Untangling the Causal Effects of Sex on Judging

Christina L. Boyd University at Buffalo, SUNY
Lee Epstein Northwestern University School of Law
Andrew D. Martin Washington University in St. Louis

We explore the role of sex in judging by addressing two questions of long-standing interest to political scientists: whether and in what ways male and female judges decide cases distinctly—"individual effects"—and whether and in what ways serving with a female judge causes males to behave differently—"panel effects." While we attend to the dominant theoretical accounts of why we might expect to observe either or both effects, we do not use the predominant statistical tools to assess them. Instead, we deploy a more appropriate methodology: semiparametric matching, which follows from a formal framework for causal inference. Applying matching methods to 13 areas of law, we observe consistent gender effects in only one—sex discrimination. For these disputes, the probability of a judge deciding in favor of the party alleging discrimination decreases by about 10 percentage points when the judge is a male. Likewise, when a woman serves on a panel with men, the men are significantly more likely to rule in favor of the rights litigant. These results are consistent with an informational account of gendered judging and are inconsistent with several others.

Ever since Jimmy Carter set out to diversify the federal bench, scholars have been exploring the effects of sex on judging. The result is now a voluminous body of literature,[1] which focuses on two chief questions: whether and in what ways male and female judges decide cases distinctly—"individual effects"—and whether and in what ways serving with a female judge causes males to behave differently—"panel effects."[2]

We too take up these important questions. In so doing, we follow the lead of others writing in this area and attend to the dominant extant accounts of why we might expect to observe either or both sex-based effects, including accounts that stress information, representation, and socialization. We depart from existing work in two ways. First, while most studies explore sex-based effects in a limited number of legal areas, we examine 13, ranging from disability law to piercing the corporate veil to, of course, sex discrimination. Analyzing a diverse set of disputes, we believe, permits a more comprehensive assessment of the implications of the various theoretical accounts. Second, while most previous work relies on variants of standard regression analysis, we turn instead to semiparametric matching methods, which follow from

Christina L. Boyd is Assistant Professor in the Department of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (cLboyd@buffalo.edu). Lee Epstein is Henry Wade Rogers Professor, Northwestern University School of Law, 357 E. Chicago Ave., Chicago, IL 60611 (lee-epstein@northwestern.edu). Andrew D. Martin is Professor in the Department of Political Science and School of Law, Washington University in St. Louis, Campus Box 1063, One Brookings Drive, St. Louis, MO 63130 (admartin@wustl.edu).

Winner of the 2008 Pi Sigma Alpha award for the best paper presented at the annual meeting of the Midwest Political Science Association. We thank the Center for Empirical Research in the Law, the Weidenbaum Center at Washington University, the National Science Foundation, the Baldy Center for Law & Social Policy, and the Northwestern University School of Law for supporting our research; Cass Sunstein, David Schkade, Lisa M. Ellman, and Andres Sawicki for sharing their data; Shari Diamond, Sarah Fischer, William Landes, Kevin Quinn, Richard Posner, Nancy Staudt, Kim Yuracko, the editor and anonymous reviewers of the *American Journal of Political Science,* and participants at faculty workshops at Dartmouth College, Stony Brook University, the University of Chicago, the University of Illinois, the University of Pennsylvania, and Washington University for providing useful comments; and Delia Bailey, Kathryn Jensen, Hyung Kim, Zachary Levinson, Jessica Silverman, and Jennifer Solomon for supplying excellent research assistance. The project's web site (http://epstein.law.northwestern.edu/research/genderjudging.html) houses a full replication archive, including the data and documentation necessary to reproduce our results.

[1] An appendix on our web site describes the results of some 30 studies on the topic. We should note that our focus is on sex, but, of course, the federal bench has been diversified on the dimensions of race and color. The methodological approach we advocate here would be equally suitable for exploring the effect of these characteristics on judges or, for that matter, legislators, advisors, attorneys, litigants, and voters.

[2] Our phrasing is not accidental. For the reasons we supply in the second section, only the second question lends itself to causal inference.

American Journal of Political Science, Vol. 54, No. 2, April 2010, Pp. 389–411

 ISSN 0092-5853

a formal framework of causal inference. For the reasons we outline below, these tools are better suited to the twin tasks at hand: estimating individual and panel effects on the federal appellate bench.

Our application of these methods unearths neither individual nor panel effects in 12 of the 13 areas of the law. Only in cases implicating sex discrimination do we observe sex-based effects: the probability of a judge deciding in favor of the party alleging discrimination decreases by 10 percentage points when the judge is a male. Likewise, when a woman serves on a panel with men, the men are significantly more likely to rule in favor of the rights litigant. More generally, our findings are consistent with informational accounts of gendered judging and are inconsistent with several others.

Seen in this way, our study adds theoretically and substantively to a burgeoning body of literature of interest to social scientists, judges, and policy makers alike. Given that our results reinforce the findings in several existing studies (e.g., Crowe 1999; Davis, Haire, and Songer 1993; Peresie 2005), however, our most important contribution may be methodological. The matching methods we deploy here hold a good deal of promise, we believe, to advance our understanding of judicial behavior—not to mention of sex (and race) effects in the other institutions of government.

What We Know about How Women and Men Judge

Almost from the day Justice O'Connor announced her retirement from the U.S. Supreme Court, pressure mounted on President George W. Bush to nominate a woman. Various news sources reported that elites on the left and right thought the seat should be "reserved" for a female, and the public concurred. Even the first lady ventured an opinion, saying that she "would really like [the President] to name another woman to the Supreme Court."

Whether Bush acceded to this pressure with his (unsuccessful) nomination of Harriet Miers is a matter of some debate. But the entire episode raises the question of why the pressure was there in the first place: why did elites and the public alike support appointing a woman to replace O'Connor? One answer centers on "social legitimacy," or the belief that "democratic institutions in heterogeneous societies ought to reflect the make-up of society" (Cameron and Cummings 2003, 28). On this account, elected officials should work to ensure the commensurate representation of women on the nation's highest court in part because they now constitute over one-half of the U.S. population and nearly one-third of all lawyers in the country.[3]

Another set of responses centers less on the sheer presence of female judges and more on "their participation and their perspective" (Sherry 1986); that is, on whether males and females behave differently (individual effects) and whether females influence their male colleagues (panel effects). Falling into this set, as we show in Table 1, are different voice, representational, informational, and organizational accounts of sex-based judging. Note that while three of the four posit differences in the behavior of male and female judges, their underlying mechanisms and, ultimately, their empirical implications, are distinct.

In light of the prominence of these accounts—one or more appears in virtually every study of gendered judging (see, e.g., Baldez, Epstein, and Martin 2006; Brudney, Schiavoni, and Merrit 1999; Clark 2004; Farhang and Wawro 2004; Martin, Reynolds, and Keith 2002; Peresie 2005; Sherry 1986; Sullivan 2002)—they require little elaboration. Briefly, the first, the *different voice* approach, follows from Gilligan's (1982) seminal work.[4] This account stresses divergencies between males and females—primarily that they develop distinct worldviews and see themselves as differentially connected to society. As a result, we would not expect much in the way of panel effects; given their differences, male and female judges are unlikely to influence one another. Individual effects, however, should be quite extensive, emerging across virtually all areas of the law. Indeed, if Gilligan's work has any implications for judging, it is that female judges bring a "feminine perspective" to the bench—one that "encompasses all aspects of society, whether or not they affect men and women differently," and not only "the political agenda associated with feminism" (Sherry 1986, 160; see also Davis 1992; Steffensmeier and Herbert 1999).

For *representational* accounts, that "political agenda" moves to the fore. The idea here, tracing to Pitkin's

[3] Other forms of this argument center on the "inherent unfairness" of only men occupying seats of power; on the desirability of input from all parts of a diverse society; and on the courts' need for legitimacy, which cannot be achieved if a "segment of the population is excluded from membership" (see, e.g., Epstein, Knight, and Martin 2003; Maule 2000, 296–97).

[4] *In a Different Voice* has faced its share of criticism on any number of grounds—sociological, biological, psychological, and methodological. And yet, as Beiner writes, despite the critiques, Gilligan's "theory no doubt continues to be taught, discussed, and tested because something about it rings true, or at least true based on some stereotyped notion of the way in which women behave" (2002, 602). Based on our inventory of the literature, Beiner has it exactly right.

TABLE 1 Accounts of Sex Effects on Judging

Account	Premise	Empirical Implications	
		Individual Effects	Panel Effects
Different Voice	Males and females develop distinct worldviews and see themselves as differentially connected to society	Yes, across a range of issues	None expected
Representational	Female judges serve as representatives of their class and so work toward its protection in litigation of direct interest	Yes, but only on issues of concern to women broadly speaking	None expected
Informational	Women possess unique and valuable information emanating from shared professional experiences	Yes, but only on issues on which female judges may possess valuable expertise, experience, or information	Yes, but only on issues on which female judges may possess valuable expertise, experience, or information
Organizational	Male and female judges undergo identical professional training, obtain their jobs through the same procedures, and confront similar constraints once on the bench	No. Male and female judges are more alike than dissimilar and face common professional constraints	None expected

"Individual Effects" are whether and in what ways male and female judges decide cases distinctly; "Panel Effects" are whether and in what ways serving with a female judge causes her male colleagues to behave differently.

(1967) work, is that female judges serve as representatives of their class and work toward its protection in litigation of direct interest—or, as Cook famously put it, "the organized campaign to place more women on the bench rest[ed] on the hope that women judges will seize decision-making opportunities to liberate other women" (1981, 216; see also, e.g., Allen and Wall 1993; Martin and Pyle 2005; Tobias 1990).[5] Consequently, this account too posits individual effects, but they should manifest themselves in a smaller set of cases—only those involving issues "where the policy consequences are likely to have immediate and direct impact on significantly larger numbers of women than men" (Carroll 1984, 308). Common examples of such "women's issues" in the law include abortion, affirmative action, sex discrimination in employment, and sexual harassment.[6]

To the extent that *informational* accounts suggest the emergence of individual effects in a few legal areas, they converge with representational theories. But the similarities end there. The logic behind informational or expertise approaches is not that women represent a particular class but rather that they possess unique and valuable information emanating from shared professional experiences (Cameron and Cummings 2003; Gryski, Main, and Dixon 1986; Peresie 2005). Accordingly, sex-based effects are likely to manifest themselves in an even more circumscribed set of cases—primarily sex discrimination in the employment context.[7] But the effects themselves

[5]Some recent research on world legislatures has found that women are not always alone in advocating for women's issues and interests (e.g., Dahlerup 2006).

[6]Worth noting is the existence of a robust debate over what constitutes a women's issue (compare, e.g., Thomas 1994 and Reingold 2000) such that some analysts would dispute the categories we list in the text. Within the literature on judging, however, it is not uncommon to adopt a rather narrow definition of "women's issue," as we do here (see, e.g., Martin and Pyle 2000; Segal 2000; Walker and Barrow 1985).

[7]When presenting our paper to various professional audiences, interesting debates ensued over whether we should limit the empirical implication here to sex-based employment discrimination or expand it to include abortion and sexual harassment as well. Those advocating greater inclusiveness emphasize that female judges may have stronger priors as a result of their experience with harassment or abortion. Those advocating less inclusiveness suggest that only

are likely to be broader, not only increasing the odds of a pro-plaintiff decision by female judges in employment litigation but also by the male judges with whom they sit. The reason is straightforward enough: because, under this approach, female judges possess information that their male colleagues perceive "as more credible and persuasive" than their own knowledge about sex discrimination, females can directly or even indirectly alter the choices made by males (i.e., induce them to decide sex discrimination cases differently than they otherwise would; Peresie 2005, 1783; see also, e.g., Baldez, Epstein, and Martin 2006; Cameron and Cummings 2003; Ostberg and Wetstein 2007; Sullivan 2002).[8]

Finally, we turn to approaches that emphasize the commonalities between male and female judges, or what some call *organizational* accounts (e.g., Steffensmeier and Herbert 1999). While not necessarily denigrating the importance of diversity for, say, promoting social legitimacy, these analysts suggest that we are unlikely to observe any sex-based effects in the courts. After all, they argue, male and female judges undergo identical professional training, obtain their jobs through the same procedures, and confront similar constraints once on the bench (see, e.g., Kritzer and Uhlman 1977; Sisk, Heise, and Morriss 1998). These commonalities should be sufficient "to overcome any biological, psychological, or experienced-based differences between the sexes" (Steffensmeier and Herbert 1999, 1165).

However different these accounts (and however distinct their empirical implications), scholars have devised remarkably similar designs and employed nearly identical methods to explore them. Virtually all quantitative work in this area:

1. asks the same research questions: Does gender *cause* judges to behave differently (individual effects)? And, more recently, does the presence of a female judge *cause* male judges to act differently (panel effects)?;
2. makes use of a dichotomous regression model (typically logit or probit), with the judge's vote (e.g., for or against the plaintiff in sex discrimination cases) serving as the dependent variable;
3. captures the effect of sex in the same way, as a dummy variable for the sex of the judge (for individual effects) or a series of dummy variables for the sex of panel members (for panel effects); and
4. attends to (approximately) the same covariates (i.e., confounding factors), chiefly attributes of the judge (e.g., ideology, age, judicial experience, race) and characteristics of the case (e.g., direction of lower court decision, year of decisions).

Despite the similarities in approach, the resulting research findings have been somewhat mixed. By our count, social scientists and legal academics have produced over 30 systematic, multivariate analyses of the extent to which female judges make decisions distinct from their male colleagues (individual effects) or cause male judges to behave differently than they otherwise would (panel effects).[9] Of these, roughly one-third purport to demonstrate clear panel or individual effects, a third report mixed results, and the final third find no sex-based differences whatsoever.

Drawing Causal Inferences about Sex and Judging

Why the mixed findings is of less immediate interest to us than the question of how best to isolate sex effects, if in fact they exist. In what follows, we undertake this challenge,

in the area of employment discrimination are female judges likely to have common experiences emanating from their work—both before and after ascending to the bench—in a male-dominated occupation (see, e.g., Avery, McKay, and Wilson 2008; Posner 2008). They also point to public opinion data indicating no significant differences between males and females on abortion but considerable differences on the question of whether more should be done to eliminate gender discrimination in the workforce. The data also show that a majority of women have faced discrimination in employment. To us, those advocating the narrower approach to information accounts have the better theoretical case. But, for the purpose of our empirical assessment, the difference is less important because we can distinguish between representational (which include abortion and harassment) and informational accounts on the basis of panel effects (see Table 1).

[8]This account is similar to cue taking in Congress, such that legislators may rely on cues in the form of information from "expert" colleagues to help with their voting decisions (see, e.g., Bianco 1997; Fowler 2006; Matthews and Stimson 1975). On these accounts, the information need not take the form of direct persuasion on the part of the expert (here, a female judge); her vote or even her presence may be enough.

Another possible mechanism is that a male judge alters his votes in the presence of female*s* but for collegial or strategic reasons (for more on both, see, e.g., Sunstein et al. 2006). Our emphasis on female"s" is purposeful: testing either or both comprehensively is possible only if two females sat on a panel with one male in a nontrivial fraction of panels (in which case we would expect the male to refrain from dissenting). But this type of mixed panel is rarely present in our dataset (see note 10). As a result, we can only explore this idea unidirectionally: that the female would not dissent (i.e., would not cast a pro-plaintiff vote) in the presence of two males, all else being equal.

[9]We focus here, and in the online appendix, on studies relying on quantitative evidence. There are also scores of descriptive studies, and they too reach competing conclusions. Compare, e.g., Artis (2004) and Bussel (2000).

individual effects analyses: sex discrimination. Consistent with informational accounts, for not one sex discrimination model displayed in Figure 6 does the 95% confidence interval come near the zero line (indicating no difference between male judges serving on all-male and mixed-sex panels). Rather, we observe causal effects ranging from 0.12 to 0.14—meaning that the likelihood of a male judge ruling in favor of the plaintiff increases by 12% to 14% when a female sits on the panel.[35]

Not only is this a fairly large difference but, at least from the perspective of litigants, it is also quite consequential, as Figure 7 shows. Notice that for all-male panels the probability of supporting the plaintiff in a sex discrimination dispute never exceeds 0.20—not even for the most liberal of male judges. But for mixed-sex panels, the probability never falls below 0.20 for even the most conservative males. For males at relatively average levels of ideology, the likelihood of a liberal, pro-plaintiff vote increases by almost 85% when sitting with a female judge.

Seen in this way, the results for sex discrimination panel effects mirror our findings for individual effects: for both, we find evidence of statistical significance and substantive importance. In fact, the only difference of note between the two sets of results centers on matters of methodology. In the case of individual effects we observe disparate results between the traditional regression-based analyses on the unmatched data and the analyses on the matched data; for panel effects, no such differences emerge.

Why? The most plausible answer, as we hinted earlier, is that random assignment to panels, while an imperfect selection mechanism, produces data that reasonably meet the assumption of independent assignment to treatment. This implies, in turn, that panel data will be close to balanced, or, at the least, more balanced than under the complete absence of randomization.[36] But it does *not* imply, to reiterate, that balancing via matching is per se unnecessary for panel data. Quite the opposite. The danger of assuming a balanced dataset is far greater than the perils of semiparametric balancing; the former can easily lead to severe errors of inference, while the latter cannot (see, e.g., Ho et al. 2007; Greiner 2006). Scholars should be no more willing to deploy regression-based tools to analyze nonexperimentally generated data than they would be to use, say, linear regression to estimate a model with a binary dependent variable (regardless of whether it yields results no different than a probit model). Best practice, of course, demands that we always use the most appropriate tool at our disposal. For even if the most and least suitable methods supply the same answer for a set of analyses of a particular set of data—as was the case here for panel effects—this will not always or even usually hold.

[35] On its face, this causal effect of panel composition is quite substantial, perhaps surprisingly so. Think about it this way. Because panels with female judges are significantly more plaintiff friendly than all-male panels, defendants should be more likely to settle after they observe assignment to a mixed-sex panel. To the extent that this form of selection bias exists, it ought to mitigate against a finding of a strong causal panel effect. As a result, our findings, however substantial, may actually *underestimate* the impact of panel composition on outcomes.

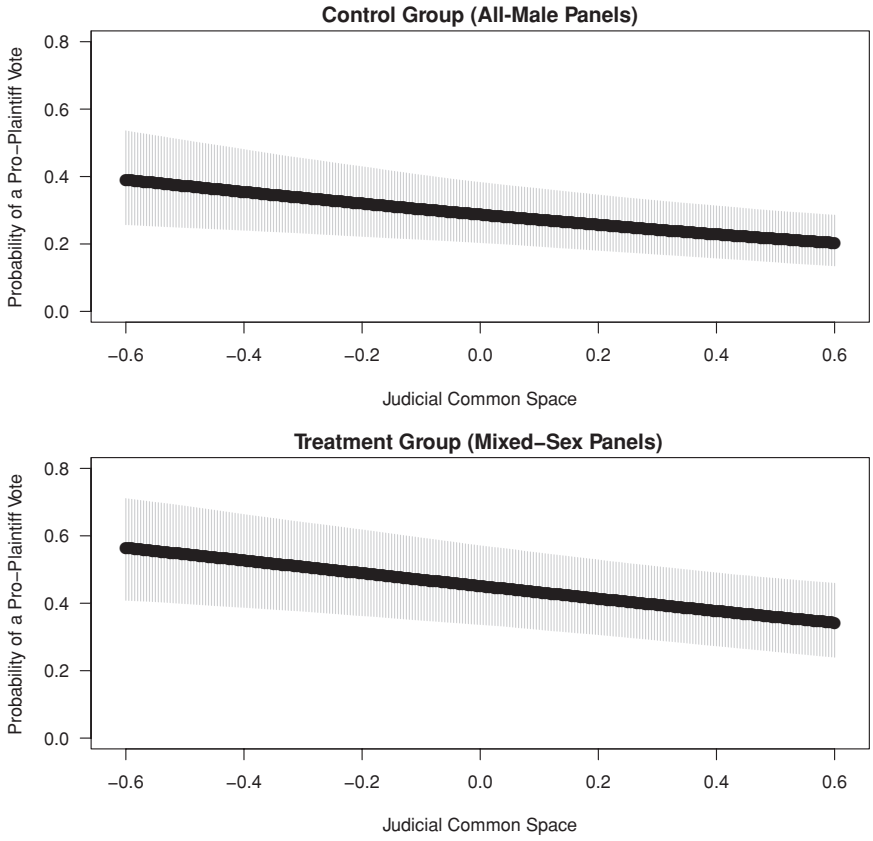
[36] To see this point, compare, e.g., the left-hand panels of Figures 2 and 3. This point also helps explain why, in certain issue areas in our study, further balancing of the original data turned out to be unnecessary.

Discussion

Ever since the campaign to place women on the federal bench began in earnest, supporters have emphasized both the symbolic and the practical implications of appointing female judges. While the first is primarily a matter for normative theorists, the second is susceptible to empirical scrutiny. And that is what we have attempted to give it here. Drawing on empirical expectations from four accounts (different voice, representational, informational, and organizational), we proceeded from a formal framework for causal inference to answer questions that have long dominated scholarly and policy discourse over the role of sex in judging.

The results of this exercise are now reasonably clear: the presence of women in the federal appellate judiciary *rarely* has an appreciable empirical effect on judicial outcomes. Rarely, though, is not never. Based on an account that isolates the analysis to judge-vote observations with a nearest-neighbor match, we observe consistent and statistically significant individual and panel effects in sex discrimination disputes: not only do males and females bring distinct approaches to these cases, but the presence of a female on a panel actually *causes* male judges to vote in a way they otherwise would not—in favor of plaintiffs. Characterized in this way, our results are consistent with an informational account of gendered judging; they also serve to reinforce other studies that identified gender effects in the employment area. Finally, our results may provide empirical fodder for a class of normative claims supportive of diversity on the bench; namely, "the greater the diversity of participation by [judges] of different backgrounds and experiences, the greater the range of ideas and information contributed to the institutional process," and the higher the likelihood of altered deliberations in

FIGURE 7 Predicted Probabilities of Pro-Plaintiff Votes in Title VII Sex Discrimination Cases as a Function of the Judicial Common Space (Ideology) for All-Male (Control) and Mixed-Sex (Treatment) Panels



The Judicial Common Space runs from most liberal (here, −0.6) to most conservative (0.6). These estimates are from the weighted logistic regression model on the matched data. All continuous variables are held at their sample means; other variables are at their sample modes. The vertical grey lines denote 95% confidence intervals.

response (Epstein et al. 2003, 944; see also Cameron and Cummings 2003).

While we hope our study goes some distance toward answering important questions in the literature, we also think that the very questions we addressed here continue to deserve a prominent place on the scholarly agenda. It seems entirely worthwhile, for example, to consider the extent to which our findings transport to other collegial courts, both here and abroad, and to other stages in the litigation process. We also can imagine extending the analyses to cover other attributes, including race, religion, and age.

We certainly commend these challenges to scholars working in the fields of public law, gender politics, and race and ethnicity. Going forward, we also encourage the use of the general framework and methods deployed here—as do a growing number of other political scientists who too now call for a reconsideration of the field's traditional and dominant approach to inference (e.g., Epstein et al. 2005; Greiner 2008; Ho et al. 2007). To them, reliance on regression analyses of unmatched data far too often leads to unreliable and misleading results. In light of the findings here, along with promising developments in the statistical sciences aimed at improving the conclusions we can draw from observational data, their message seems especially timely.

This is almost certainly true for the burgeoning scholarship on the extent to which female legislators better represent women's interests compared to their male counterparts (e.g., Dodson 2008; Reingold 2000; Swers 2002)—an

area in which the same sort of imbalances we identified may well be present. But it also may hold for research outside the gender (or race) realm. In one of the few previous studies on judicial behavior that adopted a potential-outcomes framework—Epstein and colleagues' (2005) analysis of the effect of war on Supreme Court decisions—the authors found imbalance on the key causal variable: liberal courts, relative to conservative courts, were more likely to decide cases during war times. Had Epstein et al. failed to correct for this imbalance via propensity score matching, they would have reached the highly misleading conclusion that the Court was more likely to protect individual rights in the middle of a war. Of course, the extent to which imbalance plagues other research on judging or legislating is an empirical question that researchers must evaluate for their particular projects. At the very least, though, our study, in line with the few others in this area, counsels in favor of such evaluations.

Appendix: Datasets and Selected Logistic Regression Estimates

Table A1 The Issue Areas, Years, and Sample Sizes (Measured in Votes) for the Datasets

		Sample Size			
		Individual Effects		Panel Effects	
Issue Area	Years	Full Data	Matched Data	Full Data	Matched Data
Abortion	1982-2002	297	132	270	–
ADA	1998-2002	1956	890	1648	1383
Affirmative Action	1978-2002	447	178	411	–
Campaign Finance	1976-2002	165	58	149	–
Capital Punishment	1995-2002	543	289	450	346
Contract Clause	1977-2002	222	103	201	–
EPA	1994-2002	186	100	147	–
Federalism	1995-2002	816	434	679	544
Piercing the Corporate Veil	1995-2002	318	165	274	–
(Title VII) Sex Discrimination	1995-2002	1245	590	1075	843
Sex Harassment	1995-2002	1116	594	952	784
Takings Clause	1978-2002	624	279	561	278
(Title VII) Race	1985-2002	960	468	828	639

These data originated from Sunstein et al. (2006) and were supplemented by the authors. In explaining why (and how) the years studied varied depending on the issue area studied, Sunstein et al. say, "We extended the viewscreen to earlier cases when the post-1995 sample was small. In deciding how far back to look, we typically relied on starting dates marked by important Supreme Court decisions that would predictably be cited in relevant cases" (Sunstein et al. 2004, n. 35). While the Sunstein et al. article (2004) and book (2006) consider sex harassment cases both as a part of sex discrimination cases and separately, we consider them only in the latter fashion. In addition, we limit our examination of sex discrimination cases to only those brought under Title VII. Those datasets in the panel effects context that were sufficiently balanced and did not require matching (abortion, affirmative action, campaign finance, Contract Clause, EPA, and piercing the corporate veil) have sample sizes reported only for the unbalanced data.