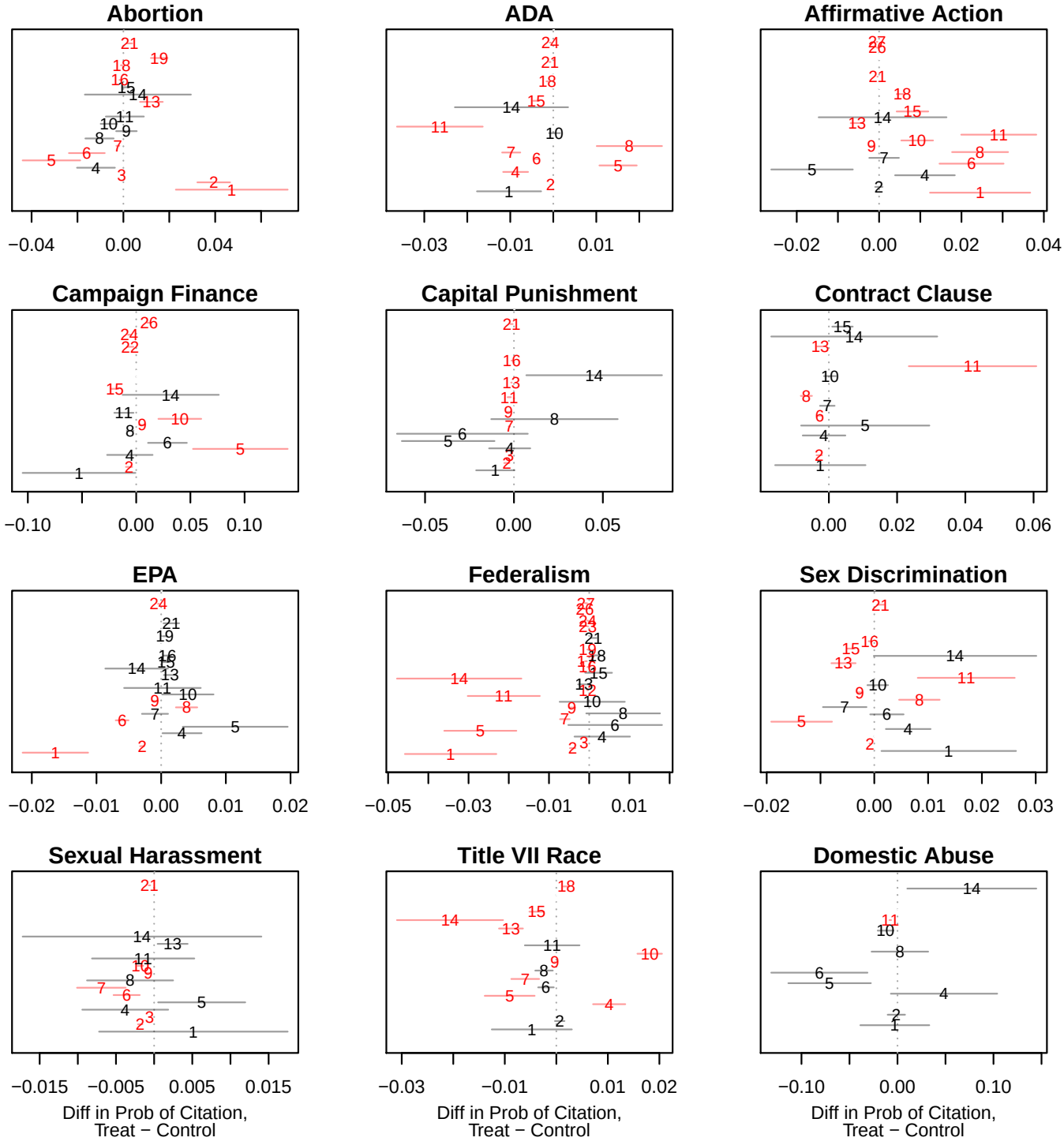


Figure 12 – Female Judges: Differential Citation of Landmark Decisions. Plots the estimated within-circuit sample-size-weighted average treatment effects for each issue area and each landmark decision category (numbered 1-149) from the Cornell Legal Information Institute. Category names provided in Appendix. Bars represent 95% confidence intervals from non-parametric bootstrap. Estimates in red are those where the null is rejected at the 0.05 level after Bonferroni correction for multiple testing.

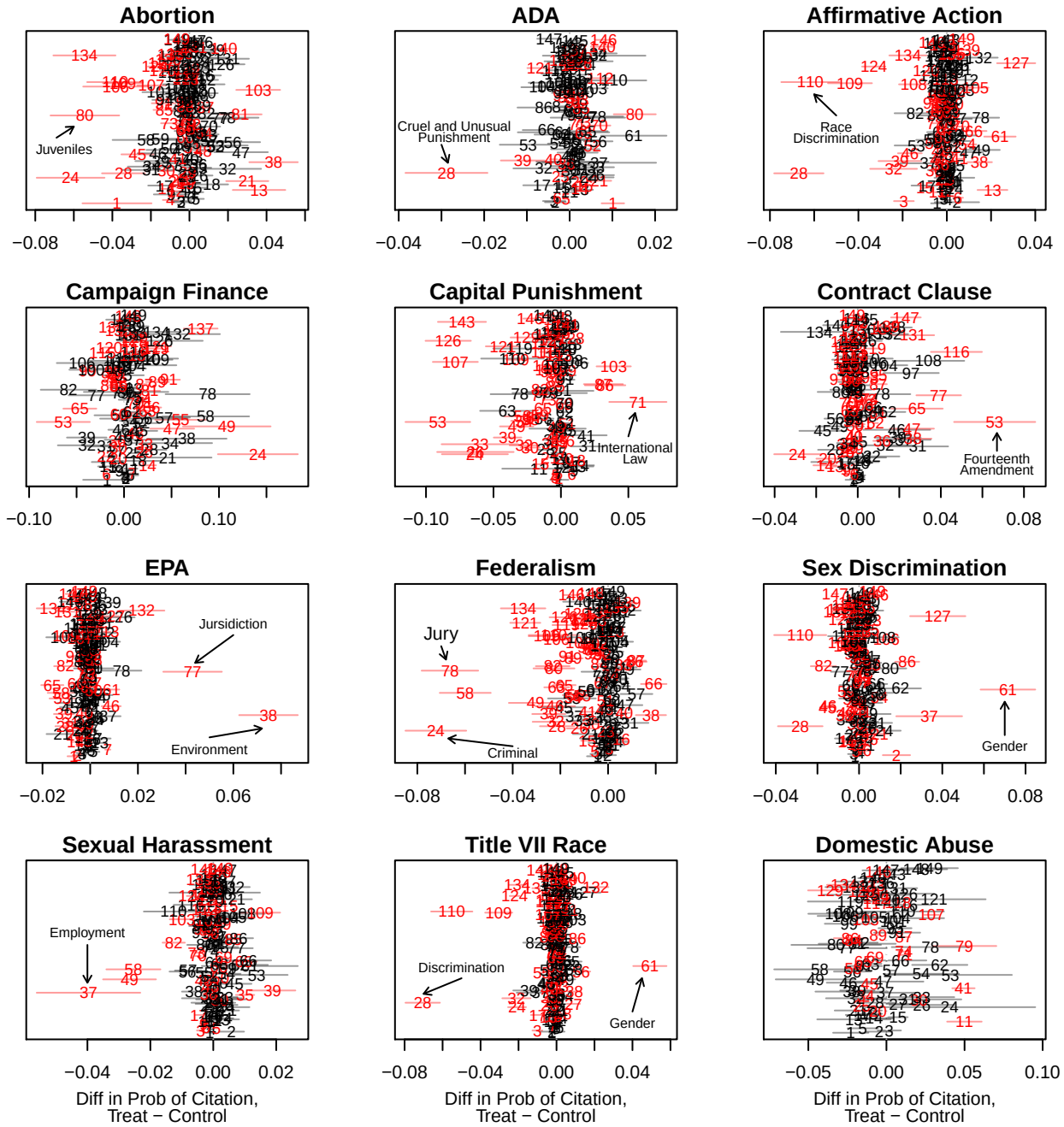
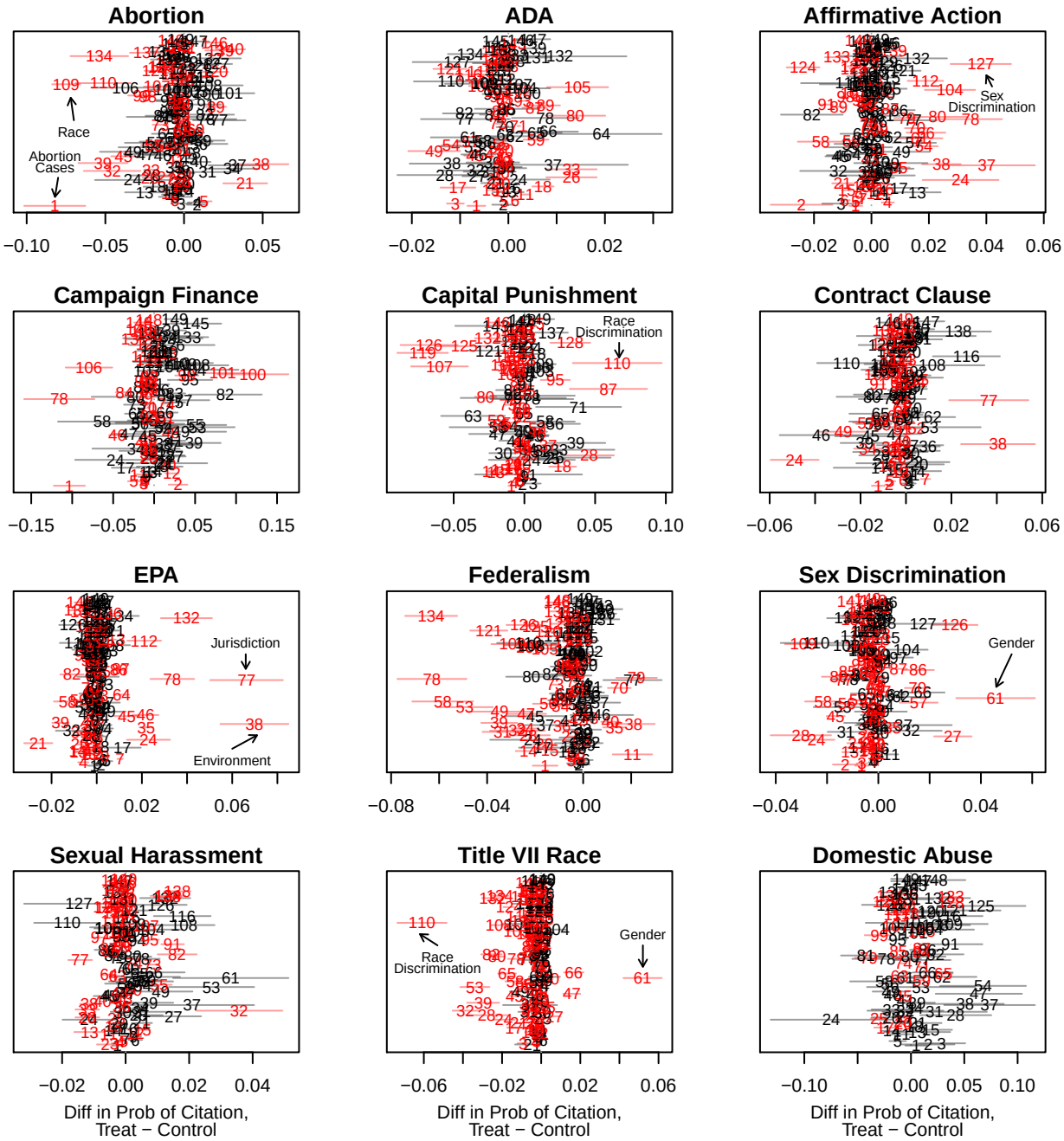


Figure 13 – Non-White Judges: Differential Citation of Landmark Decisions. Plots the estimated within-circuit sample-size-weighted average treatment effects for each issue area and each landmark decision category (numbered 1-149) from the Cornell Legal Information Institute. Category names provided in Appendix. Bars represent 95% confidence intervals from non-parametric bootstrap. Estimates in red are those where the null is rejected at the 0.05 level after Bonferroni correction for multiple testing.



Conclusion

The written word underpins our legal system, as it has in global legal systems since at least the time of Hammurabi. Typically, political science studies of legal activity focus on the disposition of cases—an area more directly conducive to empirical scrutiny, and one with great significance in and of itself. But in law, and especially in common-law systems, the words judges employ in their rulings echo long after a single case's disposition has been forgotten. In the American judicial context, legal authorities have understood this point at least since 1803, with the publication of Chief Justice John Marshall's opinion for *Marbury v. Madison*, which expressed: "Those who apply the rule to particular cases must of necessity expound and interpret that rule."[19] Legal interpretation requires a careful analysis of the language and construction of text, where "the ascertainment of the thought or meaning of the author of, or of the parties to, a legal document, as expressed therein, [are] according to the rules of language and subject to the rules of law" (Tiffany 1900).

It is for this reason that we have studied the composition of judicial text, where case outcomes from the US Courts of Appeals *are* institutionally operationalized as text. But more generally, legal text matters from a scholarly perspective because

> The precedents of the past, especially judicial precedents, come neatly packaged, with selected facts and authoritative language. Dealing with the use of past precedents thus requires dealing with the presence of the previous decision maker's words.(Schauer 1987: p. 573).

In this paper, we take advantage of modern advances in computation, data collection, and empirical methods to demonstrate a range of text-based causal effects resulting from the random assignment of female or non-white judges to cases. We provide three independent analyses to assess the effects of panel composition on legal outcomes. Each analysis provides separate insights relevant to the study of judicial outcomes. The first analysis, a study on

[19] 5 U.S. (1 Cranch) 137, 177 (1803).

the effect of raw word-usage, provides evidence for the effects of judicial identity on the the presence of unique words in legal writing on a micro-level, where the words included in the analysis were subset a priori according to a commonly-available legal dictionary. The second analysis, a study of references to constitutional amendments, provides insight into the types arguments and concepts relayed within each document, and how references to unique issues in constitutional law vary by randomly-determined panel composition. The third analysis, statistical tests over landmark case citations, provides evidence for how legal authorities link diverse topics in law into the articulation of a single case's ruling.

The random assignment of a female or non-white judge causes statistically-discernible changes in overall vocabulary, along with pronounced changes in meaningful legal words, references, and citation patterns. These effects appear strongest in issue areas salient to female and non-white judges, but are present across many issue areas. Going further, panel composition changes not only the usage of legal words, but also the rates at which constitutional amendments and landmark Supreme Court decisions are cited. We can infer from this that judicial identity matters for constructing the legal arguments that future lawyers, judges, and law students will study.

References

Angrist, Josuha D. and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ashenfelter, Orley, Theodore Eisenberg and Stewart J. Schwab. 1995. "Politics and the judiciary: The influence of judicial background on case outcomes." *The Journal of Legal Studies* 24(2):257–281.

Boos, Dennis. D. and L.A. Levanski. 2013. *Essential Statistical Inference: Theory and Methods*. Springer Texts in Statistics Springer.

Boyd, Christina L., Lee Epstein and Andrew D. Martin. 2010. "Untangling the causal effects of sex on judging." *American Journal of Political Science* 54(2):389–411.

Bueno De Mesquita, Ethan and Matthew Stephenson. 2002. "Informative Precedent and Intrajudicial Communication." *American Political Science Review* 96(04):755–766.

Cross, Frank B. 2007. *Decision Making in the U.S. Courts of Appeals*. Palo Alto, CA: Stanford University Press.

Demsar, Janez. 2006. "Statistical Comparisons of Classifiers over Multiple Data Sets." *Journal of Machine Learning Research* 7:1–30.

Ding, Peng. 2013. "Demystifying Statistical Paragons' Paradox from Causal Inference in Randomized Experiments." *Working Paper* .

Dunn, Olive Jean. 1961. "Multiple Comparisons among Means." *Journal of the American Statistical Association* 56(293):52–64.

Fisher, R.A. 1935. *The Design of Experiments.* 8th ed. New York: Hafner Press.

George, Tracey E. and Lee Epstein. 1992. "On the Nature of Supreme Court Decision Making." *The American Political Science Review* 86(2):323–337.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton & Company.

Good, Phillip I. 2005. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. 3rd ed. Springer.

Greiner, D James. and Donald B. Rubin. 2011. "Causal Effects of Perceived Immutable Characteristics." *Review of Economics and Statistics* 93(3):775–785.

Grimmer, Justin and Gary King. 2011. "A General Purpose Computer-Assisted Clustering Methodology." *Proceedings of the National Academy of Sciences* .

Haire, Susan B. and Laura P. Moyer. 2007. "Advocacy Through Briefs in the US Courts of Appeals." *S. Ill. ULJ* 32:593.

Holland, P.W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81:945–968.

Hopkins, Daniel J. and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1):229–247.

Imbens, Guido W. and Donald B. Rubin. 2010. *Causal Inference in Statistics and Social Sciences*. Unpublished Manuscript.

Kastellec, Jonathan P. 2013. "Racial Diversity and Judicial Influence on Appellate Courts." *American Journal of Political Science* 57(1):167–183.

Keele, Luke, Corrine McConnaughy and Ismail White. 2012. "Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity." *American Journal of Political Science* 56(2):484–499.

Knight, Jack and Lee Epstein. 1996. "The Norm of Stare Decisis." *American Journal of Political Science* 40(4):1018–1035.

Kritzer, Herbert M. and Mark J. Richards. 2003. "Jurisprudential Regimes and Supreme Court Decisionmaking: The Lemon Regime and Establishment Clause Cases." *Law & Society Review* 37(4):827–840.

Lauderdale, Benjamin E. and Tom S. Clark. 2012. "The Supreme Court's Many Median Justices." *American Political Science Review* 106(4):847–866.

Lehmann, Erich L. 2006. *Nonparametrics: Statistical Methods Based on Ranks*. Vol. Revised First Edition Springer.

Moyer, Laura P. 2012. "The Role of Case Complexity in Judicial Decision Making." *Law & Policy* 34(3):291–312.

Pitman, Edwin J. G. 1937. "Significance tests which may be applied to samples from any populations." *Supplement to the Journal of the Royal Statistical Society* 4(119–130).

Richards, Mark J. and Herbert M. Kritzer. 2002. "Jurisprudential Regimes in Supreme Court Decision Making." *The American Political Science Review* 96(2):305–320.

Rosenbaum, Paul R. 2010. *Observational Studies*. 2nd ed. Springer.

Scalia, Antonin and Bryan A. Garner. 2012. *Reading Law: The Interpretation of Legal Texts*. West: Thomson-Reuters.

Schauer, Frederick. 1987. "Precedent." *Stanford Law Review* 39(3):571–605.

Sen, Maya. 2012. "Is Justice Really Blind? Race and Appellate Review in U.S. Courts." *Working Paper* .

Sunstein, Cass R., David Schkade, Lisa M. Ellman and Andres Sawicki. 2006. "Are Judges Political? An Empirical Analysis of the Federal Judiciary." pp. 1–190.

Tiffany, H.T. 1900. Interpretation and Construction. In *American and English Encyclopedia of Law*, ed. David S. Garland and Lucious P. McGehee. 2d ed. 17 1, 2.

Appendix

Figure 14 – Female Treatment: Differential Word Usage. Words are plotted in alphabetical order based on the estimated effect from Equation 2. Words to the right are used more often in treated panels; those to the left more often in control panels. Words in red are those for which the null hypothesis of no effect is rejected at the .05 level after Bonferroni correction.

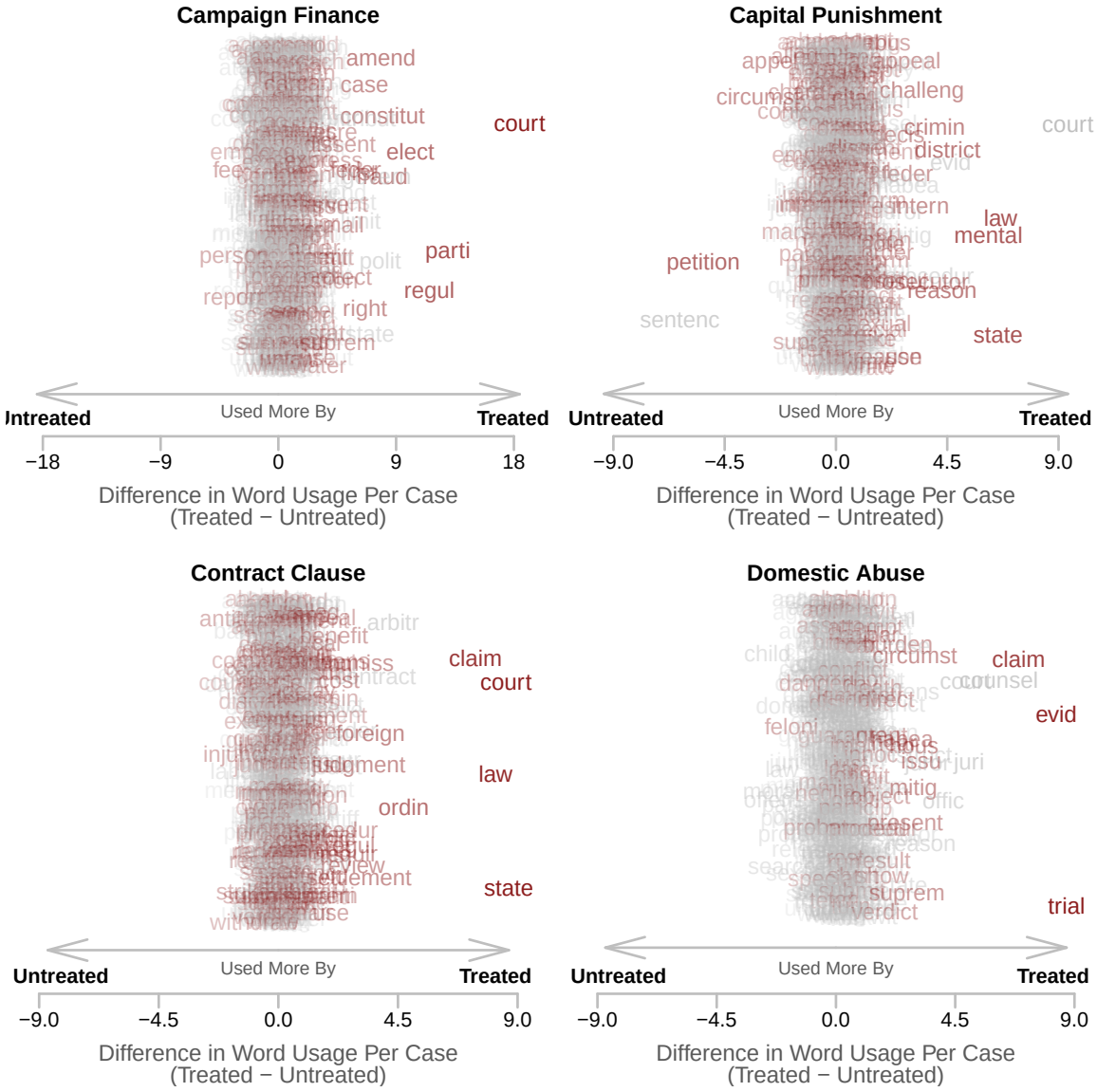


Figure 15 – Non-White Treatment: Differential Word Usage. Words are plotted in alphabetical order based on the estimated effect from Equation 2. Words to the right are used more often in treated panels; those to the left more often in control panels. Words in red are those for which the null hypothesis of no effect is rejected at the .05 level after Bonferroni correction.

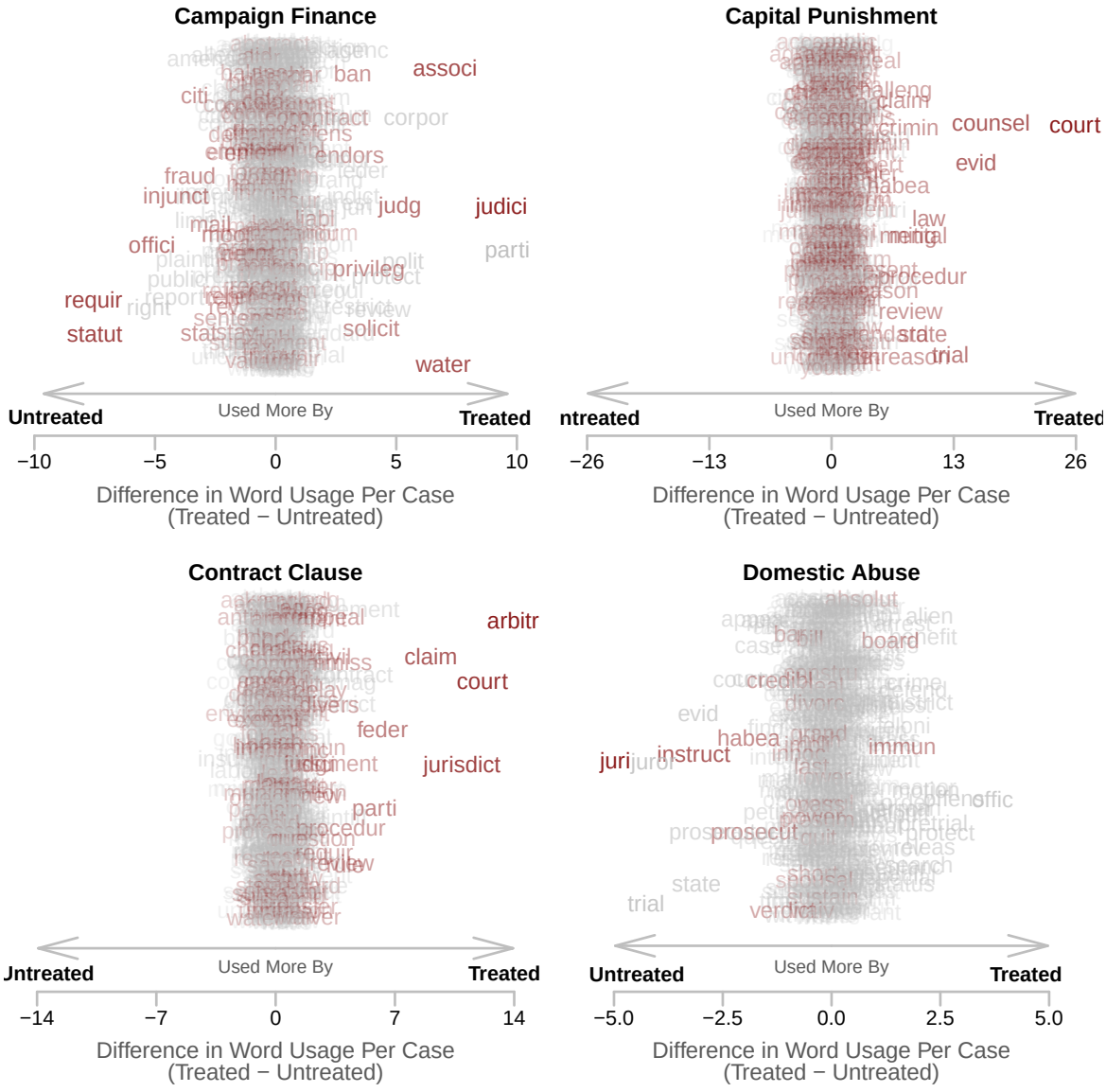


Figure 16 – List of Landmark United States Supreme Court Topics. Case types taken from the Legal Information Institute at Cornell University Law School. Topic numbers are noted to the left of each issue area. To the right of each issue area is the number of landmark Court decisions listed in that issue area.

Source: http://www.law.cornell.edu/supct/cases/topic.htm

Landmark Area

	# Cases		# Cases
1) Abortion	15	76) Judicial review	12
2) Affirmative action	12	77) Jurisdiction	69
3) Aliens	9	78) Jury	23
4) Armed services	5	79) Justiciability	48
5) Attainder	7	80) Juveniles	19
6) Attorneys	6	81) Labor	36
7) Bankruptcy	1	82) Legislative policy	4
8) Bill−of−rights	1	83) Libel	5
9) Borders	1	84) Marriage	3
10) Capital punishment	28	85) Mental health	6
11) Censorship	3	86) Mental retardation	3
12) Children	22	87) Minimum contacts	1
13) Choice of law	6	88) Monopoly	9
14) Citizenship	15	89) National power	5
15) Commander in chief	5	90) National security	24
16) Commerce clause	46	91) Native americans	8
17) Commercial speech	6	92) Necessary−and−proper	1
18) Confessions	3	93) New deal	2
19) Conflict of laws	2	94) Ninth amendment	1
20) Congress	32	95) Obscenity	19
21) Contraception	4	96) Pardon	2
22) Contract clause	8	97) Pensions	5
23) Courts	26	98) Pledge of loyalty	4
24) Criminal	193	99) Police power	12
25) Criminal procedure	8	100) Political questions	12
26) Cruel and unusual punishment	29	101) Political speech	8
27) Damages	3	102) Power to tax and spend	8
28) Discrimination	95	103) Precedent	10
29) Discrimination based on national origin	1	104) Presidency	37
30) Double jeopardy	7	105) Prison	6
31) Due process	34	106) Privacy	21
32) Education	69	107) Privileges and immunities	15
33) Eighth amendment	28	108) Property	38
34) Elections	23	109) Race	77
35) Eleventh amendment	6	110) Race discrimination	68
36) Eminent domain	3	111) Reapportionment	7
37) Employment	52	112) Regulation	21
38) Environment	9	113) Removal power	3
39) Equal protection	47	114) Reproduction	1
40) Establishment of religion	12	115) Res judicata	3
41) Evidence	8	116) Right to a hearing	11
42) Executive power	1	117) Right to bear arms	1
43) Executive priviledge	2	118) Right to confront witnesses	3
44) Extradition	2	119) Right to counsel	17
45) Federal courts	25	120) Right to travel	7
46) Federalism	30	121) Searches and seizures	49
47) Fifth amendment	29	122) Second amendment	1
48) Fighting words	2	123) Sedition	5
49) First amendment	128	124) Segregation	4
50) Flag desecration	5	125) Self−incrimination	20
51) Foreign affairs	2	126) Separation of powers	31
52) Forum	5	127) Sex discrimination	15
53) Fourteenth amendment	57	128) Sexuality	3
54) Fourth amendment	33	129) Sixth amendment	9
55) Freedom of assembly	11	130) Slavery	5
56) Freedom of association	34	131) Social security	10
57) Freedom of religion	49	132) Standing	21
58) Freedom of speech	116	133) State action	9
59) Freedom of the press	29	134) States	53
60) Full faith and credit	2	135) Sterilization	2
61) Gender	14	136) Supremacy clause	7
62) Government employment	13	137) Symbolic speech	8
63) Habeas corpus	15	138) Takings clause	9
64) Handicapped	3	139) Tax	15
65) Housing	8	140) Tenth amendment	21
66) Immunity	12	141) Testimony	1
67) Implied−powers	1	142) Thirteenth amendment	2
68) Import tariffs	1	143) Trial by jury	7
69) Incorporation	3	144) Veto	1
70) Insanity	4	145) Voting	20
71) International law	4	146) War powers	7
72) International relations	14	147) Welfare benefits	6
73) Internet	1	148) Wiretapping	6
74) Investigations	3	149) Witnesses	8
75) Involuntary servitude	2		