Untangling the Causal Effects of Sex on Judging*

Christina L. Boyd
Ph.D. Student
Department of Political Science
Washington University in St. Louis

Lee Epstein
Beatrice Kuhn Professor
School of Law and Department of Political Science
Northwestern University

Andrew D. Martin
Professor
School of Law and Department of Political Science
Washington University in St. Louis

April 24, 2007

*First version. Please send all correspondence to Lee Epstein, Northwestern University School of Law, 357 E. Chicago Ave., Chicago IL 60611. Email: lee-epstein@northwestern.edu; Phone: 312.503.1838. We thank the Center for Empirical Research in the Law, the Weidenbaum Center at Washington University, the National Science Foundation, and the Northwestern University School of Law for supporting our research; Cass Sunstein for sharing his data; Shari Diamond, Sarah Fischer, and Nancy Staudt for providing useful comments; and 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] will house a full replication archive, including the data and documentation necessary to reproduce our results.

Untangling the Causal Effects of Sex on Judging

Abstract

We enter the debate over the role of sex in judging by addressing the two predominant empirical questions it raises: whether male and female judges decide cases distinctly ("individual effects") and whether the presence of a female judge on a panel causes her male colleagues to behave differently ("panel effects"). We do not, however, rely exclusively on the predominant statistical models—variants of standard regression analysis—to address them. Because these tools alone are ill-suited to the task at hand, we deploy a more appropriate methodology—non-parametric matching—which follows from a formal framework for causal inference.

Applying matching methods to sex discrimination suits resolved in the federal circuits between 1995 and 2002 yields two clear results. First, we observe substantial individual effects: 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, we find that when a woman serves on a panel with men, the men are significantly more likely to rule in favor of the rights litigant. Both effects are is so persistent and consistent that they may come as a surprise even to those scholars who have long posited the existence of gendered judging.

Untangling the Causal Effects of Sex on Judging

> Studies show overwhelming evidence that gender-based myths, biases and stereotypes are deeply embedded in the attitudes of many male judges... Researchers have concluded that gender difference has been a significant factor in judicial decision-making.
>
> —Bertha Wilson, first female justice on the Canadian Supreme Court

> There is simply no empirical evidence that gender differences lead to discernible differences in rendering judgment.
>
> —Sandra Day O'Connor, first female justice on the U.S. Supreme Court

Suppose we identified two judges who were identical in all respects except that one was a Democrat and the other, a Republican. Further suppose that we asked both to decide a case presenting exactly the same legal issues. Under this scenario, most political theories of judging—backed by mounds of empirical evidence—would predict different outcomes. If the case involved, say, an industry attack on a labor regulation, the Democrat would rule in favor of the government and the Republican, for business. If it were an employment discrimination suit, ditto. We would guess that the Democratic would favor the plaintiff, and the Republican, the employer. In both instances, our prediction would be right far more often than wrong (see, e.g., Sunstein et al., 2006).

Now consider precisely the same hypothetical—except that the two judges differ only in their sex: one is a female, the other is a male. Theoretically, we have good reasons to believe that the woman, no less than the Democrat, would be more plaintiff friendly in the employment case than the man (see, e.g., Kay and Sparrow, 2001; Sullivan, 2002). Empirically, though, the answer is far less clear; actually it depends on the study one consults. Based on her statement, we can only guess that Justice O'Connor read papers by Ashenfelter, Eisenberg and Schwab (1995); Kulik, Perry and Pepper (2003); Manning (2004); Martinek (2003); Sisk, Heise and Morriss (1998); and Westergren (2004)—none of which found evidence of gendered judging. Justice Wilson, on the other hand, could have cited studies undertaken by Crowe (1999); Martin and Pyle (2005); McCall (2005); Peresie (2005); Segal (2000); and

Smith (2005)—all of which indicate that sex plays a significant role in judicial decisions.[1]

In what follows we jump into this debate by addressing the two predominant empirical questions it raises: first, do male and female judges decide cases differently and second, does the presence of a female judge on a panel cause her male colleagues to behave differently?[2] We do not, however, rely exclusively on the predominant statistical models—variants of standard regression analysis—to address them. Because these tools alone are inappropriate to the task at hand, we deploy non-parametric matching methods, which follow from a formal framework for causal inference.

Our application of these methods to sex discrimination cases decided by federal appellate courts reveals that Justice Wilson has the better case. 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. Both effects are is so substantial and consistent that they may come as a surprise even to those scholars who have long posited the existence of gendered judging.

1 What We (Don't) 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 replace her with a woman. Various news sources reported that elites on the left and right agreed that O'Connor's seat should be "reserved" for a female, and the public concurred. A national poll taken soon after O'Connor make public her decision to retire indicated that three out of every four respondents

[1]Our hypothetical centers on judging at the individual level. Results are equally mixed when we turn to studies that explore panel effects on collegial courts. Compare, e.g., Farhang and Wawro (2004) (finding that the presence of a female judge increases the odds of her male colleagues voting for the plaintiff in employment discrimination cases) and Cameron and Cummings (2003) (finding only trivial panel effects). For the results of other studies, see Appendix A.

[2]Our phrasing here is not accidental. For the reasons we supply in Section 2, for only the second question can we draw a causal inference.

favored replacing O'Connor with a woman.[3] Even the first lady ventured an opinion. In an interview broadcast on NBC's "Today Show," Laura Bush said that she "would really like [the President] to name another woman to the Supreme Court."[4]

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 coalesce around the idea of replacing O'Connor with a woman? One answer centers on "social legitimacy," or the notion that "democratic institutions in heterogeneous societies ought to reflect the make-up of society, if only to offset the memories of exclusion" (Cameron and Cummings, 2003, 28). Because women now constitute nearly one-third of all lawyers in the United States, elected officials ought work to ensure their commensurate representation on the nation's highest court. Or so the argument goes.[5]

Another response centers less on the sheer presence of women on the bench and more on "their participation and their perspective" (Sherry, 1986). Capturing its flavor is Cook's (1981, 216) often cited remark that "The organized campaign [beginning in the 1970s] to place more women on the bench *rests* on the hope that women judges will seize decision-making opportunities to liberate other women." In other words, whether owing to biological differences, a unique world view, or distinct cultural, social, and professional experiences, female judges are desirable because they will be more hospitable to allegations of sex discrimination than their male counterparts (e.g. Sherry, 1986; Brudney, Schiavoni and Merrit, 1999; Martin, Reynolds and Keith, 2002; Clark, 2004). Likewise, due to their distinct perspective, females, on this account, may alter the choices made by their male colleagues (i.e., induce them to decide cases differently than they otherwise would) on legal questions of

[3]The questions was: "Do you think it is essential that a woman replace Sandra Day O'Connor, that it is a good idea but not essential, that it doesn't matter to you, or that it is a bad idea?" CNN/USA Today/Gallup Poll, July 7-10, 2005.

[4]President Bush responded in kind: "The first lady gave me some good advice yesterday, which is to consider women—which, of course, I'm doing."

[5]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, 296-297).

particular concern to women (see, e.g., Peresie, 2005; Baldez, Epstein and Martin, 2006; Sullivan, 2002).

No doubt, as purely theoretical matters, these ideas are not just widespread but widely held.[6] Lining up in support of claims about the significant role of "gender difference" in judging are scores of prominent legal scholars (e.g., Sullivan, 2002; Kay and Sparrow, 2001), social scientists (e.g., Farhang and Wawro, 2004; Steffensmeier and Herbert, 1999) and, in addition to Wilson of Canada, many U.S. judges (see, e.g., Abrahamson, 1984 and, more generally, Martin, 1990).

Equally without doubt, the quest for empirical validation has proved oddly elusive. On the one hand, studies searching for gender effects evince remarkable similarities in their design and methods. As Appendix A reveals, virtually all 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).

4. attends to (approximately) the same covariates (i.e., confounding factors), chiefly attributes of the judge (e.g., their ideology, age, judicial experience, race) and characteristics of the case (e.g., legal facts, sex of the plaintiff).

On the other hand, the resulting research findings are so mixed that they practically defy characterization. By our count, social scientists and legal academics have produced nearly 30

[6]The genesis for most theoretical work in this area is Gilligan (1982), which has faced its share of criticism on any number of grounds—sociological, biological, psychological, and methodological. Nonetheless, as Beiner (2002, 602) 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." Based on our inventory of the literature, Beiner has it exactly right.

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).[7] Of those, 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.

Because some of the existing studies examine areas of the law for which even Justice Wilson and others in her theoretical camp would have difficulty sustaining claims of gender difference (e.g., disputes involving the Internal Revenue Service), uneven results are not surprising. But even investigations focusing exclusively on areas where a link between sex and judging seems quite plausible (e.g., sex discrimination) are notable for their lack of uniformity in results. Three recent examples serve to make the point. In her analysis of statutory sex discrimination disputes in the U.S. circuits between 1999-2001, Peresie (2005) found evidence of non-trivial sex effects at both the individual and panel level: female judges were more likely to find for the plaintiff, as were panels that included a female. Westergren's (2004) examination of a similar set of cases, however, revealed neither individual nor panel effects based on gender. Crowe's (1999) study of the same courts splits the difference. Like Peresie, she found that female judges were more favorable toward plaintiffs in sex-discrimination suits; and like Westergren, she unearthed no evidence that the presence of a female judge affected the decision making of males on the panel.

2 Drawing Causal Inferences about Sex and Judging

Why scholars cannot seem to come to rest over basic questions about gendered judging is of less immediate interest to us than the question of how to bring order to the "hodgepodge" that is the existing state of the literature.[8] In what follows, we undertake this challenge, not by offering a critique of each and every study, but rather by returning to first principles—

[7] In addition to these quantitative works are scores of descriptive studies. They too reach competing conclusions. Compare, e.g., Artis (2004) and Bussel (2000).

[8] We borrow this term from Boucher and Segal (1995, 826), who characterized the existing state of the literature on certiorari as a "hodgepodge at best." To the extent that they were referencing the sharp disagreements among scholars despite decades of study, precisely the same holds about the gendered judging field.

theoretical and methodological approaches to drawing causal inferences.

2.1 The Potential Outcomes Framework for Causal Inference

Of interest to us and all others working in this area is whether 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 support a plaintiff's claim that they otherwise would not.[9]

Estimating this causal effect demands counterfactual analysis.[10] 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*. And 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 in that case and that the effect was in the direction anticipated by theoretical accounts of sex difference.

More formally, we adopt the potential outcomes framework first posited by 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 our outcome variable; whether the judge voted for the plaintiff in a sex discrimination case ($Y_i = 1$) or against the plaintiff ($Y_i = 0$). Finally, each judge-vote takes place under one of two treatment conditions: the control group, denoted

[9]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 Section 2.3.

[10]We adapt the sentiments in this and the next paragraph from Epstein et al. (2005); Epstein and King (2002); King, Keohane and Verba (1994).

$T_i = 0$, 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 could have *potentially* been decided by an all-male or mixed-sex panel and the panel could have decided it for or against the plaintiff. (In a research environment with constraints—the environment under which we all operate—we are only able to observe one of the two potential outcomes. We bracket this problem for now, and return to it momentarily.)

We can under this framework and in line with the Rubin model formally define the causal effect for each observation (τ subscripted by i) as the difference between the two potential outcomes:

$$\tau_i = Y_i(T_i = 1) - Y_i(T_i = 0) \tag{1}$$

Note that we have now 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 the average treatment effect (ATE) as:

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

The obstacle, 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 *fundamental problem of causal inference* (Holland, 1986, 947), and 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); we see either the judge-vote when it takes place under the control $Y_i(T_i = 0)$ or under the treatment $Y_i(T_i = 1)$. To put it another way, we

[11] An alternative approach to this problem 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 two reasons. First, Farhang and Wawro (2004), among others, suggests that the presence or absence of a female on an appellate panel is what matters, not the actual number of women. Second and more pragmatically, our dataset lacks a sufficient number of panels with two females to perform this sort of analysis. (And, not surprisingly, we observe no panels with three females.)

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 (a mixed-sex panel, if the panel was in fact all male). 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 assign judges to treatment and control groups, we would assume that assignment is independent of all other observed pretreated covariates, denoted X_i. And, with this assumption of independent assignment met, then due solely to randomization, as the sample size grows all observed and unobserved covariates will be balanced across the treatment and control groups. An additional assumption is the "stable unit treatment value assumption" (SUTVA) (Rubin, 1974), which states, first, that the treatment regime is identical for all observations[12] and second, that the status of an observation must be independent of the potential outcomes for all other observations.[13]

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:

$$\bar{\tau} = E[Y_i(T_i = 1)] - E[Y_i(T_i = 0)] = E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \tag{3}$$

Unfortunately, of course, in most studies of judging—including ours—executing an experiment of this sort is nearly as impossible as rerunning history.[14] Rather, judicial specialists,

[12]The idea here is that all members of the treatment group receive the same treatment. In our study, the assumption is easily met given our definition of treatment.

[13]The idea is that a member of the treatment group must not affect others being treated. In our study, a violation of this assumption would occur if male judges on a mixed-sex panel affected one another's behavior due solely to their service on mixed-sex panels.

[14]Readers might wonder why the assignment wheel used in the U.S. Courts of Appeals does not operate as a randomization mechanism. Even if it were true that case assignment in the circuit courts was random—less and less likely given the growing number of senior-status judges—we cannot view the wheel as a randomization device because of inherent stratification in the federal judiciary. We expect that across circuits, profound imbalances may exist on crucial covariates such as ideology and judicial experience. 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.

again us included, typically 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. Should we have the appropriate pre-treatment covariates—for a study of panel effects, judge-specific and case-specific covariates that *precede* panel assignment[15]—we can assume that conditional on the them, assignment to treatment is unconfounded; that is, the probability of being assigned to the treatment group is not correlated with the outcome variable after controlling for the covariates.

With this obstacle hurdled, and two additional assumptions met,[16] we can proceed to estimate the average treatment effect $\bar{\tau}$:

$$\bar{\tau} = E[Y_i(T_i = 1)|X_i] - E[Y_i(T_i = 0)|X_i] = E[Y_i|X_i, T_i = 1] - E[Y_i|X_i, T_i = 0] \qquad (4)$$

But how ought we estimate it? This question has been the subject of virtually no debate within public law circles. Instead, a single approach—regression analysis—has long dominated efforts in the field to perform causal inference with observational data. Certainly this holds for studies of gendered judging: the vast majority, as we noted earlier and show in Appendix A, deploy linear regression models (or their variants for dichotomous dependent variables, such as logit or probit) to perform causal inference. The typical approach, to reiterate, 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., the sex of the plaintiff).

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 certain is the existence of very serious limitations—not the least of which is that linear regression assumes that assignment to a treatment group is conditionally

[15]It is crucial to include only pre-treatment covariates in any causal analysis. Post-treatment covariates may be affected by the treatment, thus confounding estimation of causal effects.

[16]The first, just as before, is SUTVA. The second is called "strong ignorability" (Rosenbaum and Rubin, 1983; Smith, 1997; Dehejia and Wahba, 1999), which implies that assignment to treatment is unconfounded and that overlap exists between the treatment and control groups.

independent of the other covariates.[17] In an experimental setting, where treatment assignment is randomized, this assumption is not terribly strong. 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, imbalances may emerge, which can seriously undermine the integrity of regression results. That is because performing causal inference requires researchers to limit their analyses to the range of values for which they have data in both the treatment and the control group.[18] When they do not—as a result of imbalance in the covariates—analysts wind up comparing the equivalent of apples and oranges.

Because the regression model readily extrapolates beyond the range of the observed data, this is likely an all-too-frequent occurrence in analyses of legal decisions—and, perhaps especially in work on gendered judging. In virtually every study in this field the researcher takes into account, in addition to the judges' sex, their ideology (or political party). This is a sensible choice of course: 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. Courts of Appeals judges who voted in sex discrimination cases between 1995-2002,[19] and using Judicial Common Space scores (Epstein et al., 2007) to measure their ideology, we can see that women skew to the left. Men, on the other hand, are more evenly dispersed between liberal and conservative groupings.[20]

Data of this sort are so profoundly imbalanced that regression analysis could produce profoundly misleading results. In concrete terms, because the range or distribution of ideol-

[17]Another limitation is that while our definition of a causal effect in Equation 1 does not require that the effects are constant across all observations, the linear regression model does (Rubin, 1973; Winship and Morgan, 1999; Greiner, 2006). 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 difficulty (see Greiner, 2006; Ho et al., 2007, for illustrations).

[18]This is the notion of common support, which is part of the assumption of "strong ignorability" (see note 16, and Smith, 1997; King and Zeng, 2006, p. 349).

[19]We use these data, which come from Sunstein et al. (2006), for our analyses as well. For more details on how Sunstein and his colleagues collected them, see Appendix B.

[20]Worth noting, ideology is not the only covariate for which observe substantial differences between the treatment and control groups. Others include the year of confirmation and political party.

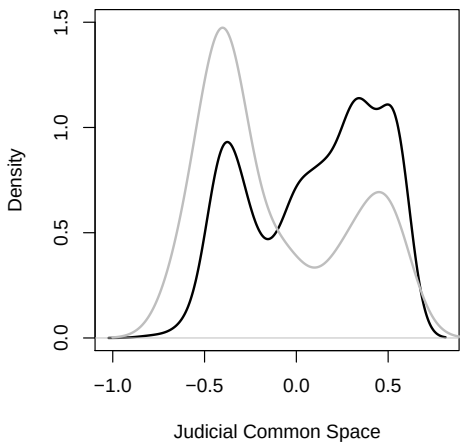

Figure 1: The ideology of U.S. Court of Appeals judges who voted in Title VII sex discrimination cases, 1995-2002, $N = 1245$. This is a kernel density plot that depicts the marginal distribution of the judge's ideology (measured using the Judicial Common Space), from most liberal to most conservative. The black line depicts the density for males judges; the grey line for female judges. Case data come from Sunstein et al. (2006) and ideology, from Epstein et al. (2007)

ogy 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 (for examples, see Greiner, 2006; Ho et al., 2007))! 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.[21]

2.2 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

[21]The profound 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.

often makes assumptions that are unjustified in the study of judging (Epstein et al., 2005). Indeed, it is entirely possible that the use of linear regression models to analyze data with a profound lack of balance in the covariates may explain the hodgepodge that the literature on sex and judging has become.

If naively using linear regression can lead to misleading inference, especially when we expect imbalance in and non-overlap of the covariates, what are the viable alternatives? The most promising is non-parametric matching—an approach intuitively 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., pre-treatment covariates),[22] 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 the statistical sciences. And, actually, one form of matching—exact matching—has even found its way into the literature on gendered judging (see, e.g., Walker and Barrow, 1985; Segal, 2000). The idea in matching of this sort 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 matching itself can become increasingly implausible.[23] 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 (say, 0.525 and 0.342 on the Judicial Common Space scale). To find an exact match

[22] See note 15.

[23] 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 Section 2.3.

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 brought by a female litigant and resolved in 1995 by a panel with two relatively right-of-center men—but on which a female judge, and not three males, sat. Because such an exact match might 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—would grow exponentially as we added more covariates, such as case facts and additional judge attributes.

To avoid wasting data, we must thus 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 pre-treatment 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 the pre-treatment covariates X_i, the idea is to obtain a single variable—the estimated propensity score—that serves as summary of the covariates on the treatment and control groups. With the propensity scores in hand, researchers can deploy them to match observations (using a variety of strategies discussed below) without making any of the strong parametric assumptions necessitated by linear regression.

2.3 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 ought 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 must 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). Or, under the common reframe of proponents of the Rubin causal model, "no causation without manipulation." In practice, this means that attributes, such as a judges' sex, *cannot be viewed as causes*. As Cox (1992, 296) puts it, "...in most situations gender is not a causal variable but rather an intrinsic property of the individual."

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 pre-treatment covariates, the average treatment effect is possible to estimate. 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—Do male and female judges decide cases differently?—presents two problems. First, because the treatment is the sex of the judge, most scholars would say that it fails to meet the standard of "no causation without manipulation." Second, all the other relevant covariates—whether centering on the judge's attributes (e.g., ideology, age) or the case's (e.g., sex of the plaintiff)—occur *after* the sex of the judge is determined. With only post-treatment 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 ought still harness the power of matching methods. As Rubin (2006, 3) 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.

We could make precisely the same claim about male-female judge pairs.

3 Implementing Propensity Score Matching: Covariates, Data, and Methods

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 animate the covariates, estimating the

propensity scores, and matching observations. Once we have non-parametrically processed the dataset in this way, we can, using simple difference of proportions tests as well as full, parametric models, summarize the difference in judging for the first question, and estimate the causal effect for the second (Ho et al., 2007).

None of the four steps presents much difficulty. To choose covariates, we took cues from the large and well-established literature on judging in the U.S. Courts of Appeals and incorporated both judge-based attributes (e.g., ideology as measured by the Judicial Common Space scores [or political party affiliation] and year of birth) and case-specific factors (e.g., year of decision, sex of plaintiff).[24]

Our data also come from a tried-and-true source, the Sunstein et al. (2006) project on the politics of judging. 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 all federal courts of appeal judges in Title VII sex discrimination cases between 1995 and 2002.[25] Sunstein and his colleagues graciously shared their data with us, and we in turn used the Zuk, Barrow and Gryski (2004) and Federal Judicial Center databases to code judge attributes (e.g., sex and age) of interest to us.

With the data in hand, we turned to the final steps: estimating propensity scores for each judge-vote in the Title VII cases (for individual and panel effects) and matching the

[24]To be transparent, the propensity score model for individual effects included: ideology, ideology^2, confirmation year, confirmation year^2, minority judge, minority judge $\times$ ideology, minority judge $\times$ confirmation year, and circuit dummies. When we turned to matching the observations, we exact matched on the year of the decision (to capture, e.g., any broad time effects, such as alterations in precedent at the circuit or Supreme Court level) and on the direction of the decision of the court below (to attend to the circuit courts' propensity to affirm). The propensity score model for panel effects included: ideology, ideology^2, minority judge $\times$ ideology, and circuit dummies (with exact matching on lower court direction, term, and judicial experience). Finally, the logit models we ultimately estimated (see Appendix C) incorporate, in addition to the primary variable of interest: the ideology (or party) of the judge, his or her year of birth, and whether she or he is a minority, as well as various case-specific factors: whether the plaintiff(s) was (were) a female, whether the plaintiff had been fired, whether the matter before the circuit was largely procedural, and whether there was a claim of pregnancy discrimination. We also include circuit and year fixed effects.

[25]See Appendix B for more details. Suffice it to note here that Sunstein and his colleagues (2006, 20-21) explored over 20 areas of the law, including the environment, campaign finance, contracts, and standing. Matching is possible in each and every one of these areas but in light of the theoretical literature on gendered judging—which emphasizes likely differences in male and female approaches to issues of concern to women—we explore only sex discrimination litigation.

observations (again, both 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 top panels of Figures 2 and 3 display the results of this exercise—the estimated propensity scores for the individual and panel analyses, respectively. Beginning with Figure 2, note the lack of common support: we observe no female judges in a broad propensity score area (roughly beyond -4). The problem, of course, is a lack of balance on many of the covariates, including party, ideology, and confirmation year (see Table 1).

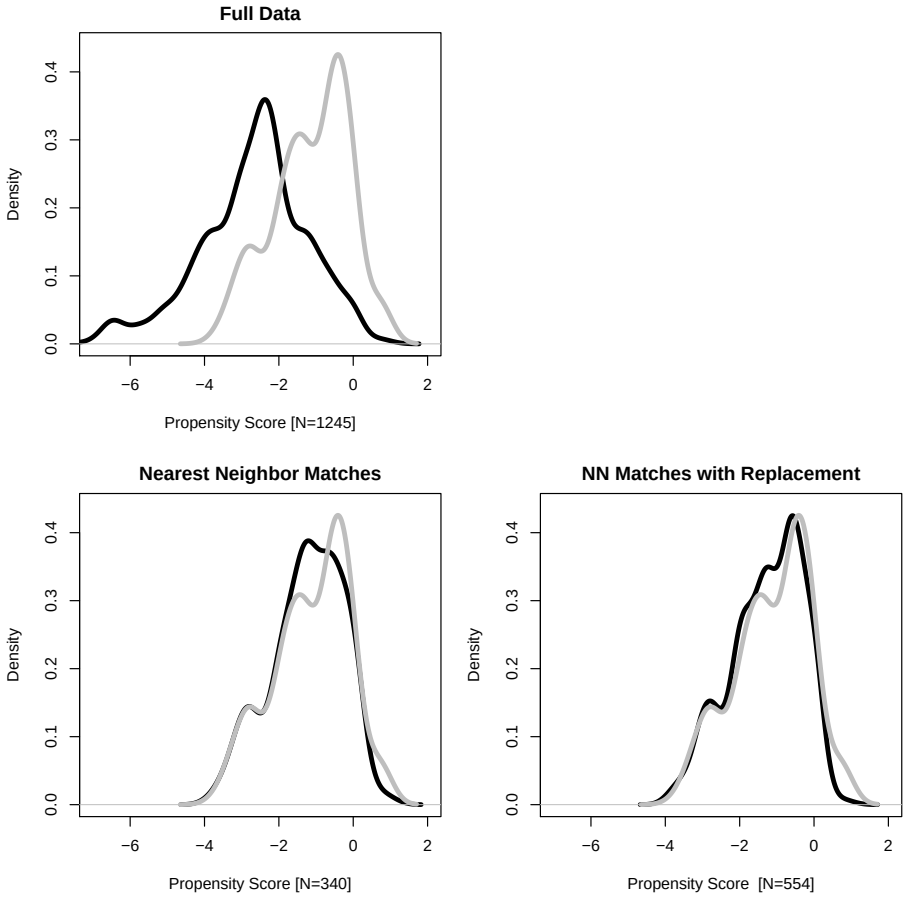

Figure 2: Kernel density plots of the estimated propensity score for the individual Title VII sex discrimination analysis. The black lines depict the density for males judges; the grey lines for female judges. The top panel is for the full dataset. The lower left panel is for nearest-neighbor matching without replacement. The lower right panel is for nearest-neighbor matching with replacement.

Turning to the panel effects, Figure 3 seems to indicate that common support is not much

	Nearest Neighbor Matching Without Replacement						
	Full Data				Matched Data		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	−1.13	−2.75	1.58	93.95	−1.13	−1.23	0.07
Minority Judge	0.12	0.09	0.00	20.44	0.12	0.15	0.00
Judge Party	0.68	0.32	0.00	90.33	0.68	0.65	0.00
Judicial Experience	0.45	0.45	0.00	na	0.45	0.42	0.00
Judicial Common Space	−0.12	0.10	0.16	87.38	−0.12	−0.10	0.03
Confirmation Year	1990.38	1984.58	6.00	93.81	1990.38	1990.02	1.00
	Nearest Neighbor Matching With Replacement						
	Full Data				Matched Data		
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
Judge Party	0.68	0.32	0.00	83.89	0.68	0.62	0.00
Judicial Experience	0.45	0.45	0.00	na	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

Table 1: Matching summary statistics for the individual Title VII sex discrimination analysis. The total number of observations in the full data is $N = 1245$. The top panel shows results from matching without replacement, resulting in $N = 340$ total observations. The bottom panel shows results from matching with replacement, resulting in $N = 554$ total observations. eQQ Med is the median difference in the empirical quantile-quantile plot (an eQQ Med of zero is ideal).

of an issue. And yet, as Table 2 shows, there are clear imbalances in the propensity score (and other covariates as well), with the mode for male judges below -2 and the modal score for female judges at -0.5.

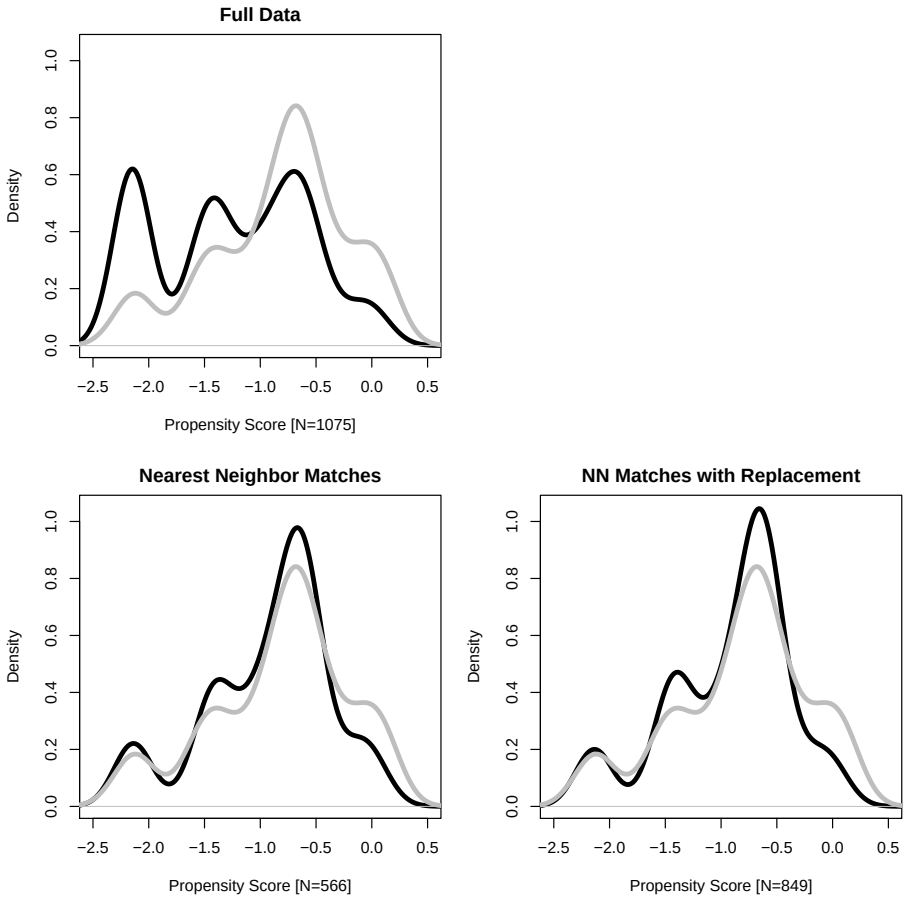

Figure 3: Kernel density plots of the estimated propensity score for the panel effects Title VII sex discrimination analysis. The black lines depict the density for all-male panels (control); the grey lines for mixed-sex panels (treatment). The top panel is for the full dataset. The lower left panel is for nearest-neighbor matching without replacement. The lower right panel is for nearest-neighbor matching with replacement.

To bring balance to the datasets, we moved to the last task: matching observations using "nearest-neighbor" matching. This approach selects for each "mixed-sex" observation (or female judge, for the individual analysis) the "all-male" observation (or male judge) that has the closest propensity score.[26] Nearest-neighbor matching can be implemented by matching observations from the control group (e.g., male judges on all-male panels) once

[26]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.

	Nearest-Neighbor Matching Without Replacement						
	Full Data				Matched Data		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	−0.83	−1.25	0.41	76.95	−0.83	−0.92	0.07
Judicial Experience	0.43	0.46	0.00	100.00	0.43	0.43	0.00
Minority Judge	0.08	0.10	0.00	17.88	0.08	0.07	0.00
Judge Party	0.34	0.31	0.00	57.74	0.34	0.35	0.00
Judicial Common Space	0.09	0.11	0.02	58.41	0.09	0.08	0.01
Confirmation Year	1984.66	1984.55	0.00	na	1984.66	1983.76	1.00
	Nearest-Neighbor Matching With Replacement						
	Full Data				Matched Data		
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
Judge Party	0.34	0.31	0.00	62.43	0.34	0.33	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	na	1984.66	1983.76	1.00

Table 2: Matching summary statistics for the panel effects Title VII sex discrimination analysis. The total number of observations in the full data is $N = 1075$. The top panel shows results from matching without replacement, resulting in $N = 566$ total observations. The bottom panel shows results from matching with replacement, resulting in $N = 850$ total observations. eQQ Med is the median difference in the empirical quantile-quantile plot (an eQQ Med of zero is ideal).

(matching without replacement) or multiple times (matching with replacement).[27] We used both. Finally, for the individual- and panel-effects analyses, we forced all observations to match exactly on the year the panel decided the case and on the direction of the lower court decision; for the latter we also exact matched on judicial experience (see note 24).

The bottom panels of Figures 2 (individual effects) and 3 (panel effects) visually depict the results of this matching exercise. For the individual analysis, observe that our matched data come only from the region of common support, from -4 to 2. Note too that the shape

[27]There are a number of ways to approach matching with or without the propensity score (for a review see Diamond and Sekhon, 2005; Ho et al., 2007), and many debates exist in the literature about which approach is best. For our analyses, we have estimated the propensity score in a number of ways, and matched using different methods, all of which yield results comparable to those reported here.

of the propensity score distribution for the treatment and control groups is essentially the same, with the nearest-neighbor with replacement producing slightly better matches in the local modes in the propensity score distribution. For the panel effects analysis in Figure 3, the distribution of the propensity score for the treatment and control groups also appears quite similar, suggesting that balance has been achieved.

Reinforcing these visual displays are the summary statistics for each matching exercise in Tables 1 and 2. The left-hand columns in both highlight the profound imbalance in many covariates in the full, unmatched dataset. But note the percent reduction statistics and the eQQ medians (a 100% percent reduction and an eQQ median score of zero is indicative of perfect balance), both of which show that for nearly all covariates matching greatly improved balance.[28]

4 Empirical Results

With the now-balanced datasets in hand (along with weights necessary for subsequent analyses), we can turn 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 means test (e.g. Smith, 1997)—or, in our case, a difference of proportions test—because the data are now balanced. Others recommend proceeding as judicial specialists typically would: parametrically processing the now non-parametrically balanced database (Robins and Rotnitzky, 2001; Ho et al., 2007). This strategy has the advantage of being "doubly robust" (Robins and Rotnitzky, 2001): if either the dataset is sufficiently balanced or the parametric model is properly specified, accurate estimates of the causal effect will result.

In what follows, we do both. To investigate the extent to which we observe individual and panel effects, we estimate various models—simple and complex—hoping to unearth

[28]Balance slightly worsens for the minority judge variable in the individual analysis and for confirmation year in the panel effects analysis after matching with replacement, but these differences are not statistically significant. Worth noting too is that our matched datasets have far fewer observations than our full dataset. Although 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 Section 2.1.

consistent results regardless of the method. As it turns out, this is (almost) precisely what obtains.

4.1 Individual Results

Beginning with the question of whether male and female judges differ in their decisions over Title VII cases, we took nine different approaches. The first three were logistic regressions using the full (unbalanced) dataset: a bivariate, with the sex of the judge as the only covariate (the equivalent of a difference of proportions test); and two fully specified models, one incorporating the judges' political party affiliation and the other substituting ideology for partisanship.[29]

We display the resulting average treatment effects (ATEs) for each in Figure 4.[30] Note that the reported effects using the full (unbalanced) dataset do not lend much support to the "in a different voice" theory of judging: the 95% confidence intervals surrounding the ATEs include zero for the fully specified estimates, meaning that we cannot eliminate the possibility of no difference between men and women.

Because this (non)-result may be due to imbalances in the full data set, we moved to the matched data and estimated the ATE for pairs constructed (via nearest neighbor matching) with and without replacement. For both estimations we performed identical analyses, exploring the effect of gender alone (using a weighted logistic regression with only the treatment on the right hand side—the equivalent of a difference of proportion test) and using fully specified logistic regressions with the same compliment of covariates. The results, as we show in Figure 4, are also nearly identical. Matching with or without replacement and incorporating controls uncovers an ATE of about 0.10, regardless of whether we estimate the full model with party or ideology. In other words, *the probability of a judge deciding a sex discrimination case in favor of the plaintiff decreases by about 10 percentage points when the*

[29]Because these variables are so highly correlated, we only include one at a time in the analysis. In general, our preference is to use the continuous Judicial Common Space scores (Epstein et al., 2007). And that seems to be the preference of other scholars as well; indeed, many explicitly say that they incorporate the judge's party to serve as a proxy for ideology. In addition to the treatment and the ideology (or partisanship), all fully specified models house the covariates identified in Appendix C, along with fixed effects for the decision year and the circuit.

[30]Appendix C supplies the statistical estimates for all nine models.

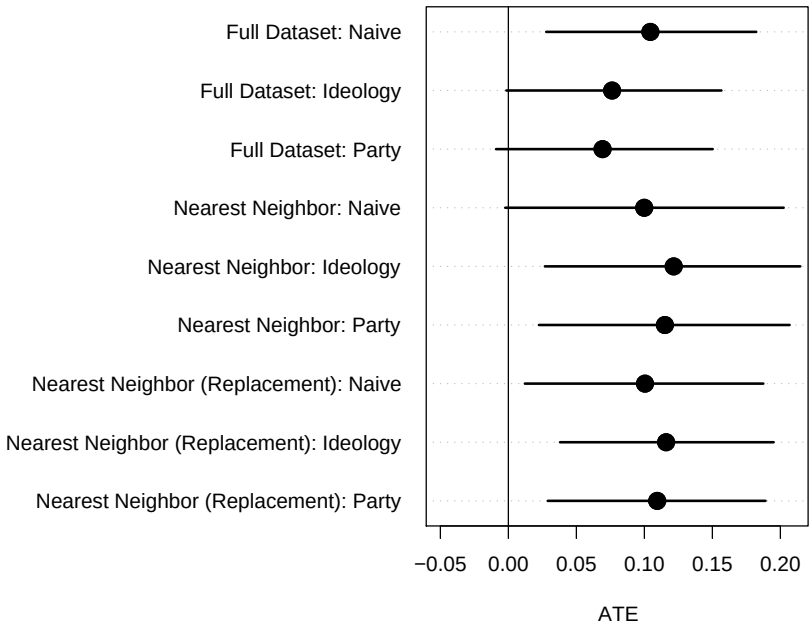

Figure 4: Dotplot of average treatment effects (ATEs) for individual effects in Title VII sex discrimination cases. The lines represent 95% confidence intervals for the average treatment effect. The first three models are logistic regression models fit to the full dataset (N=1245). The naive model includes only the judge's sex as a covariate. The other two models include the judge's sex, a number of controls (see note 24), and a measure of ideology (either party affiliation or Judicial Common Space [JCS] score). The next three models show the ATE after nearest-neighbor matching on the estimated propensity score without replacement (N=340). The first is for a difference of proportions analysis. The next two are for logistic regression models with the judge's sex, a number of controls, and either the partisan affiliation or JCS score. The final three models show the ATEs for nearest-neighbor matching with replacement (N=554).

judge is a male.

This is a rather large effect—and, of course, one that would have gone undetected under the standard operating procedure in this field (i.e., estimating a logit model with unbalanced data). It is also, substantively speaking, an important effect. Figure 5 nicely makes this point. Here we depict the predicted probabilities of men and women casting pro-plaintiff votes as a function of their ideology (and based on the "with replacement" analysis). Observe that the estimated probability of a female judge casting a pro-plaintiff is close to 0.40 at the highest levels of liberalism; for even the most left-of-center male, that figure just barely exceeds 0.20.

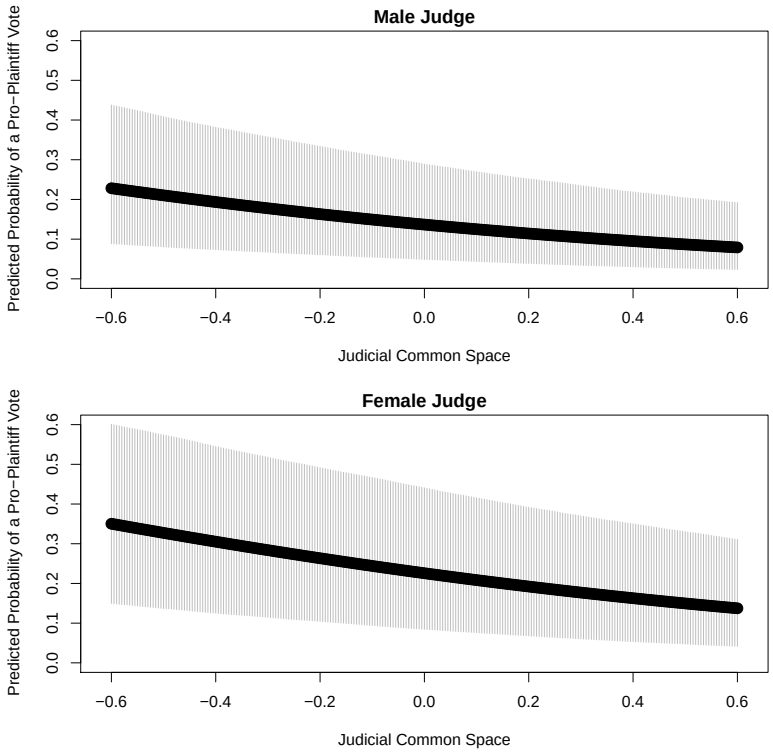

Figure 5: Predicted probabilities of pro-plaintiff votes in Title VII sex discrimination cases as a function of the Judicial Common Space (ideology) for male and female judges, individual effects. The Judicial Common Space runs from most liberal (here, -0.6) to most conservative (0.6). These estimates are from the weighted logistic regression model on the nearest neighbor with replacement matched data (see Appendix C). All continuous variables are held at their sample means; other variables are at their sample modes. The vertical grey lines denote 95% confidence intervals.

4.2 Panel Effects

If our descriptive analysis of the differences between male and female judges encourages proponents of gender-difference theories, our assessment of panel effects brings even more good news. As we show in Figure 6, for not one model does the 95% confidence interval come near the zero line (indicating no difference between all-male and mixed-sex panels).[31] Rather, we observe large causal effects, ranging from 0.12 to 0.16 —*meaning that the probability of a male judge ruling in favor of the plaintiff increases by 12% to 16% when a female sits on the panel.*

[31]For the statistical results, see Appendix C, Table 5.

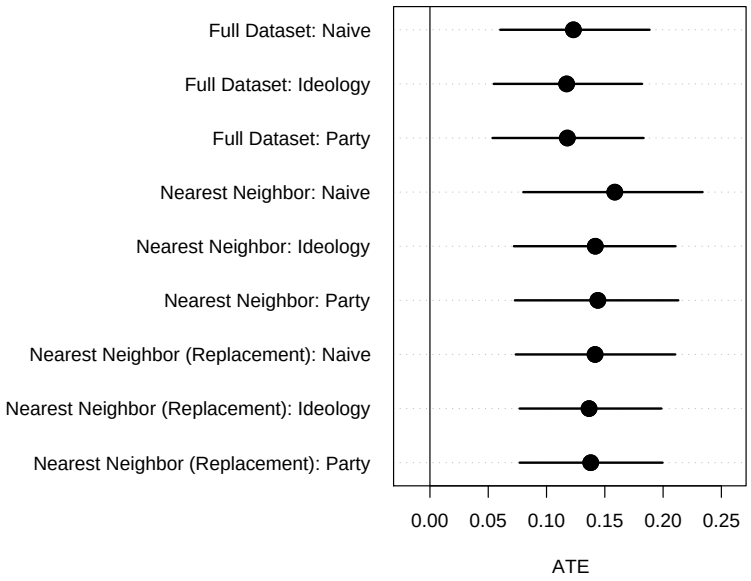

Figure 6: Dotplot of average treatment effects (ATEs) for panel effects in Title VII sex discrimination cases. The lines represent 95% confidence intervals for the average treatment effect. The first three models are logistic regression models fit to the full dataset (N=1245). The naive model includes only the judge's sex as a covariate. The other two models include the judge's sex, a number of controls (see note 24), and a measure of ideology (either party affiliation affiliation or Judicial Common Space [JCS] score). The next three models show the ATE after nearest-neighbor matching on the estimated propensity score without replacement (N=566). The first is for a difference of proportions analysis. The next two are for logistic regression models with the judge's sex, a number of controls, and either the partisan affiliation or JCS score. The final three models show the ATE for nearest-neighbor matching with replacement (N=850).

This causal effect of panel composition is not just strong but it is consistent as well. Regardless of the particular form of the analysis, the sample of data, or the matching method, it persists. It is also 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,[32] it ought to mitigate against a finding of a strong causal panel effect. Consequently, our findings, however substantial, may actually *underestimate* the impact of panel composition on outcomes.

[32]We leave this matter for future research, though initial evidence seems to suggest that, in fact, female judges encourage settlement (see, e.g., Boyd, 2006).

Finally, the causal effect of panel composition—like the female-male individual differences—is quite consequential. Underscoring this point is Figure 7, in which we display the predicted probabilities of males in the control and in the treatment groups voting to support plaintiffs, by their ideology. Notice that in the former (i.e., all-male panels) the probability never exceeds 0.20 for even the most liberal males but for mixed sex panels the probability never falls below 0.20 for even the most conservative males. For males at more average levels of ideology, the likelihood of supporting the plaintiff increases by almost 85 percent when they sit with a female judge.

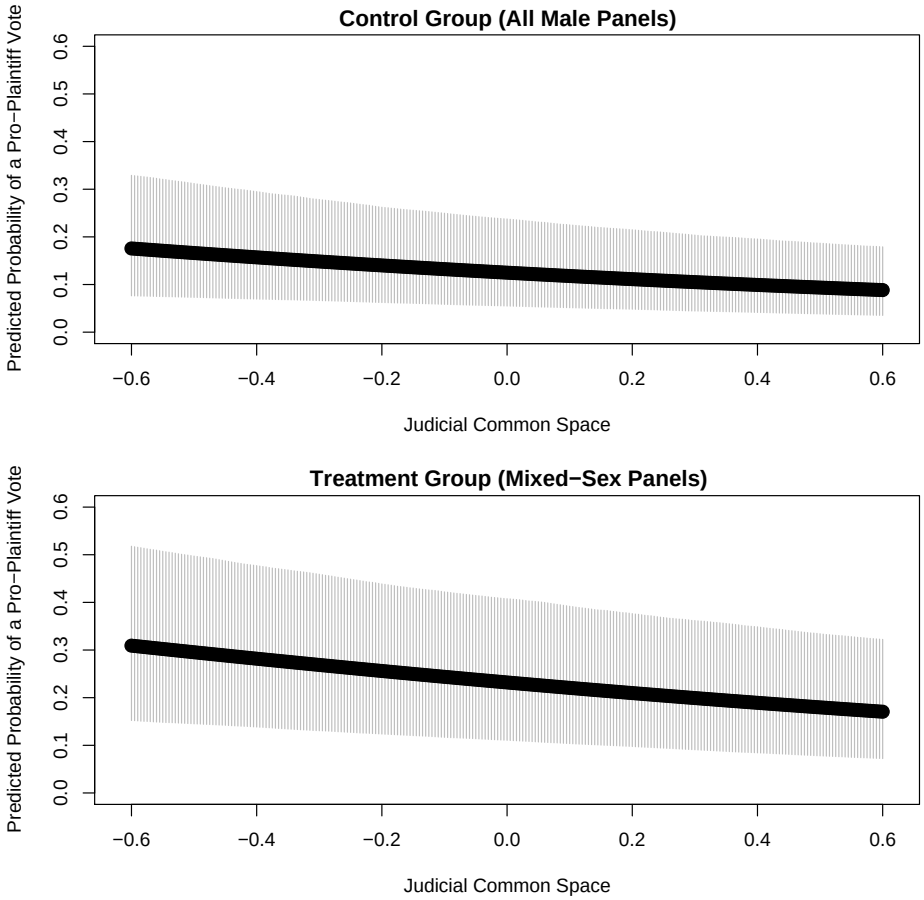

Figure 7: Predicted probabilities of pro-plaintiff votes in Title VII sex discrimination cases as a function of the Judicial Common Space (ideology) for all-male (control) and mixed-sex (treatment) panels. The Judicial Common Space runs from most liberal (here, -0.6) to most conservative (0.6). These estimates are from the weighted logistic regression model on the nearest neighbor with replacement matched data (see Appendix C). All continuous variables are held at their sample means; other variables are at their sample modes. The vertical grey lines denote 95% confidence intervals.

5 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 a matter for normative theorists, the second is susceptible to empirical scrutiny. And that is what we have attempted to give it here. Proceeding from a formal framework for causal inference, we have deployed the best available methods and data

to answer questions that have long dominated scholarly and policy discourse over the role of sex in judging. The results of this exercise are clear: proponents of gender difference have the better case. Not only do males and females bring distinct approaches to sex discrimination 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.

Seen in this way, our study, we hope, brings closure to a decades-old debate. But just as surely it opens up a new line of inquiry. Because individual effects are, in some sense, dependent on the gender make-up of a panel, scholars should turn their attention to explaining those very panel effects. A number of promising leads come from studies of the partisan composition of panels (see e.g., Cross and Tiller, 1998; Farhang and Wawro, 2004; Cameron and Cummings, 2003; Sunstein et al., 2006)—most of which center on the judges themselves, such as learning and suppressing. Another possibility is framing. On this account, the effects we observe do not implicate judges and their hierarchy as much as they do lawyers and the way they *frame* their oral arguments in response to those judges. When confronted with a mixed-sex panel, if defense attorneys, out of fear of alienating the female judge, dial down their claims, this may account for our results.

Bits and pieces of evidence from the extant literature tend to support this conjecture of framing (Beiner, 2002; Seron, 1997; Peresie, 2005) but we know of no systematic and rigorous tests. Conducting them should now be a high priority for analysts concerned with gendered judging, as should assessments of the other plausible mechanisms.

Also deserving a prominent place on the scholarly agenda are further studies of the very questions we considered here. While we believe we have provided conclusive support for gender effects in the cases where they may be most likely to emerge—sex discrimination suits[33]—it seems worthwhile to consider several other areas of the law. Coming immediately to mind, given the extant literature depicted in Appendix A, are criminal law cases and disputes implicating the family. Moreover, we could easily imagine extending analyses to cover other courts, both here and abroad, as well as to other attributes, including race, religion, and age.

[33]Then again, reconsider our discussion of selection bias in Section 4.2.

We certainly commend these challenges to scholars working in the fields of public law, gender politics, and race and ethnicity. Going forward, we also commend the framework and methods deployed here—as do other political scientists, who too urge a reconsideration of the traditional and dominant approach to inference (e.g., Greiner, 2006; Ho et al., 2007; Epstein et al., 2005). Their message seems especially timely in light of promising developments in the statistical sciences aimed at improving the conclusions we can draw from observational data.

References

Abrahamson, Shirley S. 1984. "The Woman Has Robes: Four Questions." *Golden Gate Law Review* 14:489–499.

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

Ashenfelter, Orley, Theodore Eisenberg and Stewart J. Schwab. 1995. "Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes." *Journal of Legal Studies* 24:257–281.

Baldez, Lisa, Lee Epstein and Andrew D. Martin. 2006. "Does the U.S. Constitution Need an ERA?" *Journal of Legal Studies* 35:243–283.

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

Bogoch, Bryna and Rachelle Don-Yechiya. 1999. *Gender in Justice: Bias Against Women in Israeli Courts*. Jerusalem, Israel: Jerusalem Institute for Israel Research.

Boucher, Jr., Robert L. and Jeffrey A. Segal. 1995. "Supreme Court Justices as Strategic Decision Makers: Aggressive Grants and Defensive Denials on the Vinson Court." *Journal of Politics* 57:824–837.

Boyd, Christina L. 2006. "The Effect of Judge Sex on Case Disposition Method in Federal District Courts." Paper presented at the 2006 annual Conference for Empirical Legal Studies, Austin TX.

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* 60:1675–1766.

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

Cameron, Charles and Craig Cummings. 2003. "Diversity and Judicial Decision Making on the U.S. Courts of Appeals." Unpublished ms., available: http://www.yale.edu/coic/CameronCummings.pdf.

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

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.

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

Cross, Frank B. and Emerson H. Tiller. 1998. "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals." *Yale Law Journal* 107:2155–2176.

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. thesis, University of Chicago.

Davis, Sue, Susan Haire and Donald R. Songer. 1993. "Voting Behavior and Gender on the U.S. Courts of Appeals." *Judicature* 77:129–133.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94:1053–1062.

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 Meeting, Florida State University.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal and Chad Westerland. 2007. "The Judicial Common Space." *Journal of Law, Economics & Organization,* forthcoming.

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

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

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* 91:903–966.

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* 20:299–330.

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

Fox, Richard and Robert Van Sickel. 2000. "Gender Dynamics and Judicial Behavior in Criminal Trial Courts: An Exploratory Study." *The Justice System Journal* 21:261–280.

Garrison, Marsha. 1995. "How do Judges Decide Divorce Cases? An Empirical Analysis of Discretionary Decision Making." *North Carolina Law Review* 74:401–552.

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

Greiner, D. James. 2006. "Causal Inference in Civil Rights Litigation." Unpublished manuscript, available at: http://people.iq.harvard.edu/∼jgreiner/Papers.html.

Gryski, Gerard, Eleanor C. Main and William J. Dixon. 1986. "Models of State High Court Decision Making in Sex Discrimination Cases." *Journal of Politics* 48:143–155.

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,* forthcoming.

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

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–970.

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

Kay, Herma Hill and Geraldine Sparrow. 2001. "Does Gender Make a Difference?" *Wisconsin Women's Law Journal* 16:1–13.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14:131–159.

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

Kulik, C.T., E.L. Perry and Molly Pepper. 2003. "Here Comes the Judge: The Influence of Judge Personal Characteristics on Federal Sexual Harassment Court Decisions." *Law and Human Behavior* 27:69–86.

Manning, Kenneth L. 2004. "Como Decide? Decision-Making by Latino Judges in the Federal Courts." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago IL.

Martin, Elaine. 1990. "Men and Women on the Bench: Vive La Difference?" *Judicature* 73:204.

Martin, Elaine and Barry Pyle. 2000. "Gender, Race and Partisanship on the Michigan Supreme Court." *Albany Law Review* 63:1205–1236.

Martin, Elaine and Barry Pyle. 2005. "State High Courts and Divorce: The Impact of Judicial Gender." *University of Toledo Law Review* 36:923–947.

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

Martinek, Wendy L. 2003. "The Effect of Confirmation Politics on the United States Courts of Appeals Decision Making." Paper presented at the annual meeting of the American Political Science Association, Philadelphia PA.

Massie, Tajuana, Susan W. Johnson and Sara Margaret Green. 2002. "The Impact of Gender and Race in the Decisions of Judges on the United States Courts of Appeals." Paper presented at 2002 annual meeting of Midwest Political Science Association, Chicago IL.

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

McCall, Madhavi. 2005. "Court Decision Making in Police Brutality Cases, 1990-2000." *American Political Research* 33:56–80.

Peresie, Jennifer L. 2005. "Female Judges Matter: Gender and Collegial Decisionmaking in the Federal Appellate Courts." *Yale Law Journal* 114:1759–1790.

Robins, James M. and Andrea Rotnitzky. 2001. "Comment on the Bicket and Kwon article, 'Inference for Semiparametric Models: Some Question and an Answer'." *Statistica Sinica* 11:920–936.

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):516–524.

Rubin, Donald B. 1973. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159–183.

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

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

Schneider, Daniel M. 2002. "Assessing and Predicting Who Wins Federal Tax Trial Decisions." *Wake Forest Law Review* 37:473–529.

Segal, Jennifer A. 2000. "Representative Decision Making on the Federal Bench: Clinton's District Court Appointees." *Political Research Quarterly* 53:137–150.

Seron, Carroll. 1997. "A Report of the Perceptions and Experiences of Lawyers, Judges and Court Employees Concerning Gender, Racial and Ethnic Fairness in the Federal Courts of the Second Circuit of the United States." *Annual Survey of American Law* 1997:419–528.

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

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

Smith, Fred O. 2005. "Gendered Justice: Do Male and Female Judges Rule Differently on Questions of Gay Rights?" *Stanford Law Review* 57:2087–2134.

Smith, Herbert L. 1997. "Matching With Mulitple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27(1):325–353.

Songer, Donald R. and Kelley A. Crews-Meyer. 2000. "Does Judge Gender Matter? Decision Making in State Supreme Courts." *Social Science Quarterly* 81:750–762.

Songer, Donald R., Sue Davis and Susan Haire. 1994. "A Reappraisal of Diversification in the Federal Courts: Gender Effects in the Courts of Appeals." *Journal of Politics* 56:425–439.

Steffensmeier, Darrell and Chris Herbert. 1999. "Women and Men Policymakers: Does the Judge's Gender Affect the Sentencing of Criminal Defendants?" *Social Forces* 77:1163–1196.

Sullivan, Kathleen M. 2002. "Constitutionalizing Women's Equality." *California Law Review* 90:735–764.

Sunstein, Cass R., David Schkade and Lisa Ellman. 2004. "Ideological Voting on Federal Courts of Appeals: A Preliminary Investigation." *Virginia Law Review* 90:301–354.

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

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

Westergren, Sarah. 2004. "Gender Effects in the Courts of Appeals Revisited: The Data Since 1994." *Georgetown Law Journal* 92:689–708.

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

Zuk, Gary, Deborah J. Barrow and Gerard S. Gryski. 2004. "Appeals Court Attribute Database." Available at: http://www.as.uky.edu/polisci/ulmerproject/auburndata.htm.

Appendix A. Summary of Multivariate Analyses of the Effect of Sex on Judging

Study (Court)	Methods & Design	Findings
Ashenfelter, Eisenberg and Schwab (1995) (U.S. District Courts)	Regression analysis of civil rights and prisoner cases (1980-1981). Covariates include political party and prior judicial experience.	No difference between males and females.
Baldez, Epstein and Martin (2006) (state courts of last resort)	Regression analysis of sex discrimination decisions (1960-1999). Covariates include ideology and case characteristics.	The more female justices on the court, the more likely the court was to rule for the plaintiff.
Bogoch and Don-Yechiya (1999) (Israeli district and magistrate's courts)	Regression analysis of "serious violent crime" cases (1988-1993).	No differences between male and female justices on the decision to convict; females tend to give lower jail sentences but not when they sit on panels with males.
Boyd (2006) (U.S. District Courts)	Regression analysis of personal injury and civil rights terminations. Covariates include judicial experience, race, and ideology.	Female judges settle cases more than males.
Brudney, Schiavoni and Merrit (1999) (U.S. Courts of Appeals)	Regression analysis of labor law decisions (1986-1993). Covariates include political party, gender, race, and career experience.	Republican females more likely to support unions that Republican males but no difference between Democratic men and women.
Cameron and Cummings (2003) (U.S. Courts of Appeals)	Regression analysis of affirmative action cases (1971-1999). Covariates include the race, age, birth cohort, ideology and case specific factors.	Sex of the judge has no clear impact on the judge's own behavior, nor does it give rise to panel effects.
Crowe (1999) (U.S. Courts of Appeals)	Multivariate analysis of sex and race discrimination litigation (1981-1996). Covariates include political party and race.	Females more likely to support plaintiff in sex discrimination cases; no panel effects based on gender.
Davis, Haire and Songer (1993); Songer, Davis and Haire (1994) (U.S. Courts of Appeals)	Regression analysis of search and seizure, obscenity, and employment discrimination cases (1981-1990). Covariates include political party.	Females more likely to support plaintiff in employment discrimination litigation but no differences in other areas of the law.
Farhang and Wawro (2004) (U.S. Courts of Appeals)	Regression analysis of employment discrimination litigation (1998-1999). Covariates include ideology and race.	Males sitting on a panel with a female are more likely to find for the plaintiff.

Fox and Sickel (2000) (state trial courts)	Regression analysis of selected criminal law cases. Covariates include age and judicial style.	Males are more likely to find for the defendant.
Garrison (1995) (state trial courts)	Regression analysis of divorce-related cases (1978 and 1984). Judge related covariates include political party, age, religion, previous experience, and eduction.	No differences between male and female judges on alimony awards but differences on child support.
Gryski, Main and Dixon (1986) (state courts of last resort)	Regression analysis of sex discrimination cases (1971-1981). Covariates include the political party, age, and tenure.	No differences between male and female judges.
Kulik, Perry and Pepper (2003) (U.S. District Courts)	Regression analysis of sexual harassment litigation (1981-1996). Covariates include political party affiliation, race and age.	No differences between male and female judges.
Manning (2004) (U.S. District Courts)	Regression analysis of age discrimination cases (1984-1995). Covariates include age, race, and political party.	No differences between male and female judges.
Martin and Pyle (2000) (Michigan Supreme Court)	Regression analysis of discrimination, divorce, and feminist issues (1985-1998). Covariates include race and political party.	Females more likely to vote liberally in divorce cases; no difference between men and women in other issue areas.
Martin and Pyle (2005) (state courts of last resort)	Regression analysis of divorce cases (1998-1999). Covariates include age, race, party, and judicial experience.	Females tend to rule for mothers over fathers; a female justice on the court increases the likelihood of males ruling in favor of the mother.
Martinek (2003) (U.S. Courts of Appeals)	Regression analysis of civil rights and liberties cases (1977-1996). Covariates include race.	No differences between male and female judges.
Massie, Johnson and Green (2002) (U.S. Courts of Appeals)	Regression analysis of civil rights, liberties, and justice cases (1977-1996). Covariates include political party.	Females are more conservative in criminal cases and more liberal in civil rights and liberties suits; males sitting on a panel with a female tend to vote more similarly to the female.
McCall (2005) (state courts of last resort)	Regression analysis of police brutality cases (1998-1999). Covariates include ideology and selection system used in the state.	Females are more likely to find the defendant's favor.
Peresie (2005) (U.S. Courts of Appeals)	Regression analysis of statutory sex discrimination cases (1999-2001). Covariates include party of the appointing president, ideology of the judge, and background experience.	Female judges more likely to find for the plaintiff; males sitting on a panel with a female are more likely to find for the plaintiff.

Schneider (2002) (Tax Court and U.S. District Courts)	Regression analysis of disputes with the IRS (1979-1998). Covariates include race, political party, and length of service.	Female Democrats more likely to rule in favor of the taxpayer.
Segal (2000) (U.S. District Courts)	Matching analysis on women's policy issues, with (13 Clinton appointee) pairs based on sex, race, and district.	Females less likely than males to support women's rights issues.
Sisk, Heise and Morriss (1998) (U.S. District Courts)	Regression analysis of cases questioning the constitutionality of the sentencing guidelines (1988). Covariates include political party and race.	No differences between male and female judges.
Smith (2005) (federal and state appellate courts)	Regression analysis of gay rights litigation (1983-2003). Covariates include political party and birth cohort.	Female judges more willing to strike down laws adverse to gays.
Songer and Crews-Meyer (2000) (state courts of last resort)	Regression analyses of obscenity and death penalty cases (1982-1993). Covariates include political party.	Democratic females are more liberal in both areas than Democratic males but gender differences (individual and panel effects) did not emerge for Republicans.
Steffensmeier and Herbert (1999) (state trial courts)	Regression analyses of criminal sentencing cases (1991-1993). Covariates include case and defendant characteristics.	Females more likely to incarcerate a defendant and hand out longer sentences.
Walker and Barrow (1985) (U.S. District Courts)	Matching analysis across many areas of the law, with (29 Carter appointee) pairs based on race, sex, and district.	Some differences based on gender but none in cases implicating "women's issues."
Westergren (2004) (U.S. Courts of Appeals)	Regression analysis of (statutory) sex discrimination cases (1994-2000). Covariates include political party and race.	No differences between male and female judges.

Appendix B. The Sunstein et al. Database

The sex discrimination cases included in this study were originally compiled by Sunstein, Schkade and Ellman (2004) and are limited to three-judge panel published opinions issued in the U.S. Courts of Appeals between 1995 and 2001. The Lexis-Nexis search used to retrieve the cases was "sex! discrimination".

With the Sunstein et al. database in hand, we coded all opinions for the presence of common facts and legal issues; we also controlled for the basis of the suit by limiting our analysis to Title VII claims (78% of the cases in the dataset). Finally, we amassed background information on the individual panelists in each case. Judge-specific variables were collected largely from Zuk, Barrow and Gryski (2004) and the Federal Judicial Center's Biographical Directory of Federal Judges (2007). The notable exception was the judges' ideology. For courts of appeals judges, we retrieved their Judicial Common Space scores from the Epstein et al. (2007) project website (at `http://epstein.law.northwestern.edu/research/JCS.html`). For district court judges sitting by designation in the courts of appeals, we computed their scores using the methodology developed and described in Epstein et al. (2007).

Appendix C. Logistic Regression Estimates for Individual and Panel Analyses

	Full Dataset			Nearest Neighbor			Nearest Neighbor (Replacement)		
Covariates	Naive	Ideology	Party	Naive	Ideology	Party	Naive	Ideology	Party
Intercept	−0.678	16.078	18.236	−0.658	93.151	88.363	−0.662	72.542	69.727
	(0.065)	(13.204)	(13.094)	(0.162)	(30.982)	(31.166)	(0.103)	(23.381)	(23.500)
Sex	0.441	0.381	0.349	0.422	0.658	0.632	0.426	0.634	0.609
	(0.167)	(0.200)	(0.200)	(0.224)	(0.261)	(0.262)	(0.186)	(0.218)	(0.220)
Ideology or Party		−0.669	0.555		−0.924	0.945		−1.077	1.020
		(0.199)	(0.146)		(0.367)	(0.291)		(0.282)	(0.225)
Year of Birth		−0.009	−0.010		−0.049	−0.046		−0.038	−0.037
		(0.007)	(0.007)		(0.016)	(0.016)		(0.012)	(0.012)
Minority Judge		0.360	0.283		0.123	−0.019		0.398	0.268
		(0.222)	(0.226)		(0.383)	(0.391)		(0.287)	(0.292)
Female Plaintiff		0.235	0.233		0.508	0.495		0.725	0.722
		(0.180)	(0.180)		(0.335)	(0.337)		(0.276)	(0.277)
Plaintiff Terminated		−0.296	−0.293		−0.099	−0.116		−0.258	−0.270
		(0.138)	(0.138)		(0.269)	(0.271)		(0.211)	(0.212)
Mostly Procedural		0.072	0.077		0.683	0.690		0.063	0.058
		(0.219)	(0.219)		(0.415)	(0.417)		(0.317)	(0.320)
Pregnancy		0.799	0.786		0.087	0.070		0.618	0.587
		(0.267)	(0.267)		(0.515)	(0.519)		(0.405)	(0.408)
Lower Court Direction		1.168	1.168		1.050	1.078		1.118	1.142
		(0.151)	(0.151)		(0.331)	(0.334)		(0.254)	(0.255)
N:	1245	1245	1245	340	340	340	590	590	590
Log-Likelihood:	−803.422	−714.884	−713.366	−225.764	−194.109	−191.872	−345.578	−287.788	−285.670

Table 4: Logistic regression estimates for the Title VII sex discrimination cases, individual effects. Average treatment effects reported in Figure 4 are derived from these estimates. Standard errors are in parentheses. Estimates of circuit and year fixed effects are not reported to conserve space. The naive models include only the judge's sex as a covariate. The other models include the judge's sex, either ideology or party, and the other reported covariates.

	Full Dataset			Nearest Neighbor			Nearest Neighbor (Replacement)		
Covariates	Naive	Ideology	Party	Naive	Ideology	Party	Naive	Ideology	Party
Intercept	−0.827	8.249	11.207	−1.002	4.887	8.412	−0.917	11.292	14.953
	(0.077)	(14.103)	(13.924)	(0.134)	(19.533)	(19.321)	(0.094)	(15.827)	(15.639)
Treatment	0.535	0.599	0.605	0.710	0.784	0.804	0.625	0.767	0.774
	(0.143)	(0.165)	(0.165)	(0.180)	(0.203)	(0.204)	(0.152)	(0.174)	(0.175)
Ideology or Party		−0.646	0.506		−0.787	0.726		−0.672	0.555
		(0.225)	(0.162)		(0.306)	(0.226)		(0.258)	(0.189)
Year of Birth		−0.005	−0.007		−0.004	−0.006		−0.007	−0.009
		(0.007)	(0.007)		(0.010)	(0.010)		(0.008)	(0.008)
Minority Judge		0.393	0.335		0.496	0.412		0.849	0.789
		(0.251)	(0.255)		(0.397)	(0.398)		(0.336)	(0.338)
Female Plaintiff		0.176	0.173		0.684	0.674		0.666	0.661
		(0.197)	(0.197)		(0.271)	(0.271)		(0.236)	(0.235)
Plaintiff Terminated		−0.302	−0.294		−0.042	−0.021		−0.142	−0.133
		(0.151)	(0.151)		(0.210)	(0.211)		(0.173)	(0.174)
Mostly Procedural		0.073	0.080		0.392	0.379		0.060	0.059
		(0.243)	(0.243)		(0.337)	(0.337)		(0.305)	(0.304)
Pregnancy		0.891	0.879		0.338	0.318		0.432	0.409
		(0.293)	(0.294)		(0.429)	(0.428)		(0.347)	(0.347)
Lower Court Direction		1.215	1.215		1.102	1.091		1.153	1.147
		(0.163)	(0.163)		(0.247)	(0.248)		(0.203)	(0.203)
N:	1075	1075	1075	566	566	566	843	843	843
Log-Likelihood:	−679.831	−605.226	−604.524	−357.833	−308.007	−306.160	−517.052	−440.333	−439.923

Table 5: Logistic regression estimates for the Title VII sex discrimination cases, panels effects. Average treatment effects reported in Figure 6 are derived from these estimates. Standard errors are in parentheses. Estimates of circuit and year fixed effects are not reported to conserve space. The naive models include only the treatment (mixed-sex panel) as a covariate. The other models include the treatment, either ideology or party, and the other reported covariates.