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).

not by offering a critique of the existing approaches, but rather by returning to first principles—theoretical and methodological approaches to drawing causal inferences.

The Potential Outcomes Framework for Causal Inference

Of interest to us and many others working in this area is whether and in what ways gender leads judges to behave differently. For a panel of judges hearing a case on an intermediate appellate court, for example, we aspire to estimate the extent to which the presence of a female judge causes male judges to vote in a particular direction when they otherwise would not.[10]

Estimating this causal effect demands counterfactual analysis (see, generally, Epstein et al. 2005; Epstein and King 2002; King, Keohane, and Verba 1994). We want to learn how a male judge would vote on a panel with a female judge *but for the presence of the female judge*. Undertaking it requires us to determine the effect of a female judge for any given panel composition, along with any other relevant (i.e., confounding) case and judge factors (such as the sex of the litigant and the ideology of the judge).

This task would be straightforward enough in a research environment lacking constraints. We would create an all-male panel and ask it to decide a sex discrimination case; then we would rerun history, holding everything constant except the absence of a female judge, and ask the panel to decide the same case. If we observed the men voting against the plaintiff when serving on the all-male panel but supporting the plaintiff when serving with a woman, then we might conclude that the female had an effect on the panel and that the effect was in the direction anticipated by at least one theoretical account of sex difference.

For a more formal accounting of this type of analysis, we adopt the potential outcomes framework posited by Neyman (1935) and Rubin (1973, 1974), thoroughly reviewed in Holland (1986), and recently applied in political science by Imai (2005) and Epstein et al. (2005). Under this framework, let the unit of analysis for our panel-effect example be the judge-vote cast by a male judge, and $i = 1, \ldots, N$ index each observation. Further, let Y_i denote an outcome variable; say, whether the judge voted for ($Y_i = 1$) or against ($Y_i = 0$) the plaintiff in a discrimination suit. Finally, each judge-vote takes place under one of two treatment conditions: the control group, denoted $T_i = 0$, includes the panels where the other two judges are male (an all-male panel); the treatment group, denoted $T_i = 1$, consists of those panels with at least one female judge (a mixed-sex panel).[11] Note that this notation is in terms of potential outcomes: the case *potentially* could have been decided by an all-male or mixed-sex panel, and the panel could have decided it for or against the plaintiff.

Under this framework and consistent with the Neyman-Rubin model, we can now formally define the causal effect for each observation (τ subscripted by i) as the difference between the two potential outcomes:

$$\tau_i = Y_i(T_i = 1) - Y_i(T_i = 0) \quad (1)$$

Observe that we have explicitly incorporated the counterfactual state of the world—or the treatment effect—for each observation. Because we observe only one of the two states of the world on the right-hand side of equation (1), this formulation consists of the difference between a factual and counterfactual. To summarize that effect—the causal effect of sex—across a number of observations, we can estimate $\bar{\tau}$, the average treatment effect (ATE), as:

$$\bar{\tau} = E[Y_i(T_i = 1)] - E[Y_i(T_i = 0)] \quad (2)$$

The difficulty, of course, is that in the real world of research we cannot rerun history to estimate the counterfactual and obtain τ_i and its summary $\bar{\tau}$. This is known as the *fundamental problem of causal inference* (Holland 1986, 947). It simply means that, for any given observational unit, we will never observe the outcome under both the treatment (a mixed-sex panel) and the control (an all-male panel). Instead, we see the judge-vote either when it takes place under the control $Y_i(T_i = 0)$ or the treatment $Y_i(T_i = 1)$. To put it another way, we can only observe the factual (e.g., if the panel was, in fact, all male, then we observe an all-male panel) and not the counterfactual (e.g., observing a mixed-sex panel, if the panel was in fact composed of all males). Consequently, and *depending on the research setting*, we must make certain assumptions to estimate τ_i.

Consider, first, the experimental setting. Were we able to randomly select judges and in turn assign them, again randomly, to treatment and control groups, we would assume that assignment is independent of all other observed

[10] Throughout this section, we focus on panel (rather than individual) effects because sex cannot be treated as a causal variable for purposes of investigating whether male and female judges decide cases differently. For more on this point, see the section "Sex as a Causal Variable" below. We also note that our research design does not attempt to account for any effects that might be due to accumulated "exposure" to diversity when individual judges repeatedly serve with one another (see Berger, Conner, and Fisek 1974).

[11] An alternative approach is to define two treatment groups: one with just one female on the panel and another with two females on the panel. We define the treatment as we do for several reasons, not the least of which is purely pragmatic: our datasets lack a sufficient number of panels with two females to perform this sort of analysis. (And, not surprisingly, we observe no panels with three females.)

pretreated covariates (denoted X_i). And then—with the assumption of independent assignment met—as the sample size grows, all observed and unobserved covariates will be balanced across the treatment and control groups due solely to the presence of randomization. Present also is the "stable unit treatment value assumption" (SUTVA; Rubin 1974), which states that the potential outcome of one unit does not depend on the treatment assignment of another unit.[12]

Because most experimental settings easily meet SUTVA and the assumption of independent assignment to treatment, researchers can estimate the average treatment effect by doing nothing more complicated than computing the differences of means:

$$\begin{aligned}\bar{\tau} &= E[Y_i(T_i = 1)] - E[Y_i(T_i = 0)] \\ &= E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \end{aligned} \quad (3)$$

Unfortunately, of course, in most studies of judging—including ours—executing an experiment of this sort is nearly as impossible as rerunning history. While it is true that the U.S. appellate courts use a "wheel" to assign judges to panels, logic and practice counsels against deeming it a mechanism for true random selection.[13] As a result, judicial specialists, again us included, must work with observational data, which substantially complicate the inferential task. One obstacle is that the assumption of independent assignment to treatment rarely, if ever, holds. This is not insurmountable, however, if we can condition on our observed covariates (X_i) and if the assumption of conditional ignorability holds. Should we have the appropriate pretreatment covariates—for a study of panel effects, judge-specific and case-specific covariates that *precede* panel assignment[14]—we can then assume that conditional on them, assignment to treatment is unconfounded; that is, after controlling for the covariates, the probability of being assigned to the treatment group is not correlated with the outcome variable.

With this obstacle hurdled, and the additional assumptions of SUTVA and strong ignorability met,[15] we can proceed to estimate the ATE $\bar{\tau}$:

$$\begin{aligned}\bar{\tau} &= E[Y_i(T_i = 1)|X_i] - E[Y_i(T_i = 0)|X_i] \\ &= E[Y_i|X_i, T_i = 1] - E[Y_i|X_i, T_i = 0] \end{aligned} \quad (4)$$

But how ought we estimate this effect? This question has been the subject of virtually no debate within public law and gender politics circles. Instead, a single approach has long dominated efforts in these fields to perform causal inference with observational data—including efforts to study gendered judging: linear regression models (or their variants for dichotomous dependent variables, such as logit or probit). The typical approach, as we mentioned earlier, is to regress an outcome variable of interest (usually the judge's vote, either for or against the sex-discrimination plaintiff) on a dichotomous sex variable and a handful of controls, including additional information about the judges (e.g., their ideology) and the cases (e.g., direction of lower court decision).

To be sure, linear regression provides analysts with a particular type of statistical control, and, if certain assumptions are met, the model will provide reliable inferences about causal effects. But equally as apparent are several very serious limitations—not the least of which is that linear regression assumes the presence of a precise functional form for the relationship between the treatment and outcome, measured covariates and the treatment, and measured covariates and the outcome.[16] In an experimental setting, where treatment assignment is randomized, this assumption is easily met. For observational data, however, we cannot depend on random assignment to ensure that our covariates are systematically unrelated to our treatment variable. As a result,

[12]This assumption likely will be violated in the descriptive individual effects analysis that follows (and relaxing it is beyond the scope of our project). Importantly, though, if this violation occurs in our data, any estimated effects would be attenuated *toward* zero. Thus, any resulting estimates of the average treatment effects will be overly conservative.

[13]Even if it were true that assignment in the circuit courts was random—less and less likely given the growing number of senior-status judges—we confront the problem of inherent stratification in the federal judiciary. We expect that across circuits, profound imbalances may exist on crucial covariates (e.g., ideology). Only if cases were randomly assigned across all circuits (such that any case could be assigned to any three judges) would we expect all other covariates to be balanced. And even in that case, Ho et al. (2007) suggest that using matching methods to balance covariates is appropriate in experimental settings to mitigate against possible confounders.

[14]It is generally important to include only pretreatment covariates in any causal analysis. Posttreatment covariates may be affected by the treatment, thus confounding estimation of causal effects. In narrow circumstances posttreatment variables can be included to obtain conditional ignorability.

[15]"Strong ignorability" (Dehejia and Wahba 1999; Rosenbaum and Rubin 1983; Smith 1997) implies that assignment to treatment is unconfounded and that overlap exists between the treatment and control groups.

[16]Another limitation is that our definition of a causal effect in equation (1) does not require constant effects across all observations, but the linear regression model does (Greiner 2008; Rubin 1973; Winship and Morgan 1999). In some cases this strong linearity assumption might be justified, but there is no reason to assume ex ante that it holds. And, if it does not, we can inappropriately estimate the causal effect without much effort (for illustrations, see Greiner 2008; Ho et al. 2007).

imbalances frequently emerge. Since performing causal inference requires researchers to limit their analyses to the range of values for which they have data in the treatment *and* the control groups,[17] the presence of imbalances can undermine the integrity of regression results. Without accounting for these imbalances in the covariates, analysts wind up comparing the equivalent of apples and oranges.

Because the regression model too readily extrapolates beyond the range of the observed data, this may well be a rather frequent occurrence in analyses of legal decisions—and perhaps especially in work on gendered judging. To see why, consider that in virtually all studies of this sort the researcher takes into account, in addition to the judges' sex, their ideology. This is a sensible choice: we know that ideology is an important determinant of judicial decisions. But since female judges are, on average, far more liberal than their male colleagues, it is also a problematic choice. Figure 1 nicely illustrates the point. Looking at U.S. Court of Appeals judges who voted in disputes over sex discrimination in employment (Title VII) or the Americans with Disabilities Act (ADA) and using their Judicial Common Space scores (Epstein et al. 2007; Giles, Hettinger, and Peppers 2001) to measure ideology, we can see that the men are rather evenly dispersed between liberal and conservative groupings. Women, in contrast, noticeably skew to the left.

Data of this sort are so imbalanced that regression analysis could produce profoundly misleading results. In concrete terms, because the range or distribution of ideology is, at least for now, sufficiently different between male and female judges serving on the federal courts, a linear regression model of their votes on their sex and ideology might well estimate a significant and negative treatment effect (men are more likely to cast left-of-center votes), when, in reality, the treatment effect is positive![18] Under such circumstances, the only way to ensure a reliable estimate of the average treatment effect is to obtain balance on the covariates; i.e., to compare apples and apples.[19]

Matching Methods for Performing Causal Inference

In the simple example depicted in Figure 1 it is easy to spot the imbalance, but when we incorporate more covariates, as we typically do, that task becomes essentially impossible. More generally, while regression can be a useful and appropriate tool in some settings, it often makes assumptions that are unjustified in the study of judging (Epstein et al. 2005).

If naively using linear regression can lead to misleading inference, especially when we expect imbalance in and nonoverlap of the covariates, what are the viable alternatives? The most promising is semiparametric matching, where the idea is to estimate equation (4) only when units are matched on all covariates. The intuition behind this approach is easy to grasp: while we can neither rerun history to see if male judges would decide the same case differently on an all-male versus mixed-sex panel nor run an experiment to test the same, we can match cases and judges that are as similar as possible (except of course on the key causal variable, the presence or absence of a female judge) to make the same causal inference. In other words, once we have conditioned on all the relevant confounding factors (i.e., pretreatment covariates; see note 14), we can attribute any remaining differences in the proportion of votes cast for or against plaintiffs to the presence of a female judge.

While matching methods are only beginning to make headway in political science (see, e.g., Epstein et al. 2005; Imai 2005), they have gained considerable traction in statistics and other related fields. And, actually, one form of matching—exact matching—has even found its way into the literature on gendered judging (see, e.g., Segal 2000; Walker and Barrow 1985). With exact matching, the idea is to estimate equation (4) only when units are matched on all covariates.

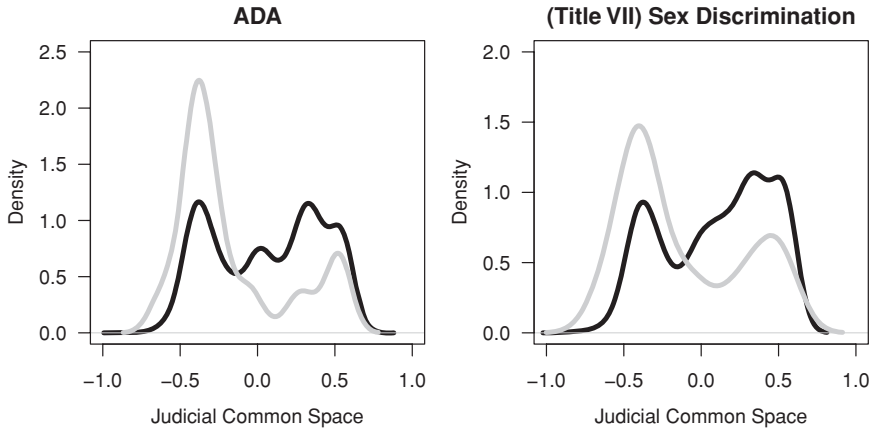
Exact matching has the benefit of increasing the plausibility of the assumption of strong ignorability. But it introduces other problems, primarily the "curse of dimensionality": as the number of covariates increases, exact

[17]This is the notion of common support, which is part of the assumption of "strong ignorability" (see note 15, King and Zeng 2006, and Smith 1997, 349).

[18]For other examples of this general phenomenon, see Greiner (2008) and Ho et al. (2007).

[19]The lack of balance depicted in Figure 1 shores up yet another problem with using linear regression to estimate equation (4). Linear regression allows us to assess the effect of the treatment on the outcome, holding all else constant. But all else is likely not constant when comparing the treatment and control groups, unless, of course, they are balanced. This ceteris paribus assumption is justified when treatment assignment is random and independent, as in an experimental setting, but it likely does not hold in most observational studies. It is also worth noting that while there are a variety of diagnostic tools associated with traditional regression analyses, because imbalance has to do with the linear extrapolation the model uses to estimate treatment effects and not model fit or correlation among covariates, these tools are not designed to detect a profound lack of balance that can result in the misestimation of treatment effects (see, e.g., Greiner 2008 and Ho et al. 2007).

FIGURE 1 The Ideology of U.S. Court of Appeals Judges Who Voted in Title VII Sex Discrimination and Americans with Disabilities Act (ADA) Cases



Each panel displays a kernel density plot that depicts the marginal distribution of ideology (measured using the Judicial Common Space), from most liberal to most conservative, of the participating U.S. Court of Appeals judges. The black line represents male judges and the grey line represents female judges. Case data come from Sunstein et al. (2006) and ideology, from Epstein et al. (2007).

matching itself can become increasingly implausible.[20] To see the problem, suppose we began with the first sex discrimination case decided by an appellate court panel in 1995. Further suppose that the suit was decided in favor of the female plaintiff by a mixed-sex panel on which the men had fairly conservative ideological scores. Finally, assume that the panel's male judges were confirmed to the bench in 1950 and 1966 and the female judge was confirmed in 1979. To find an exact match for this case we would need to identify a dispute and a panel that had the same values on all the potentially confounding variables—in this example, a suit resolved in 1995 by a panel with two relatively right-of-center men with these precise confirmation years—but on which a female judge, and not three males, sat. Because such an exact match may not exist in our database, we would be forced to discard this dispute, and likely countless others, from our analysis. And the problem—the curse really—only grows exponentially as we add more covariates, such as additional judge attributes and the direction of the lower court decision.

To avoid unnecessarily wasting data, we create matches that are not exact but are as close to exact as possible. The approach we take is to match on a one-dimensional summary of the pretreatment covariates known as the propensity score (Rosenbaum and Rubin, 1983, 1984). By calculating the predicted values from a logistic regression of the treatment indicator T_i on only the pretreatment covariates X_i, the idea is to obtain a single variable—the estimated propensity score—that serves as a summary of the covariates on the treatment and control groups. With the propensity scores in hand, we can utilize them to match observations (using a variety of strategies discussed below) without making any of the strong parametric assumptions necessitated by linear regression.

Sex as a Causal Variable

Estimating propensity scores and executing matching are tasks that require the researcher to make a series of choices, and momentarily we explain ours. But first we must deal with a final conceptual complication—one that implicates the specific research questions we ask and the precise inferences we can draw. Simply put, a crucial and by now obvious feature of the potential outcomes framework is that for a treatment to be a cause there should be "*potential* (regardless of whether it can be achieved in practice or not) for exposing or not exposing each unit to the action of a cause" (Holland 1986, 946). In practice,

[20] An additional problem with the exact-matching gender studies is that they violate the mantra of "no causation without manipulation." For more on this point, see the section "Sex as a Causal Variable" below.

this means that attributes, such as a judge's sex, *cannot be viewed as causes*. As Cox tells us, in most cases, sex "is not a causal variable but rather an intrinsic property of the individual" (1992, 296). Drawing inferences about sex, race, and other immutable characteristics is methodologically quite challenging and is only now starting to receive attention in the literature on causal inference (see, e.g., Greiner and Rubin 2009; Imai and Yamamoto 2010).

Where does this leave us with the two research questions of interest? The second question—does the presence of a female judge on a panel cause male judges to behave differently?—lends itself to causal analysis. In principle, a case could have been heard by a panel with only men or a panel with one or more women. As a result, panel composition is (experimentally speaking) subject to manipulation, and with suitable pretreatment covariates, it is possible to estimate the average treatment effect. To put it another way, because the values of our observed covariates are determined before the panel is assigned, we can assess the extent to which the presence of a female judge *causes* male judges to behave differently.

Our first research question (and the one that predominates in the existing literature)—do male and female judges decide cases differently?—presents two problems. First, because the treatment is the sex of the judge, most would say that it fails to meet the "no causation without manipulation" standard. Second, the other covariates relevant to this question—whether centering on the judge's attributes (e.g., ideology and age) or the case's details (e.g., direction of the lower court decision)—occur *after* the sex of the judge is determined. With only posttreatment covariates, we cannot estimate a causal effect.

The conclusion is thus inescapable: the question of whether sex *causes* judges to behave differently is ill posed. Instead, our data can only be informative on the descriptive—though nonetheless interesting—matter of whether male and female judges decide cases differently. This does not imply, we hasten to note, a return to regression analysis without first balancing the database. Quite the opposite: to perform better *descriptive* inference, we still should harness the power of matching methods. As Rubin himself observed,

> [E]ven though it may not make sense to talk about the 'causal' effect of a person being a white student versus being a black student, it can be interesting to compare whites and blacks with similar background characteristics to see if there are differences in academic achievement, and creating matched black-white pairs is an intuitive way to implement this comparison. (2006, 3)

Implementing Propensity Score Matching

With that important caveat now noted, we turn to the implementation of propensity score matching—a task performed in four steps: selecting appropriate factors on which to match cases and judges, amassing the data necessary to assess the various accounts of gendered judging (see Table 1), estimating the propensity scores, and matching observations. Once we have semiparametrically processed the dataset in this way, we can summarize the difference in judging for the first question and estimate the causal effect for the second (Ho et al. 2007).

Beginning with the first step, choosing covariates, we took cues from the large and well-established literature on judging in the U.S. Courts of Appeals (e.g., Cross 2007; Hettinger, Lindquist, and Martinek 2004; Scherer 2005) and incorporated both judge-based attributes (e.g., ideology and age) and case-specific factors (e.g., year of decision and the direction of the lower court decision).[21]

Our data on the votes cast by judges come from the Sunstein et al. (2006) project on the federal appellate courts. To determine whether Democratic judges reach more liberal decisions than Republicans, and whether the partisan composition of a panel affects votes as well, the Sunstein team developed a database containing the decisions of federal appellate court judges in 13 areas: abortion, affirmative action, disability law (ADA cases), campaign finance, capital punishment, the Contract Clause, environmental protection (EPA cases), federalism, piercing the corporate veil, sex discrimination in employment (Title VII), sexual harassment, the Takings Clause, and race discrimination (Title VII).[22] Not only are these data of an extremely high quality in terms of their accuracy,

[21] In addition to the primary variable of interest, the logit models we estimated incorporate the ideology of the judge, his or her year of birth, and whether she or he is a minority. We also include year fixed effects and a variable for the distance between the ideologies of the judge and the circuit's median judge. In addition, for the individual effects models only, we incorporate the sex of the case's majority opinion writer. See footnote 24 for the contents of our propensity score models.

[22] Full details on the searches used by Sunstein et al. to identify the cases for the original datasets are available in Sunstein et al. (2006, note 20–35). The appendix provides explicit detail on the number

detail, and thoroughness, but, also, fortunately, given the range of areas covered, they are extremely well suited to assessing the various theoretical accounts of sex-based judging. Under differences accounts, to reiterate, we anticipate individual effects across most of, if not all, 13 areas. For representational approaches, we also expect individual differences, but they should be largely cabined to abortion, affirmative action, sex discrimination, and sexual harassment cases. If informational accounts are afoot, we ought to observe both individual and panel effects but in an even more circumscribed category of cases, sex discrimination in employment.[23]

With the data in hand, we complete our final steps: estimating propensity scores for each judge-vote in the cases (for individual and panel effects) and matching the observations (again, for individual and panel effects). For both the individual and panel effects analyses, we used a logistic regression of the treatment indicator on a number of covariates to estimate the propensity score. The right-hand panels of Figure 2 depict the distribution of the propensity scores prior to matching for the Title VII sex discrimination and Americans with Disabilities Act (ADA) cases (see note 22).[24] For the sex discrimination scores, note the lack of common support: we observe no female judges in a broad propensity score area (roughly beyond -4). A similar, if not as severe, issue exists for the ADA propensity scores. The problem for both, and the other 11 datasets as well, is a lack of balance on many covariates, as Table 2 indicates. Note, though, that the matching procedure was successful in remedying the imbalances. A visual inspection of the left-hand panels of Figure 2 suggests as much, and Table 1 confirms what our eyes tell us. The percent reduction statistics and the eQQ medians both show that for nearly all covariates, matching greatly improved balance.[25]

Turning to panel effects, at first blush the left-hand panels of Figure 3 seem to indicate that common support for the ADA and Title VII sex discrimination cases is not much of an issue.[26] On further inspection, though, the range of the propensity scores are more evenly spread for male judges (the black line) than for the females (grey line)—a fact that Table 3 confirms (note the presence of imbalances in the scores and other covariates). Matching markedly improves balance, as the percent reductions and eQQ medians in Table 3 and the right-hand panels of Figure 3 indicate.[27] In other words, after matching, the distribution of the propensity scores for the treatment and control groups in each dataset appears quite similar, suggesting that balance has been achieved.

Performing inference required one final step: matching observations. For this task, we used "nearest-neighbor" matching with replacement; that is, for each "mixed-sex" observation (or female judge, for

of observations and years of inclusion for each of the datasets in our study.

Where appropriate to detail individual results in the remainder of the article, we are consistent and generally limit our discussion and visual displays to two exemplars: ADA and sex discrimination cases. We should note that while we would like to discuss and visualize all 13 issue areas, space limitations prevent us from providing a wholesale accounting of all summary statistics and results in the main text; as such, we implore the reader to consult the project's web site for similar information on each issue area.

[23] One might argue that informational accounts cover other areas in which female judges could plausibly possess unique expertise (e.g., abortion and sexual harassment) and so their empirical implications are indistinguishable from approaches stressing representation. Even if one finds this claim compelling, it is important to keep in mind a fundamental difference between the two accounts—not over individual effects but over panel effects, which representational approaches do not anticipate.

[24] For the ADA individual effects data, our estimated propensity score model contains judge party, year of birth, year of birth-squared, minority judge, judicial experience, and circuit dummies. For the Title VII sex discrimination cases, our propensity score model includes ideology, ideology-squared, confirmation year, confirmation year-squared, ideology $\times$ confirmation year, minority judge, minority judge $\times$ ideology, minority judge $\times$ confirmation year, and circuit dummies. Each model also contains exact matching on year of decision and lower court decision. While space limitations prevent us from reporting the contents of the other estimated propensity score models here, these details will be provided on the project's web site. For each issue area we chose propensity score models that provided the best balance between the treatment and control groups. One implication of this is that the same specification was not used in each issue area.

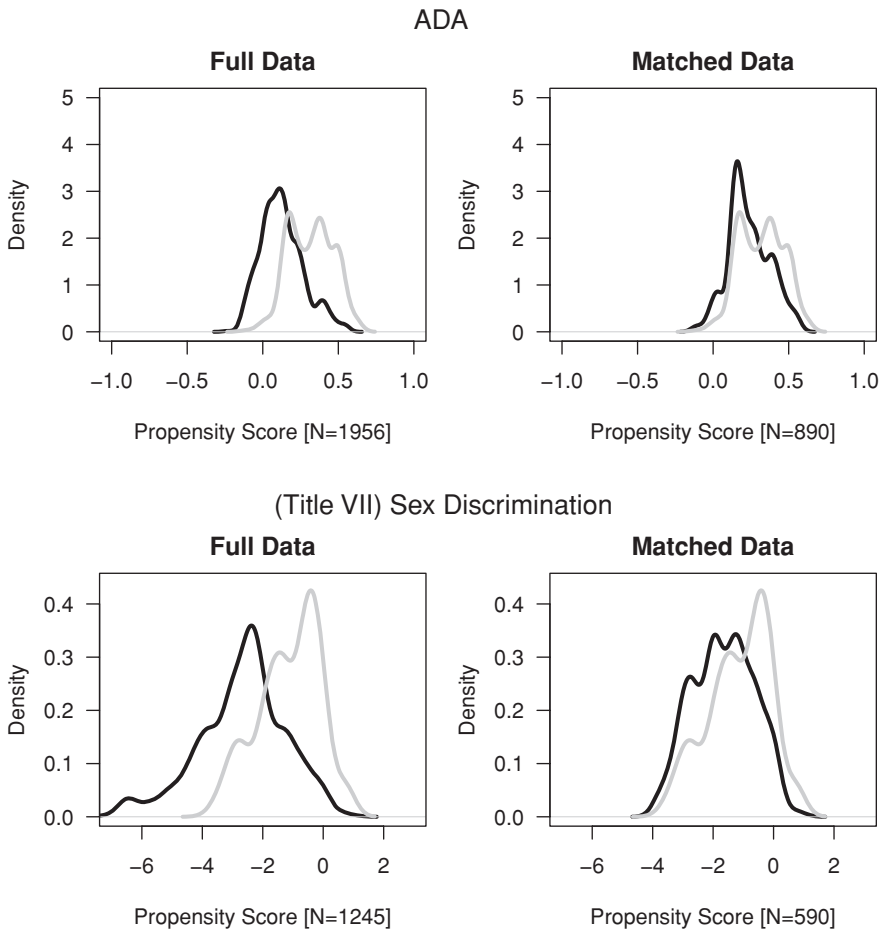
[25] For the percent reduction in the difference of means between the treatment and control groups, a reduction of 100% indicates perfect balance. The eQQ median is the median difference in the quantile-quantile plot for each variable; an eQQ median of zero is indicative of perfect balance (see, e.g., Ho et al. 2007).

Also worthy of explanation is why the matched datasets have fewer observations than the full datasets. While it may seem counterintuitive, balanced data that are comparable—even if smaller in number—are preferable to a complete sample for the purpose of estimating causal effects. To work with the full dataset would likely force us to rely on strong model assumptions to extrapolate, as we discussed in the section "The Potential Outcomes Framework for Causal Inference" above.

[26] The estimated ADA panel effects propensity score model includes ideology, ideology-squared, year of birth, year of birth-squared, ideology x year of birth, minority judge, judicial experience, and circuit dummies. The same model for the Title VII sex discrimination data includes ideology, ideology-squared, minority judge x ideology, minority judge, and circuit dummies. Each model also includes exact matching on year of decision and lower court direction. See also footnote 24.

[27] Note that in the case of confirmation year in Title VII sex discrimination decisions, eQQ median actually increases. This difference is not, however, statistically significant.

FIGURE 2 Kernel Density Plots of the Estimated Propensity Score for the ADA and Title VII Sex Discrimination Individual Effects Analyses



The black lines depict the density for all-male panels (control); the grey lines for mixed-sex panels (treatment). Each left-hand panel represents the full datasets while the right-hand panels display the propensity scores for only the matched data.

the individual analysis), the "all-male" observation (or male judge) that has the closest propensity score is selected.[28] We implemented this approach by matching observations from the control group (e.g., male judges on all-male panels) multiple times ("with replacement").[29]

[28]We used the MATCHIT package in R written by Ho et al. (2006) to perform the matching. MATCHIT implements a variety of matching methods, including nearest-neighbor matching, and provides tools for assessing balance.

[29]Nearest neighbor matching also can be implemented without replacement. Debates ensue over which of the many matching approaches is best (for reviews, see Diamond and Sekhon 2005; Ho et al. 2007). For our analyses, we estimated the propensity score in a number of ways and matched using different methods. Regardless of the approach, we obtain results comparable to those reported in the text.

Empirical Results

With the balanced datasets in hand (along with weights necessary for subsequent analyses), we turned to the task of assessing the impact of the variables of interest. In terms of implementing it, scholars are of two minds. Some suggest that researchers can estimate the causal effect with little more than a difference of proportions test (e.g., Smith 1997) because the data are now balanced. Others recommend proceeding in the typical fashion by parametrically processing the now balanced database (e.g., Ho et al.

TABLE 2 Matching Summary Statistics for the Individual Effects Analyses for ADA and Title VII Sex Discrimination Cases

	ADA Cases						
	Full Data ($N = 1956$)				Matched Data ($N = 890$)		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	0.32	0.13	0.19	94.89	0.32	0.31	0.09
Minority Judge	0.09	0.11	0.00	·	0.09	0.12	0.00
Judicial Experience	0.47	0.47	0.00	·	0.47	0.48	0.00
Judicial Common Space	−0.17	0.06	0.17	98.04	−0.17	−0.17	0.06
Confirmation Year	1991.14	1985.17	5.00	92.60	1991.14	1990.70	2.00
	(Title VII) Sex Discrimination Cases						
	Full Data ($N = 1245$)				Matched Data ($N = 590$)		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	−1.13	−2.75	1.58	91.67	−1.13	−1.27	0.57
Minority Judge	0.12	0.09	0.00	30.39	0.12	0.14	0.00
Judicial Experience	0.45	0.45	0.00	·	0.45	0.43	0.00
Judicial Common Space	−0.12	0.10	0.16	81.48	−0.12	−0.08	0.11
Confirmation Year	1990.38	1984.58	6.00	98.12	1990.38	1990.27	2.00

The left portion of each table provides results for the full, unmatched data, while the right portion displays results after matching has taken place. eQQ med is the median difference in the empirical quantile-quantile plot (an eQQ med of zero is ideal).

2007). We do both with the hope of unearthing consistent results, and, as it turns out, this is (almost) precisely what obtains.

Individual Results

We begin our analysis with the question of whether male and female judges differ in their decisions over cases in the 13 issue areas. Returning briefly to Table 1, the four accounts of gendered judging present relatively diverse empirical expectations with respect to individual effects: for "different voice," we should see effects across most, if not all, of the issue areas; for representational, effects should be limited to abortion, affirmative action, Title VII sex discrimination, and sexual harassment; for informational, we expect only Title VII sex discrimination cases to produce effects (but see note 23); and for organizational, no effects at all are anticipated.

To assess these accounts, we estimated four different models for each of the 13 datasets. The first two are the conventional tests in this literature: logistic regressions using the full *unbalanced* dataset—specifically a bivariate, with the sex of the judge as the only covariate (the equivalent of a difference of proportions test); and a fully specified model incorporating the judges' political ideology.[30]

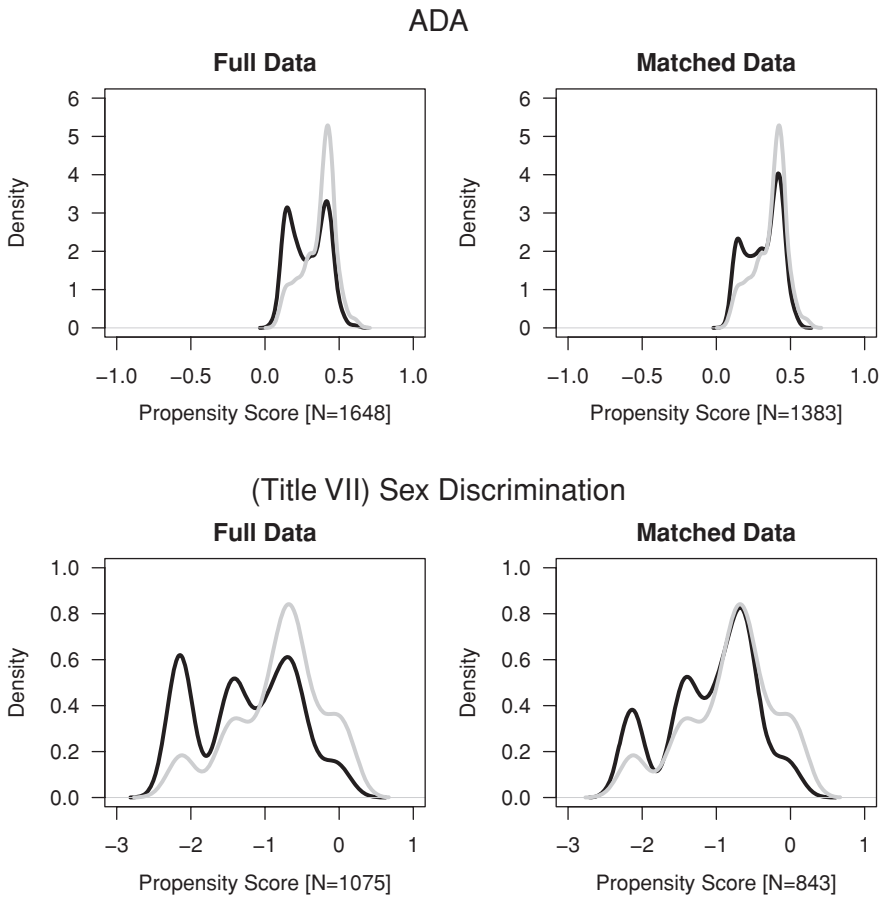
We plot the resulting individual effects ATEs for all 13 issue areas in Figure 4.[31] For each, we constructed the top two models using the full, unmatched data and the bottom two models, from the matched data.[32] Note that

[30]Some scholars use political party as a proxy for ideology (e.g., Sunstein et al. 2006). Although our general preference is to use the continuous Judicial Common Space scores (Epstein et al. 2007), our tests (unreported here) incorporating party as a substitute for this measure indicate that it makes no difference in the overall substantive results of the models. (Because party and ideology are so highly correlated, we only include one in the analysis.) In addition to the treatment and the ideology, all fully specified models house the covariates identified in Table 5 in the appendix, along with fixed effects for the decision year.

[31]We report the Title VII sex discrimination logistic regression estimates (individual and panel effects) in the appendix, Table 5. Because of space limitations, we provide all other estimates on the project's web site.

[32]In our multivariate analysis of the matched campaign finance data, the results failed to converge, and as such, we do not report ATEs for this model.

FIGURE 3 Kernel Density Plots of the Estimated Propensity Score for the ADA and Title VII Sex Discrimination Panel Effects Analyses



The black lines depict the density for all-male panels (control); the grey lines for mixed-sex panels (treatment). Each left-hand panel represents the full datasets while the right-hand panel displays the propensity scores for only the matched data.

almost without exception, female and male judges do not reach different decisions. To be sure, for the ADA and capital punishment cases, the results of the naive, unmatched analyses seem to indicate that female judges are more liberal. But this turns out to be an artifact of imbalances in the data; for both the naive and multivariate models, the matched data analyses reveal no significant effects. Note too (and in contrast to some existing empirical results), the matched data findings show that the judges' sex has no bearing on the direction of their votes in sexual harassment, affirmative action, or abortion litigation. These findings might give pause to proponents of representational and, especially, "different voice" accounts of gendered judging.

One exception to this general finding of "no difference" emerges, however, and it tends to support informational approaches while discounting organizational theories: female and male judges differ significantly in their treatment of Title VII sex discrimination suits. On average, the probability of female judges voting in favor of the plaintiff in a sex discrimination case is around 0.10 higher than it is for male judges—a difference with meaning, as Figure 5 indicates.

There we depict the predicted probabilities of men and women casting liberal (pro-plaintiff) votes in sex discrimination cases as a function of their ideology and the gender of the majority opinion's author. Note that the estimated probability of a female judge voting in favor of the plaintiff (when a female judge is the majority opinion writer) is over 0.61 at the highest levels of liberalism; for even the most left-of-center male, that figure is closer to 0.50. When the case has a male majority

TABLE 3 Matching Summary Statistics for the Panel Effects Analyses for ADA and Title VII Sex Discrimination Cases

	ADA Cases						
	Full Data ($N = 1648$)				Matched Data ($N = 1383$)		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	0.36	0.29	0.06	93.37	0.36	0.36	0.04
Minority Judge	0.12	0.10	0.00	97.32	0.12	0.12	0.00
Judicial Experience	0.43	0.49	0.00	93.15	0.43	0.44	0.00
Judicial Common Space	0.05	0.07	0.01	63.57	0.05	0.05	0.01
Confirmation Year	1985.70	1984.93	1.00	99.67	1985.70	1985.70	0.00
	(Title VII) Sex Discrimination Cases						
	Full Data ($N = 1075$)				Matched Data ($N = 843$)		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	−0.83	−1.25	0.41	77.55	−0.83	−0.92	0.21
Judicial Experience	0.43	0.46	0.00	100.00	0.43	0.43	0.00
Minority Judge	0.08	0.10	0.00	34.99	0.08	0.07	0.00
Judicial Common Space	0.09	0.11	0.02	59.00	0.09	0.08	0.02
Confirmation Year	1984.66	1984.55	0.00	·	1984.66	1983.76	1.00

The left portion of each table provides results for the full, unmatched data, while the right portion displays results after matching has taken place. eQQ med is the median difference in the empirical quantile-quantile plot (an eQQ med of zero is ideal). See also note 34.

opinion writer, the likelihood of a liberal male judge voting in favor of sex discrimination plaintiffs is less than 0.38.

What is especially interesting about these results, we believe, is that they may have gone undetected had we employed the standard procedure (i.e., estimating a logit model with unbalanced data). Note that in the full, unmatched sex discrimination data displayed in Figure 4, at a 0.05 level of statistical significance, no difference emerges between male and female judges.[33] Only via matching and balancing were we able to unearth what amounts to a fairly important sex-based distinction.

Panel Effects

Turning to panel effects, recall that accounts of sex-based judging are nearly of one mind. Of the four, only informational accounts suggest that a female may influence her male colleagues and then only in sex discrimination cases. As it turns out, our results are consistent with this one account; they also parallel the findings for individual effects (see Figure 4).[34]

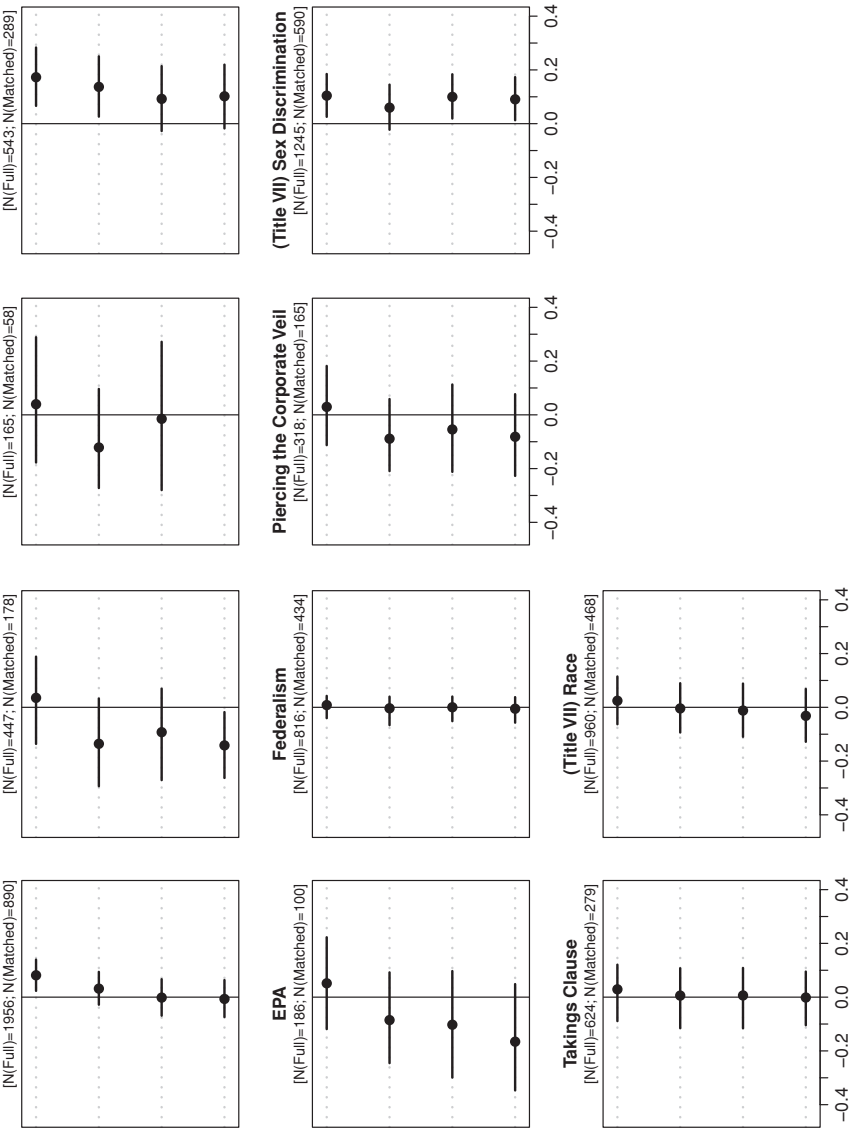
As Figure 6 indicates, for most types of disputes male judges serving on mixed-sex panels do not vote differently than male judges serving on all-male panels. As was the case for individual effects, the naive, unbalanced ADA and capital punishment analyses indicate statistical significance but yet again the matched data analyses do not support the conclusion of a genuine difference based on panel composition. More importantly, our analyses identify no significant differences in several areas (e.g., sex harassment and affirmative action) where others previously reported them (e.g., Cameron and Cummings 2003; Peresie 2005).

Where strong and systematic panel effects emerge is in precisely the same area we observed them in the

[33]The p-value on the judge-sex variable indicator is 0.161.

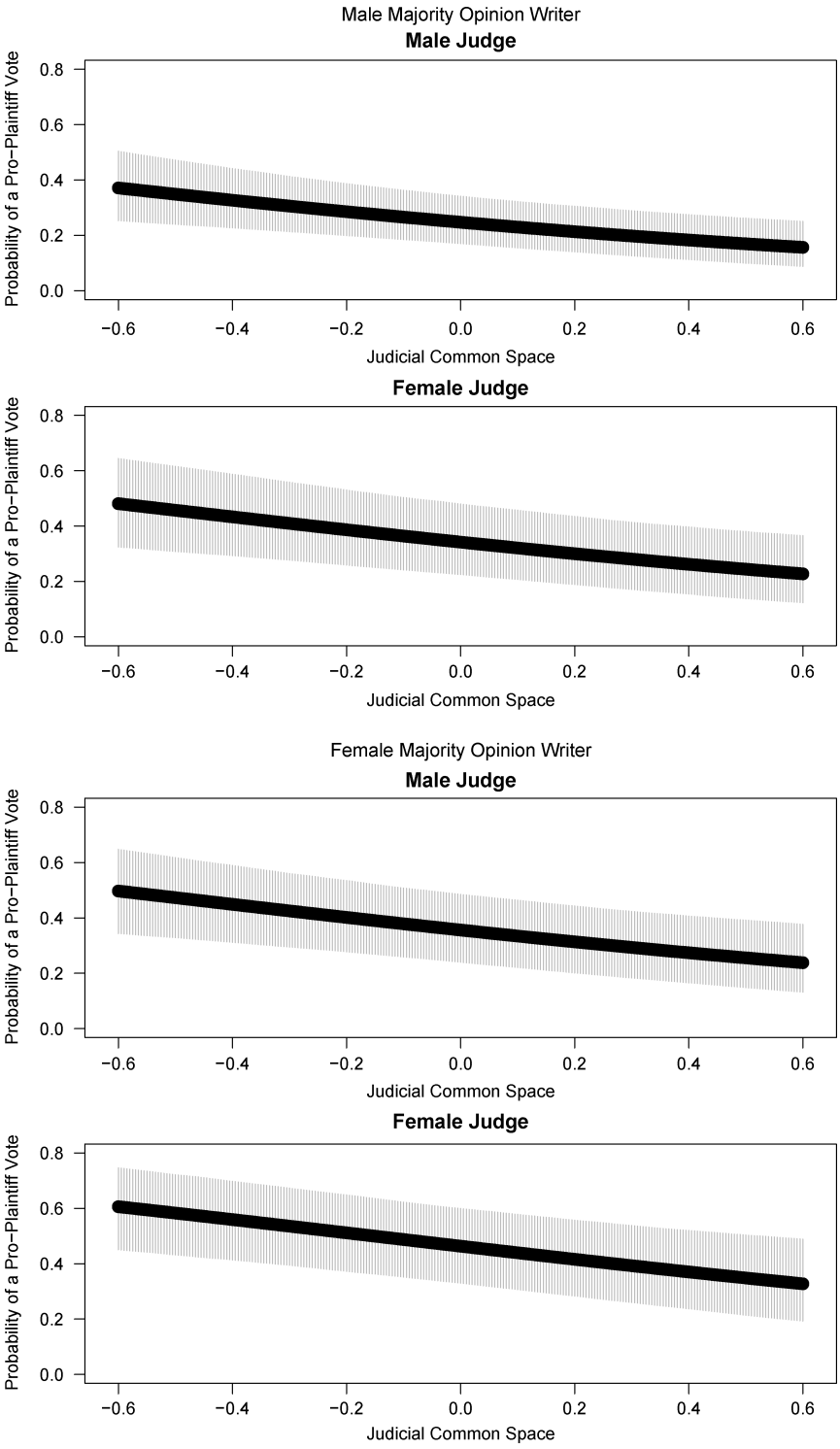
[34]In a few instances, we found that the unmatched data were sufficiently balanced. For these datasets (abortion, affirmative action, campaign finance, Contract Clause, EPA, and piercing the corporate veil) we only report average treatment effects for the unmatched data. Note, though, that only after estimating propensity score models and comparing the summary statistics across models were we able to come to the conclusion that we could appropriately analyze these datasets without matching observations.

FIGURE 4 Dotplots of Average Treatment Effects (ATEs) for Individual Effects Across 13 Issue Areas



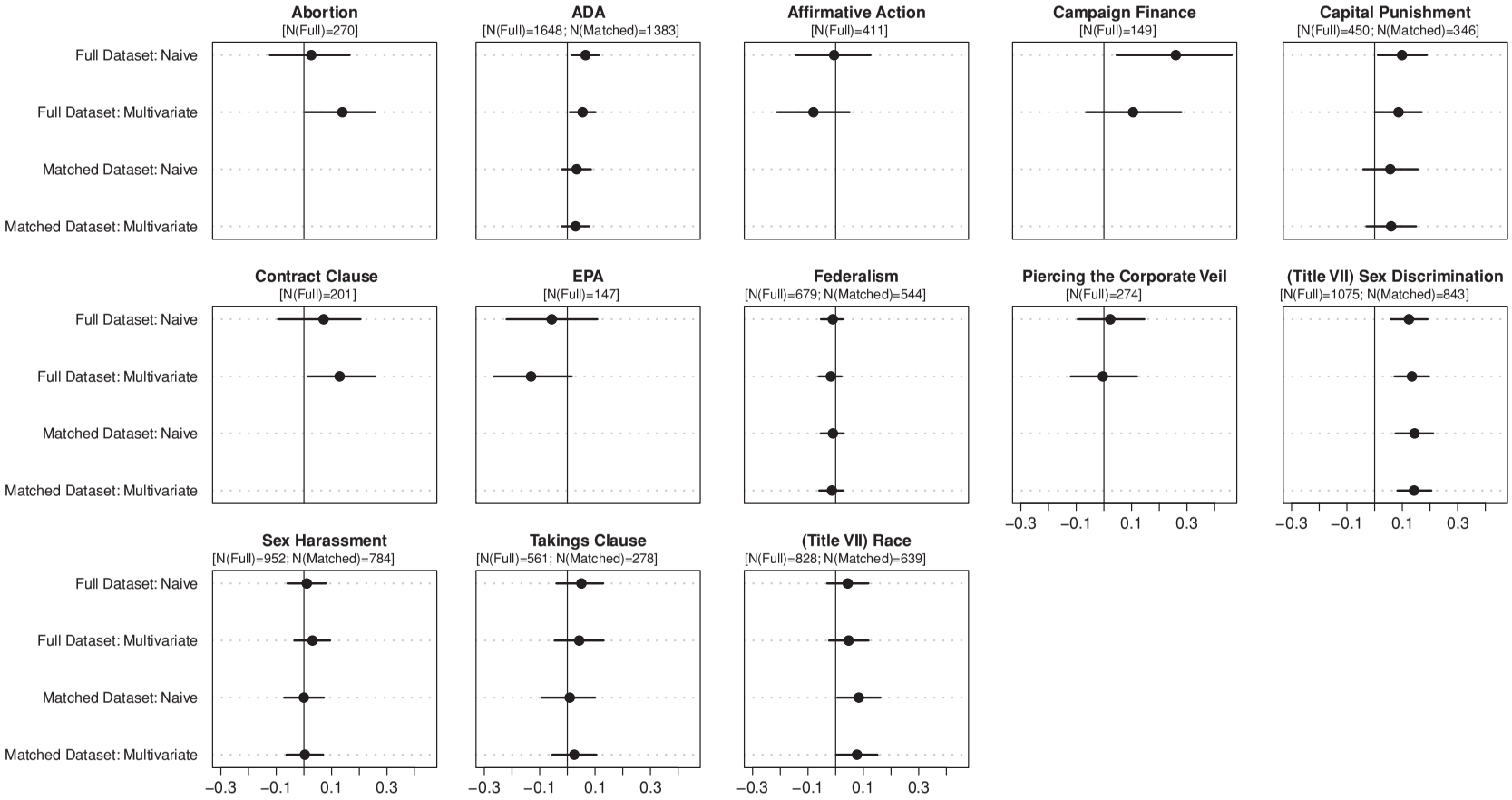
The lines represent 95% confidence intervals for the average treatment effect. For every issue area, the first two models are logistic regression models fit to each full, unbalanced dataset. The naive model includes only the judge's sex as a covariate. The other model includes the judge's sex and a number of controls, including ideology. The next two models show the ATE after nearest-neighbor matching with replacement on the estimated propensity score. The first is for a difference of proportions analysis. The second is for a logistic regression model with the judge's sex and a number of controls including ideology.

FIGURE 5 Predicted Probabilities of Pro-Plaintiff Votes in Title VII Sex Discrimination Cases as a Function of the Judicial Common Space (Ideology) and the Gender of the Majority Opinion Writer for Male and Female Judges, Individual Effects



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 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.

FIGURE 6 Dotplots of Average Treatment Effects (ATEs) for Panel Effects across 13 Issue Areas



The lines represent 95% confidence intervals for the average treatment effect. For every issue area, the first two models are logistic regression models fit to each full, unbalanced dataset. The naive model includes only the treatment as a covariate. The other model includes the treatment and a number of controls, including ideology. The next two models show the ATE after nearest-neighbor matching with replacement on the estimated propensity score. The first is for a difference of proportions analysis. The second is for a logistic regression model with the treatment and a number of controls including ideology. See also note 34.

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.

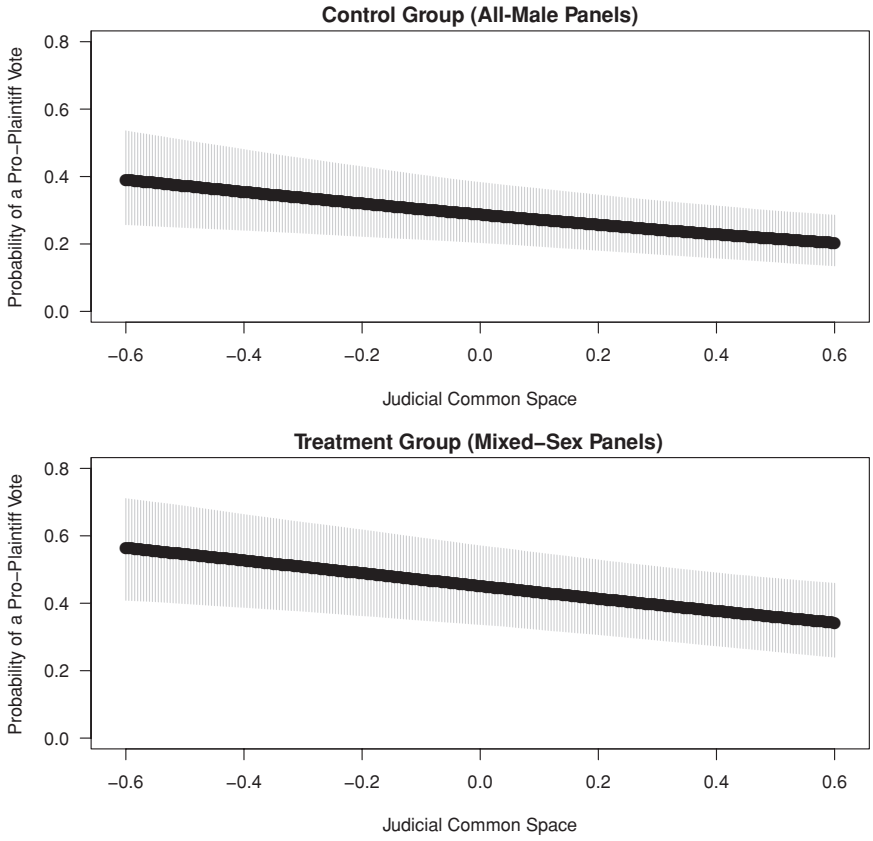
[36]To see this point, compare, e.g., the left-hand panels of Figures 2 and 3. This point also helps explain why, in certain issue areas in our study, further balancing of the original data turned out to be unnecessary.

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.

TABLE A2 Logistic Regression Estimates for the Title VII Sex Discrimination Cases, Individual and Panel Effects

	Individual Effects				Panel Effects			
Covariates	Full: Naive	Full: Multivariate	Matched: Naive	Matched: Multivariate	Full: Naive	Full: Multivariate	Matched: Naive	Matched: Multivariate
(Intercept)	−0.68*	12.68	−0.66*	72.97*	−0.83*	3.94	−0.93*	7.59
	(0.06)	(12.78)	(0.10)	(22.22)	(0.08)	(13.59)	(0.09)	(15.11)
Treatment	0.44*	0.28	0.42*	0.46*	0.54*	0.65*	0.63*	0.72*
	(0.17)	(0.20)	(0.19)	(0.22)	(0.14)	(0.15)	(0.15)	(0.16)
Judge Ideology		−0.79*		−1.06*		−0.79*		−0.75*
		(0.21)		(0.31)		(0.23)		(0.26)
Year of Birth		−0.01		−0.04*		−0.00		−0.01
		(0.01)		(0.01)		(0.01)		(0.01)
Minority Judge		0.32		0.35		0.32		0.65*
		(0.21)		(0.27)		(0.23)		(0.30)
Lower Court		1.08*		1.12*		1.10*		1.03*
Direction		(0.14)		(0.24)		(0.15)		(0.18)
Circuit Ideology		−0.11		−0.26		−0.05		−0.03
		(0.30)		(0.40)		(0.33)		(0.36)
Female Maj.		0.46*		0.51*				
Opin. Writer		(0.18)		(0.23)				
Standard errors in parentheses; *p < 0.05								
N:	1245	1245	590	590	1075	1075	843	843
Log-Likelihood:	−797.42	−700.10	−338.49	−255.48	−673.83	−590.15	−508.98	−420.95

Average treatment effects reported in Figures 4 and 6 are derived from these estimates. Standard errors are in parentheses. To conserve space, estimates of year fixed effects are not reported. The naive models include only the treatment (for individual effects a female judge, for panel effects a mixed-sex panel) as a covariate. The other models include the treatment, ideology, and other reported covariates. Similar regression tables for the 12 other issue areas are reported in the online appendix.

References

Abrahamson, Shirley S. 1984. "The Woman Has Robes: Four Questions." *Golden Gate Law Review* 14(3): 489–99.

Allen, David W., and Diane E. Wall. 1993. "Role Orientations and Women State Supreme Court Justices." *Judicature* 77(3): 156–65.

Artis, Julie E. 2004. "Judging the Best Interests of the Child: Judges' Accounts of the Tender Year Doctrine." *Law and Society Review* 38(4): 769–806.

Avery, Derek R., Patrick F. McKay, and David C. Wilson. 2008. "What Are the Odds? How Demographic Similarity Affects the Prevalence of Perceived Employment Discrimination." *Journal of Applied Psychology* 93(2): 235–49.

Baldez, Lisa, Lee Epstein, and Andrew D. Martin. 2006. "Does the U.S. Constitution Need an ERA?" *Journal of Legal Studies* 35(1): 243–83.

Beiner, Theresa M. 2002. "The Elusive (but Worthwhile) Quest for a Diverse Bench in the New Millenium." *University of California Davis Law Review* 36(3): 597–617.

Berger, Joseph, Thomas L. Conner, and M. Hamit Fisek. 1974. *Expectation States Theory: A Theoretical Research Program.* Cambridge, MA: Winthrop.

Bianco, William T. 1997. "Reliable Source or Usual Suspects? Cue-taking, Information Transmission, and Legislative Commitees." *Journal of Politics* 59(3): 913–24.

Brudney, James J., Sara Schiavoni, and Deborah J. Merrit. 1999. "Judicial Hostility Toward Labor Unions? Applying the Social Background Model to a Celebrated Concern." *Ohio State Law Journal* 60(5): 1675–1766.

Bussel, Daniel J. 2000. "Textualism's Failures: A Study of Overruled Bankruptcy Decisions." *Vanderbilt Law Review* 53(3): 887–946.

Cameron, Charles, and Craig Cummings. 2003. "Diversity and Judicial Decision Making: Evidence from Affirmative Action Cases in the Federal Courts of Appeals, 1971-1999." Paper presented at the Crafting and Operating Institutions conference, Yale University.

Carroll, Susan J. 1984. "Woman Candidates and Support for Feminist Concerns." *Western Political Quarterly* 37(2): 307–23.

Clark, Mary L. 2004. "One Man's Token Is Another Woman's Breakthrough? The Appointment of the First Women Federal Judges." *Villanova Law Review* 49(3): 487–548.

Cook, Beverly B. 1981. "Will Women Judges Make a Difference in Women's Legal Rights?" In *Women, Power, and Political Systems*, ed. Margherita Rendel. London: Croom Helm, 216–39.

Cox, D. R. 1992. "Causality: Some Statistical Aspects." *Journal of the Royal Statistical Society, Series A* 155(2): 291–301.

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

Crowe, Nancy. 1999. "The Effects of Judges' Sex and Race on Judicial Decision Making on the U.S. Courts of Appeals, 1981-1996." Ph.D. dissertation, University of Chicago.

Dahlerup, Drude. 2006. *Women, Quotas and Politics*. New York: Routledge.

Davis, Sue. 1994. "Do Women Judges Speak 'In a Different Voice?' Carol Gilligan, Feminist Legal Theory, and the Ninth Circuit." *Wisconsin Women's Law Journal* 8(1): 143–73.

Davis, Sue, Susan Haire, and Donald R. Songer. 1993. "Voting Behavior and Gender on the U.S. Courts of Appeals." *Judicature* 77(3): 129–33.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053–62.

Diamond, Alexis, and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Paper presented at the annual meeting of the Society for Political Methodology, Florida State University.

Dodson, Debra L. 2008. *The Impact of Women in Congress*. New York: Oxford University Press.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal, and Chad Westerland. 2007. "The Judicial Common Space." *Journal of Law, Economics & Organization* 23(2): 303–25.

Epstein, Lee, Daniel E. Ho, Gary King, and Jeffrey A. Segal. 2005. "The Supreme Court During Crisis." *NYU Law Review* 80(1): 1–116.

Epstein, Lee, and Gary King. 2002. "The Rules of Inference." *University of Chicago Law Review* 69(4): 1–133.

Epstein, Lee, Jack Knight, and Andrew D. Martin. 2003. "The Norm of Prior Judicial Experience and Its Consequences for Career Diversity on the U.S. Supreme Court." *California Law Review* 91(4): 903–66.

Farhang, Sean, and Gregory Wawro. 2004. "Institutional Dynamics on the U.S. Court of Appeals: Minority Representation Under Panel Decision Making." *Journal of Law, Economics & Organization* 20(2): 299–330.

Federal Judicial Center. 2007. "Federal Judges Biographical Database." http://www.fjc.gov/public/home.nsf/hisj.

Fowler, James H. 2006. "Connecting the Congress: A Study of Cosponsorship Networks." *Political Analysis* 14(4): 456–87.

Giles, Michael W., Virginia A. Hettinger, and Todd Peppers. 2001. "Picking Federal Judges: A Note on Policy and Partisan Selection Agendas." *Political Research Quarterly* 54(3): 623–41.

Gilligan, Carol. 1982. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.

Greiner, D. James. 2008. "Causal Inference in Civil Rights Litigation." *Harvard Law Review* 122(2): 533–98.

Greiner, D. James, and Donald B. Rubin. 2009. "Potential Outcomes and Immutable Characteristics." Paper presented at the Applied Statistics Workshop, Harvard University.

Gryski, Gerard, Eleanor C. Main, and William J. Dixon. 1986. "Models of State High Court Decision Making in Sex Discrimination Cases." *Journal of Politics* 48(1): 143–55.

Hettinger, Virginia A., Stefanie A. Lindquist, and Wendy L. Martinek. 2004. "Comparing Attitudinal and Strategic Accounts of Dissenting Behavior on the U.S. Courts of Appeals." *American Journal of Political Science* 48(1): 123–37.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth Stuart. 2006. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." R package version 2.2-11, http://gking.harvard.edu/matchit.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–70.

Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99(2): 283–300.

Imai, Kosuke, and Teppei Yamamoto. 2010. "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54(2): 543–60.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2): 131–59.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Kritzer, Herbert M., and Thomas M. Uhlman. 1977. "Sisterhood in the Courtroom: Sex of Judge and Defendant in Criminal Case Disposition." *Social Sciences Journal* 14(2): 77–88.

Martin, Elaine. 1990. "Men and Women on the Bench: Vive la Difference?" *Judicature* 73(4): 204–8.

Martin, Elaine, and Barry Pyle. 2000. "Gender, Race and Partisanship on the Michigan Supreme Court." *Albany Law Review* 63(4): 1205–36.

Martin, Elaine, and Barry Pyle. 2005. "State High Courts and Divorce: The Impact of Judicial Gender." *University of Toledo Law Review* 36(4): 923–47.

Martin, Patricia Yancey, John R. Reynolds, and Shelley Keith. 2002. "Gender Bias and Feminist Consciousness among Judges and Attorneys: A Standpoint Theory Analysis." *Signs* 27(3): 665–701.

Matthews, Donald R., and James A. Stimson. 1975. *Yeas and Nays: Normal Decision-Making in the U.S. House of Representatives*. New York: Wiley.

Maule, Linda. 2000. "A Different Voice: The Feminine Jurisprudence of the Minnesota State Supreme Court." *Buffalo Women's Law Journal* 9: 295–316.

Neyman, Jerzy. 1935. "Statistical Problems in Agricultural Experimentation." *Journal of the Royal Statistical Society* II(2): 107–154.

Ostberg, C. L., and Matthew E. Wetstein. 2007. "In a Different Voice: Sex Difference in Economic Cases Decided by the Canadian Supreme Court." Paper presented at the annual meeting of the Canadian Political Science Association, Saskatoon, Saskatchewan.

Peresie, Jennifer L. 2005. "Female Judges Matter: Gender and Collegial Decisionmaking in the Federal Appellate Courts." *Yale Law Journal* 114(7): 1759–90.

Pitkin, Hanna. 1967. *The Concept of Representation*. Berkeley: University of California Press.

Posner, Richard A. 2008. *How Judges Think*. Cambridge, MA: Harvard University Press.

Reingold, Beth. 2000. *Representing Women: Sex, Gender and Legislative Behavior in Arizona and California*. Chapel Hill: University of North Carolina Press.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79(387): 516–24.

Rubin, Donald B. 1973. "Matching to Remove Bias in Observational Studies." *Biometrics* 29(1): 159–83.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 6(5): 688–701.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Scherer, Nancy. 2005. *Scoring Points: Politicians, Activists, and the Lower Federal Court Appointment Process*. Palo Alto, CA: Stanford University Press.

Segal, Jennifer A. 2000. "Representative Decision Making on the Federal Bench: Clinton's District Court Appointees." *Political Research Quarterly* 53(1): 137–50.

Sherry, Suzanna. 1986. "Civic Virtue and the Feminine Voice in Constitutional Adjudication." *Virginia Law Review* 72(3): 543–616.

Sisk, Gregory C., Michael Heise, and Andrew P. Morriss. 1998. "Charting the Influences on the Judicial Mind: An Empirical Study of Judicial Reasoning." *NYU Law Review* 73(5): 1377–1500.

Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27: 325–53.

Steffensmeier, Darrell, and Chris Herbert. 1999. "Women and Men Policymakers: Does the Judge's Gender Affect the Sentencing of Criminal Defendants?" *Social Forces* 77(3): 1163–96.

Sullivan, Kathleen M. 2002. "Constitutionalizing Women's Equality." *California Law Review* 90(3): 735–64.

Sunstein, Cass R., David Schkade, and Lisa Ellman. 2004. "Ideological Voting on Federal Courts of Appeals: A Preliminary Investigation." *Virginia Law Review* 90(1): 301–54.

Sunstein, Cass R., David Schkade, Lisa M. Ellman, and Andres Sawicki. 2006. *Are Judges Political? An Empirical Analysis of the Federal Judiciary*. Washington, DC: Brookings.

Swers, Michele L. 2002. *The Difference Women Make: The Policy Impact of Women in Congress*. Chicago: University of Chicago Press.

Thomas, Sue. 1994. *How Women Legislate*. New York: Oxford University Press.

Tobias, Carl. 1990. "The Gender Gap on the Federal Bench." *Hofstra Law Review* 19(1): 171–84.

Walker, Thomas G., and Deborah J. Barrow. 1985. "The Diversification of the Federal Bench: Policy and Process Ramifications." *Journal of Politics* 47(2): 596–617.

Winship, Christopher, and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25(1): 659–706.