ECONOMIC DEVELOPMENT & INSTITUTIONS

EDI WORKING PAPER SERIES

DO JUDGES FAVOR THEIR OWN ETHNICITY AND GENDER: EVIDENCE FROM KENYA

Daniel Chen
Bilal Siddiqi
Jimmy Graham
Manuel Ramos Maqueda
Shashank Singh

August 2021

Oxford Policy Management

UKaid from the British people

UNIVERSITÉ DE NAMUR

PARIS SCHOOL OF ECONOMICS ÉCOLE D'ÉCONOMIE DE PARIS

ADE ANALYSIS FOR ECONOMIC DECISIONS

Abstract

Evidence from high-income countries suggests that judges often exhibit in-group bias, favoring litigants that share an identity with the judge. However, there is little evidence on this phenomenon from the Global South. This paper examines the extent of in-group bias along gender and ethnic lines in the Kenyan judiciary. We find that judges display both gender and ethnic in-group bias towards defendants.In contrast, we do not observe a clear in-group bias trend towards plaintiffs; the in-group effects are null, consistent with the defendant starting in a defensive position and suspected of doing wrong, which triggers in-group bias. Quantitatively, our results indicate that defendants are 4 percentage points more likely to win if they share the judge's gender and 5 percentage points more likely to win if they share the judge's ethnicity. In the textual judgments, we explore the determinants of in-group bias. We find that potentially biased decisions are associated with shorter written judgements that are less likely to be cited. Additionally, we find evidence that judges that exhibit a slant against women in their writing are more likely to make biased decisions against women. We estimate that a one standard deviation change in the measure of gender slant is associated with a 2 percentage point decrease in win probability for female defendants. These findings suggest that written judgements can be used to predict in-group bias.

About Economic Development & Institutions

Institutions matter for growth and inclusive development. But despite increasing awareness of the importance of institutions on economic outcomes, there is little evidence on how positive institutional change can be achieved. The Economic Development and Institutions – EDI – research programme aims to fill this knowledge gap by working with some of the finest economic thinkers and social scientists across the globe.

The programme was launched in 2015 and will run until 2022. It is made up of four parallel research activities: path-finding papers, institutional diagnostic, coordinated randomised control trials, and case studies. The programme is funded with UK aid from the UK government. For more information seehttp://edi.opml.co.uk.

UKaid
from the British people

1 Introduction

Judges often exhibit bias in decision-making. One particular form of judicial bias documented in recent years is in-group bias, wherein judges are more likely to rule in favor of plaintiffs or defendants that share a certain identity with the judge (Shayo and Zussman 2011; Gazal-Ayal and Sulitzeanu-Kenan 2010; Knepper 2018; Sloan 2020). There are still many unknowns regarding the scope and determinants of judicial in-group bias and the potential role for implicit bias. Indeed, the phenomenon has been studied in relatively few contexts and rarely in the Global South.

Judicial bias in general, and in-group bias in particular, have far-reaching negative consequences. Decisions may be biased against groups that are marginalized, which can exacerbate existing inequalities. This is especially true for in-group bias since privileged groups may be more likely to represent a higher proportion of judges. Moreover, bias could undermine the effectiveness and inclusivity of courts, which are widely recognized as a key component of a well-functioning economy (Rodrik 2000; Visaria 2009; Ponticelli and Alencar 2016; World Bank 2017).

This paper aims to determine the extent and predictors of bias (especially in-group bias) along gender and ethnic lines in judicial decisions in the higher courts in Kenya.

Kenya provides an ideal setting for studying judicial bias for several reasons. First, political groups in Kenya are sharply divided along ethnic lines (Asingo et al. 2018), which may increase ethnic bias in society. Second, certain ethnic groups are underrepresented in the judiciary (see below), and there is a high degree of gender inequality across a number of dimensions, including representation in the judiciary and a variety of socioeconomic outcomes (IDLO 2020; UNDP 2020). If in-group bias is widespread, it may disproportionately harm these underrepresented groups. Third, there is an ongoing debate regarding the extent to which co-ethnic bias affects decision-making in the context of Africa generally and Kenya specifically (Berge et al. 2015).

We employ several data sources to examine the extent and determinants of judicial bias in Kenya. Our main data source is the Kenyan Judiciary's publicly available database for court cases, covering mostly Superior Court cases over the period 1976-2020.[1] By scraping the metadata associated with each case, we determined key variables, such as case type and names of judges and litigants. We also used additional data sources to determine the gender and ethnicity of participants. Furthermore, we used machine learning techniques to extract other key variables, such as the outcome of the case and the degree to which judges, by associating women with either negative or stereotypical traits, exhibit gender slant against women in their writing, which serve as textual proxies for implicit bias.

To determine the causal effect of an in-group relationship between judges and litigants, we rely on the random assignment of Kenyan judges to cases. Random assignment assures us that any relationship between in-group status and case outcomes is driven by bias rather than other factors, such as self-selection of judges to certain cases. To investigate the circumstances in which bias may be most prevalent, we examine whether judges that exhibit gender bias in their written judgements are more likely to display gender bias in the direction of their decisions.

Our main finding is that judges in Kenyan display both gender and ethnic in-group bias towards defendants. Our results suggest that defendants are about 4 percentage points more likely to win if they share the judge's gender and about 5 percentage points more likely to win if they share the judge's ethnicity. In contrast, we do not observe a clear in-group bias trend towards plaintiffs; the in-group effects are null.

We also find that slant against women in written judgements is in fact associated with lower win-rates for female defendants. Once again, however, we find no relationship with outcomes for plaintiffs. We estimate that a one standard deviation change in the measure of gender slant is associated with about a 2 percentage point decrease in win probability for female defendants. We also find that potentially biased judgements are associated with shorter written judgments (for gender and ethnic bias) that are less likely to be cited (for ethnic bias). These findings may suggest that biased decisions are more likely to be of a lower quality.

These findings have important implications for the Kenyan context. As mentioned, women and certain ethnic groups are underrepresented in the judiciary. As such, they are more likely to be negatively affected by in-group bias. In practical terms, keeping in mind the main case types in the dataset, this could imply a range of negative consequences. For civil cases, which often involve disputes over money (among other topics), in-group bias would imply a financial disadvantage for women and underrepresented ethnic groups. For

[1] See http://kenyalaw.org/caselaw/.

environment and land cases, bias may make these groups more likely to lose disputes over land ownership. Similarly, for succession cases, bias could lead to women being unfairly cut out of family inheritance or property.

This paper makes several important contributions. First, it builds on the scant literature related to judicial bias in developing countries. Judicial bias has been well studied in the United States (Knepper 2018; Depew, Eren, and Mocan 2017) and Israel (Shayo and Zussman 2011; Gazal-Ayal and Sulitzeanu-Kenan 2010), but in few other countries. Studying bias in the Kenyan context builds much needed evidence for the possibility of judicial bias outside the United States. As such, this paper helps expand our understanding of the scope of judicial bias in the Global South where these data are relatively scarce. To our knowledge, only one other paper has studied in-group bias in a developing country context. In their paper, Ash, Asher, et al. (2021) find no evidence in the Indian judiciary of gender or religious in-group bias in criminal case verdicts.

Second, to our knowledge, our paper is the first to examine judicial in-group bias towards both defendants and plaintiffs. Most previous studies on the topic have focused on criminal cases, for which the plaintiff is typically the state. By including civil, criminal, and other cases in our analysis, we are able to expand the scope. In doing so, we highlight that judges may be more likely to exhibit bias towards defendants than plaintiffs. One potential explanation for this heterogeneity rests in social identity theory, which states that "individuals define their own identities with regard to social groups and that such identifications work to protect and bolster self-identity" (Islam 2014). The theory predicts that when an individual perceives a "threat" to their in-group (and, by extenstion, their self-identity), they may be more likely to exhibit bias in favor of their group as a means to defend the group (and their identity).[2] To the extent that seeing one's in-group member as a defendant (i.e. as potentially guilty) constitutes a threat to group identity, greater in-group bias towards defendants than plaintiffs is in fact consistent with social identity theory. We are not, however, able to directly test this causal mechanism.

Third, our findings contribute to the broader literature on ethnic bias. Most importantly, they bring new insights to the issue of ethnic bias in sub-Saharan Africa generally and in Kenya specifically. There is a large literature studying the extent to which ethnicity affects decision-making and preferences in sub-Saharan Africa.[3] Considering the high level of ethnic fractionalization on the continent (and in Kenya), ethnic bias has major implications for a range of outcomes, including public service provision and community mobilization (Barkan and Chege 1989; Miguel and Gugerty 2005). The findings from this literature have been mixed. For example, Burgess et al. (2015) find that districts in Kenya that share the president's ethnicity receive twice as much funding on roads. On the other hand, Berge et al. (2015) show that most participants in lab experiments in Nairobi, Kenya exhibit no ethnic bias. Our paper adds to this literature by documenting a high-stakes form of ethnic bias, i.e. judicial bias, in Kenya. However, the findings also suggest that the level of bias is relatively mild, especially compared to the magnitude in the Israeli context, where ethnic in-group bias has a roughly 18 percentage-point impact on win probability (Shayo and Zussman 2011).

Fourth, our findings contribute to a large literature on gender bias in general–spanning labor markets (Azmat and Petrongolo 2014), education (Carlana 2019), and much more–by further demonstrating the extent and impact of gender discrimination. Relatedly, our study builds on research demonstrating the importance of female representation in public positions for both reducing bias (Beaman et al. 2009) and directly improving outcomes for women (Hessami and Fonseca 2020). It builds evidence for one specific channel (i.e. outcomes in court cases) through which female representation in the public sector directly affects women.

Fifth, the paper presents a novel application of machine learning tools for investigating the determinants of bias. Most notably, we show that text analysis can be used to predict bias in judge decision-making on litigants. The only previous study of judges' textual implicit bias examined how text analysis can predict bias in the judicial profession and common law precedent (Ash, Chen, and Ornaghi 2021) rather than predicting the impact on decisions involving a minority litigant. Likewise, we show that there is a level of consistency between bias in writing and judicial decisions. We are also the first to show that biased decisions appear to be associated with lower quality written judgements.

The rest of the paper is organized as follows. Section 2 presents background information on the judiciary,

[2] As evidence, Dietz-Uhler and Murrell (1998) found that when individuals read negative review about their group (i.e. when they were exposed to a threat), they were more likely to make positive affirmations about their own group. For additional evidence see Wann and Grieve (2005) and Voci (2010).

[3] For an overview, see Berge et al. (2015).

gender, and ethnicity in Kenya. Section 3 presents the data used in analysis. Section 4 outlines the empirical strategy. Section 5 presents the results. Finally, section 6 concludes.

2 Background

2.1 The Kenyan judiciary

The Kenyan judiciary is divided into two main court types: Superior and Subordinate Courts. The vast majority of our data covers the Superior Courts, which include High Courts, which hear both criminal and civil cases and appeals from Subordinate Courts; Environment and Land Courts; Employment and Labour Relations Courts; the Court of Appeal, which hears appeals from the High Courts, Environment and Land Courts, and Employment and Labour Relations Courts; and the Supreme Court, which hears appeals from the Court of Appeal and other high-level cases. (Kenyan Judiciary 2021).

According to our data (described below), the Court of Appeals almost exclusively hears civil cases; the Environment and Land Courts are largely split between civil cases and environment and land cases; the Employment and Labour Relations Courts are largely split between labor cases and civil cases; and the High Courts frequently hear a wide range of cases, including civil cases, land and environment cases, labor cases, criminal cases, and others. We have little data on Supreme Court cases, but it appears to hear mostly civil cases. Despite these general trends, the data appears to show that the courts are generally not restricted in the cases they hear, as they all tend to hear a wide range of case types. For most cases in most courts, there is only one judge. An exception is in Courts of Appeal, where the majority of cases are composed of multi-judge panels.

The Kenya judiciary does not employ a jury system. This means that judges alone are able to decide the outcomes of cases, which implies that bias among judges can have especially serious consequences.

In August 2010, the judicial system was overhauled by the implementation of a new constitution. The Constitution led to a wide range of reforms, including in the judiciary(Akech 2010). The judicial reforms were designed to reduce executive branch control over judicial outcomes, eliminate the system of bribing judges, increase transparency in judge selection, reduce the large backlog of cases, and increase female participation (Akech 2011; Gainer 2016). The reforms included the appointment of an ombudsperson to address corruption complaints; the creation of a meritocratic judge appointment process, separate from the oversight of the executive; the design of a standardized case management system; a doubling of the judicial budget; and the creation of the requirement (applied across elective and appointive bodies throughout the government) that no more than two-thirds of Kenyan judges be of the same gender. Although not all of these reforms have been fully enacted, progress has been made on many dimensions (Gainer 2015; Mutunga 2011; IDLO 2020). These reforms could have important implications for in-group bias and its effects. For one, some of the reforms, such as the reduction of corruption and increase in meritocratic assignment, could potentially reduce overt bias. Moreover, with more women in the judiciary, the aggregate effects of in-group bias for women would be less severe.

2.2 Gender and ethnicity in Kenya

As we show in the summary statistics section, below, there has been substantial progress towards gender parity in the judiciary–though there is still a long way to go to achieve equality. Inequalities in the judiciary are reflective of broader gender inequalities in Kenyan society. According to the 2020 United Nations Development Programme's Gender Inequality Index, which scores countries based on gender gapes related to representation in government, educational attainment, and labor force participation, Kenya ranks 126th out of 189 countries. Notably, women hold only 23 percent of seats in parliament (UNDP 2020).

Also widespead are the inequalities along ethnic lines, which provided an impetus for the adoption of a new constitution and its measures to devolve power (Akech 2010). As we show below, some groups are underrepresented in the judiciary relevant to their share of the population. Moreover, economic inequalities across regions (and likewise ethnicities) are highly salient (Friedrich-Ebert-Stiftung 2012). Political allegiance is also distributed by ethnicity, with political parties and coalitions created along clear ethnic lines (Asingo et al. 2018). According to the most recent census, Kenya has over 100 ethnic groups, and the largest group

(the Kikuyu) accounts for only about 17 percent of the population (KNBS 2019). In this context of diffuse ethnic groups and ethnic-based politics and resource distribution, ethnicity is a highly salient topic.

3 Data

3.1 Overview

The main data source used in our analysis is the Kenyan Judiciary's publicly available database for court cases.[4] The database includes 159,645 cases, almost exclusively from the Superior Courts, over the period of 1976 to 2020. Kenya Law, an organization within the Kenya Judiciary, began uploading case information in 2006. They upload all cases that are sent to them from the individual courts. Judicial officers in Superior Courts have a mandate to send cases to Kenya Law, so most cases from Superior Court cases are included, especially in more recent years when compliance has been greater. Cases from Subordinate Courts are sent on a much more ad-hoc basis. According to a representative from the organization, less than 1 percent of all cases online are removed via court orders by individuals that do not want their information online. Likewise, private information is removed from some cases. For cases prior to 2006, Kenya Law has made (and continues to make) efforts to gather and upload case information. Because older case information is less likely to be available, there is less information for earlier years. As such, the database is only roughly representative of Superior Court cases after 2006, and is less representative of cases before that date or of Subordinate Court cases.

In order to build our dataset for analysis from this database, we scraped the metadata and full text decision associated with each case. In doing so, we were able to directly extract the following for most cases: the names of plaintiffs, defendants, and judges; the type of case; the court in which the case was heard; and the year the judgement was delivered. We also used the history associated with each case to determine whether it was an appeal.

To determine gender and ethnicity and remove non-human cases (i.e. cases with companies or organizations as litigants), we used the name information scraped from the database. Cases without gender or ethnicity information for judges and either plaintiffs or defendants were dropped. The process for removing non-humans and determining gender and ethnicity (as well as the reasons for missing information) are discussed in appendix A. Once gender and ethnicity was assigned to each individual, we could determine the majority genders and plurality ethnicities for the judges, defendants, and plaintiffs for each case. By majority, we mean an absolute majority, where one gender comprises more than 50 percent of the total. By plurality, we mean a simple majority, where there is more of one ethnic group than any other. If no majority could be determined for gender, the majority gender was coded as missing. If no plurality could be determined for ethnicity, the plurality ethnicity was coded as "no plurality." This difference in coding was necessary because the main specification for gender in-group analysis requires binary outcomes, while the main specification for ethnicity in-group analysis does not (see specifications in section 4, below).

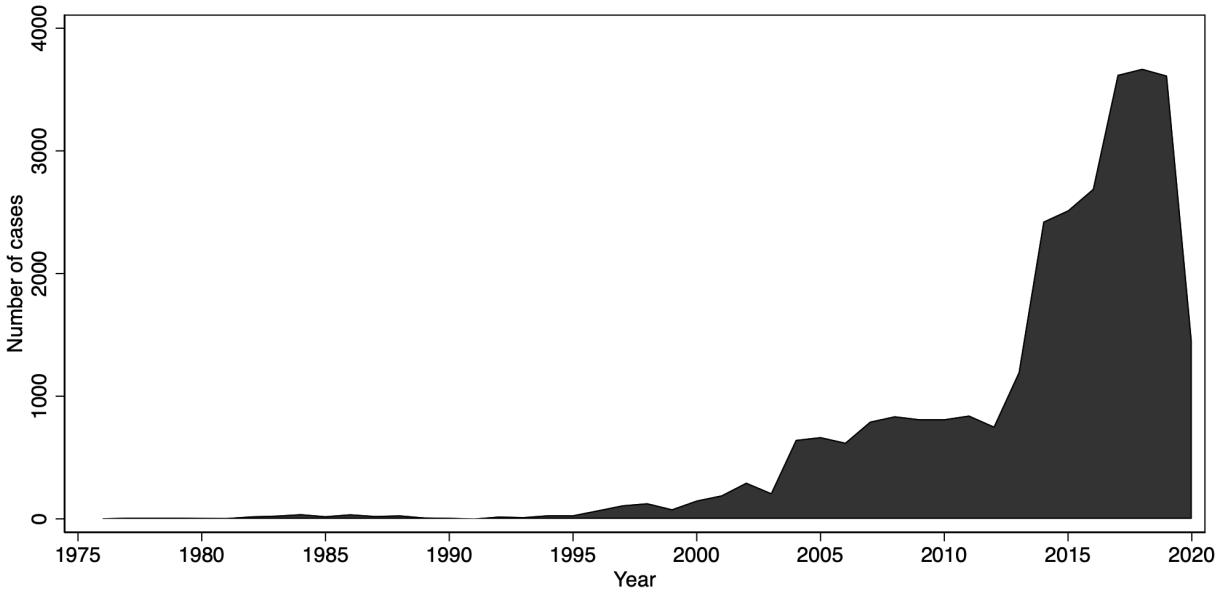
We used machine learning techniques to extract several other variables. To determine the winner of each case, we first scraped the case outcome information from the metadata. However, for 58,622 cases, the outcome was not stated. For these cases, we used a Binary Classification Machine Learning Model (described in appendix A) to analyze the text decisions of each case and determine the outcome. In the test set, the model was about 93 percent accurate. To measure the gender bias in judges' writing, we used a word embedding approach that captures the textual relationship between gendered language and either positive/negative language or career-oriented/family-oriented language. This approach allowed us to measure the extent to which judges disproportionately associate women with either negative or stereotypical qualities (i.e. a focus on family rather than career). The two variables resulting from this process are *Median slant, career vs family* and *Median slant, good vs bad*. For both measures, positive values indicate greater slant against women. A detailed description of this approach is provided in appendix A.

We also created variables measuring aspects of each written judgement that could signal quality of the judgement, including the number of cases cited in the text, the number of laws and acts cited in the text, the length of the text (measured as the number of words), and the number of times the judgement has been cited by other cases in our dataset. Appendix A describes variable construction in greater detail.

[4] See http://kenyalaw.org/caselaw/.

In total, the analysis dataset includes 29,571 cases, covering 94 courts and 352 judges. As figure 1 shows, the cases cover the years 1976 to 2020. Most of the cases in the dataset are from 2000 and after, with a sharp increase following 2012. Summary statistics of variables in the dataset are presented in appendix B. It shows that the main case types in our dataset are civil cases (46 percent), environment and land cases (32 percent), succession (9 percent), miscellaneous (8 percent), and labor relations (2 percent). All other cases comprise less than 1 percent of the total. Table B1 in appendix B shows the court types that are included in the dataset. It indicates that over 99 percent of cases are from Superior Court; in the "other" category, there are a small number of Subordinate Court cases. Most cases are from High Courts, followed by Environment and Land, Court of Appeal, and Employment and Labor. Very few Supreme Court cases are included. The analysis dataset has weak coverage of certain case types (most notably, criminal cases) not because the Kenya Law database does not include them, but because certain cases were far more likely to be dropped for reasons outlined above. For example, most of the 36,700 criminal cases that were included in the 159,645 cases in the database were dropped because they involved at least one non-human litigant.

Figure 1: Frequency of cases in the dataset over time



3.2 Summary statistics: gender

Figure C1 in appendix C shows that men comprise the majority of plaintiffs, the majority of defendants, and the majority of judges for most cases. As mentioned, these data are only roughly representative of cases in Kenya after 2006, and much less so before that date. But they still provide interesting insights into a large sample of cases. The gender gap is especially large for plaintiffs and defendants. Men comprise the majority of plaintiffs and the majority of defendants in about three times as many cases as women. In contrast, female judges comprise the majority in over half the number of cases as male judges.

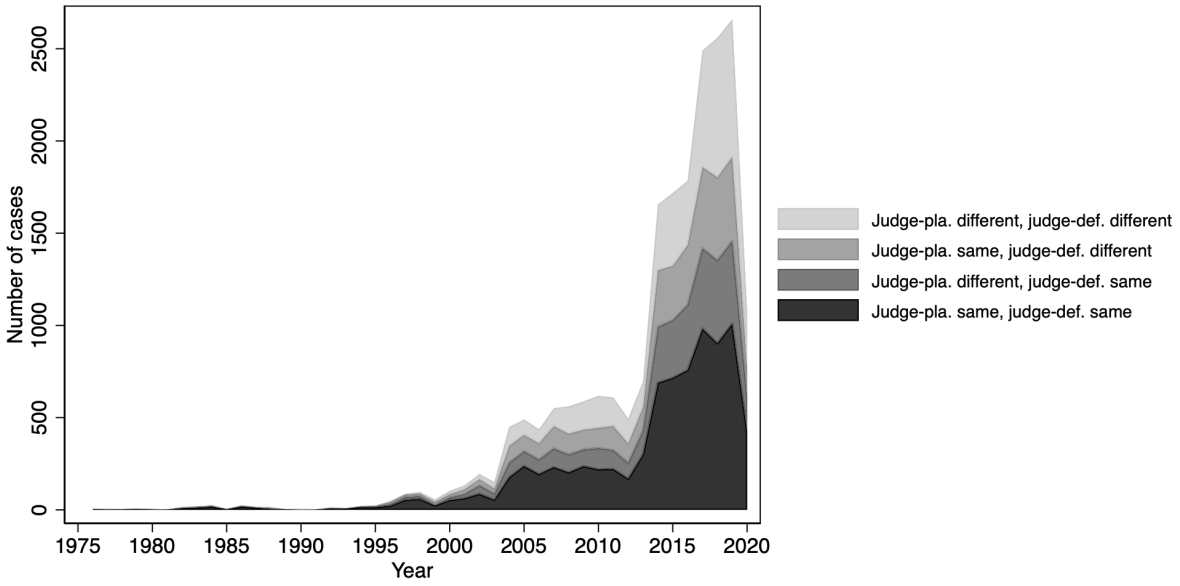
Figure C2 in appendix C shows how the gender gap has evolved over time. Since 1980, female representation has increased for all three roles (i.e. judge, plaintiff, and defendant). The increase has been most dramatic for judges, with sharp increases beginning around 2000. The increases continued after 2010, the year during which the new constitution was adopted.

Figure C3 in appendix C illustrates the gender gaps by case type and role. It shows that women are especially underrepresented in criminal cases as defendants and plaintiffs. In contrast, women are approaching parity as defendants and plaintiffs in family and succession cases. Criminal cases are more or less evenly split between male and female judges, but all other case types have a greater proportion of male judges.

Figure 2 displays the gender similarities and differences between judges and plaintiffs/defendants over time. It shows that the most common combination is for judges, plaintiffs, and defendants to all have

the same majority gender–especially in earlier years. This is due to the fact that all three positions are dominated by men. The second most common combination is for judges to have a different majority gender than both plaintiffs and defendants. The other two combinations, where the judge has the same majority gender as either the defendant or plaintiff but not both of them, are more os less equal. Overall, despite these differences, cases are relatively evenly split across the four combinations.

Figure 2: Case frequency over time by similarities/differences in majority gender across judges, plaintiffs, and defendants



pla. = plaintiff, def. = defendant.

3.3 Summary statistics: ethnicity

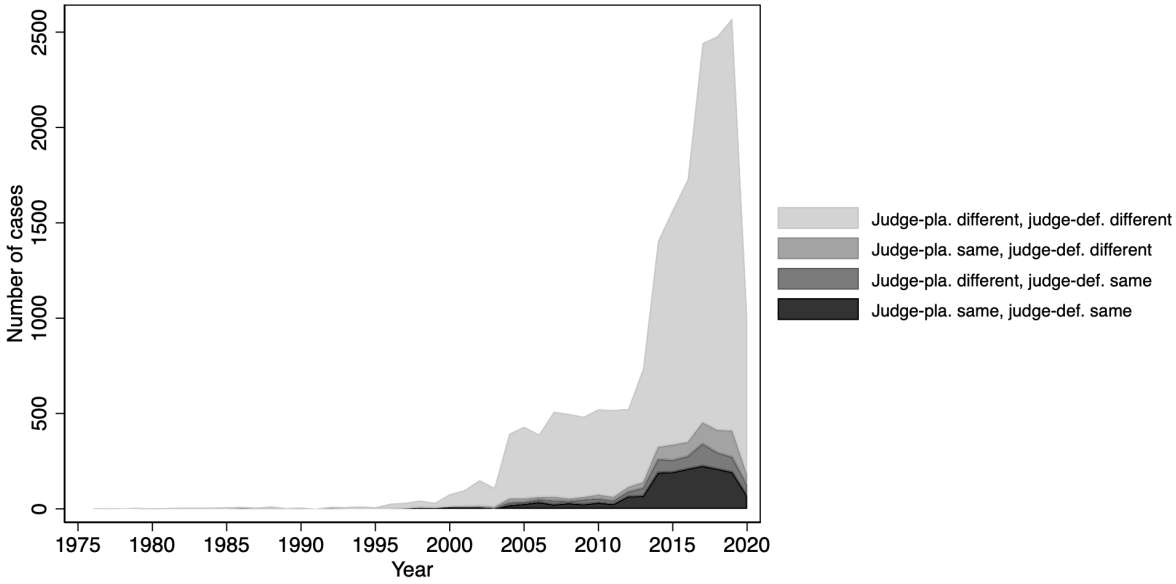
Figure C4 in the appendix depicts the proportion of cases represented by the different ethnic pluralities. It also depicts each ethnic group's proportion of the total Kenyan population, as a benchmark for equal representation. Again, these data are not entirely representative, but they at least provide insights into potential ethnic disparities in the judiciary. Several trends stand out. First, even accounting for the fact that Kikuyu is the largest ethnic group in Kenya, the group still has outsized representation in the dataset as judges, plaintiffs, and defendants. In contrast, the Turkana and Somali have notably low representation given the size of the total population of these groups. Furthermore, the Luo are overrepresented as judges, and the Meru are overrepresented as plaintiffs and defendants. Given this variation, it is clear that in-group bias would have differential effects across ethnic groups.

Table C2 in appendix C helps to ground the findings in figure C4 within the context of ethnic politics in Kenya. The table uses data from the Ethnic Politics Relations (EPR) dataset, which classifies each ethnic group based on their political power, across various periods of time. The classifications are senior partner, which indicates that representatives from the ethnic group participate as senior partners in a power-sharing agreement for control of the executive branch of government; junior partner, which indicates that representatives from the ethnic group participate as junior partners in a power-sharing agreement for control of the executive branch of government; powerless, which indicates that "representatives hold no political power at either the national or the regional level without being explicitly discriminated against;" and discriminated, which indicates that "group members are subjected to active, intentional, and targeted discrimination, with the intent of excluding them from both regional and national power" (Cederman, Wimmer, and Min 2010). The table reveals the complexity of ethnic politics in Kenya, wherein almost all of the major ethnic groups have occupied executive power as either junior or senior partners. Only the Somali group has remained outside of power as a discriminated group. Given this complexity, it is difficult to draw links between

political power and representation in the judiciary, but it is notable that the Somali group is substantially underrepresented.

Figure 3 displays the ethnic similarities and differences between judges and plaintiffs/defendants over time. By far the most common combination is for judge plurality ethnicity to be different than plurality ethnicity for plaintiffs and defendants, which is not surprisingly given the number of ethnic categories. The next most common combination is for the judge plurality ethnicity to be the same as both the defendants' and the plaintiffs' plurality ethnicity. The other two possible combinations are more or less evenly split.

Figure 3: Case frequency over time by similarities/differences in plurality ethnicity across judges, plaintiffs, and defendants



pla. = plaintiff, def. = defendant.

4 Empirical strategy

The main goal of our analysis is to determine the extent to which judges exhibit bias towards plaintiffs and defendants of the same gender and ethnicity and to determine the predictors of this bias. To isolate the causal effect of in-group identity, we rely on the as-good-as-random assignment of judges to cases, described in the following subsection and further justified by balance tests.

Random assignment is key to our empirical strategy because it assuages the concern that judge ethnicity or gender is correlated with case characteristics that affect outcomes. For example, if judges of a certain ethnic group preferred to rule on cases for which their ethnic group was less likely to be guilty, then we would expect to see indications of in-group bias, but the effect would be driven by selection bias rather than in-group bias. In addition, judges of a certain ethnicity may be more likely to rule on cases in areas of the country where crime is more or less severe. If these distributions of crime severity are correlated with the ethnic distributions of defendants and plaintiffs, then we may again falsely perceive in-group bias.

One threat to this strategy is the possibility that co-gender or co-ethnicity influences the litigants' behavior, which in turn affects the judge's ruling. For example, co-ethnicity may create greater confidence among defendants, causing them to display "better" behavior. Likewise, the judge may be more likely to say certain things to the litigants that affect their behavior. If the judges looks more favorably on this behavior, they may be more likely to rule in favor of the defendant. In this case, the judge may be displaying a certain type of "attitude" bias rather than in-group bias. Regardless, in this case, the outcome would still be altered by the in-group relationship, albeit through a mechanism other than direct in-group bias.

The rest of this section proceeds as follows. Subsection 4.1 presents the qualitative justification that

judge assignment is as good as random (at least in the most recent years). Subsections 4.2 and 4.3 present quantitative evidence through balance tests that the random assignment assumption does in fact hold across the full time period for both gender and ethnicity, respectively. The econometric approaches for examining gender and ethnic in-group bias are described in detail in subsections 4.4 and 4.5, respectively. We also aim to examine whether judge gender slant in written opinions is correlated with outcomes for female defendants and plaintiffs. The approaches to these analyses are described in subsection 4.6. Finally, subsection 4.7 presents the specifications for examining the relationship between in-group bias and textual variables other than slant, for both gender and ethnicity.

4.1 Random assignment of cases to judges

Random assignment of cases to judges is critical to our identification strategy. At least as of 2020, the World Bank Doing Business Index asserts that cases are in fact randomly assigned to judges (World Bank 2021). However, it must be noted that this randomized procedure may be a relatively new phenomenon. Indeed, introducing randomization was allegedly one of the goals of the reform team following the implementation of the 2010 Kenyan Constitution (Gainer 2015). Therefore, it is necessary to provide further evidence that case selection by judges has not been a common feature across our sample. To do so, we present balance tests in the following two subsections. Considering the possibility that assignment has become random after 2010, in addition to conducting the tests for the full sample, we also split the balance tests into before 2011 and since 2011.

4.2 Gender balance tests

To confirm that judge assignment to cases is random in terms of gender majority, we use the following balance test for the analysis sample:

$$\begin{aligned} judge\_maj\_female_{i,c,t} = \beta_1 def\_maj\_female_{i,c,t} + \\ \beta_2 pla\_maj\_female_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \end{aligned} \tag{1}$$

where $\Phi_{c,t}$ and $X_{i,c,t}$ represent the same fixed effects and controls as in the main gender specification (discussed below).

The results of the balance test are shown in table D1 in appendix D. Column (1) does not include additional controls. Column (2) includes controls for ethnicity, and column (3) adds additional controls (as listed in the table notes). The results indicate that male- and female-majority defendant groups are equally likely to be assigned male- and female-majority judge panels (including single-judge panels). Likewise, male- and female-majority plaintiff groups are equally likely to be assigned male- and female-majority judge panels. In light of the concern that cases have only become randomized after the creation of the new constitution and the accompanying judicial reforms, tables D2 and D3 present balance tests for pre-2011 and since 2011, respectively. The results are consistent with table 4.

4.3 Ethnicity balance tests

To confirm that judge assignment to cases is random in terms of ethnic majority, we use variations of the following balance test:

$$\begin{aligned} judge\_plur\_kikuyu_{i,c,t} = \beta_1 def\_plur\_kikuyu_{i,c,t} + \\ \beta_2 pla\_plur\_kikuyu_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \end{aligned} \tag{2}$$

where $\Phi_{c,t}$ and $X_{i,c,t}$ represent the same fixed effects and controls as in the main gender specification, $judge\_plur\_kikuyu_{i,c,t}$ is a binary variable indicating whether the judge plurality is the Kikuyu ethnic group, $def\_plur\_kikuyu_{i,c,t}$ is a binary variable indicating whether the defendant plurality is the Kikuyu ethnic group, and $pla\_plur\_kikuyu_{i,c,t}$ is a binary variable indicating whether the plaintiff plurality is the Kikuyu ethnic group. We run series of 12 tests, with each test using binary variables for different ethnicities.

Tables D4 through D7 in appendix D report the results of the tests. They show that defendants and plaintiffs across all ethnicities are not more likely to be assigned judges from their ethnic group. One exception is Luhya defendants, as table D5 shows. Balance tests for both pre-2011 and since 2011 are also

presented in appendix D (see tables D8 through D15). They show that there are significant coefficients for Luhya defendants in the 2011-2020 period and Kamba for the 1976-2010 period. However, in all cases, the coefficients are too small to raise serious concern about bias in case assignment. In order to nonetheless guard against this possibility, we conduct a robustness check of the main analysis that drops all Luhya and Kamba individuals. Appendix E presents these results. A comparison between these results and the main results below show that the in-group bias we observe is not driven by any possible bias in Luhya or Kamba case assignment.

Tables D4 through D15 include a full set of controls. To save space, we have not included the results without controls. However, they are qualitatively similar, with no additional significant coefficients for the variables of interest.

4.4 Main gender specifications

To estimate judicial gender bias, we use an empirical strategy that follows Shayo and Zussman (2011) and Ash, Asher, et al. (2021). We model outcome $Y_{i,c,t}$ (where Y=1 corresponds to the defendant winning the case) for case i filed in court c at time t as:

$$\begin{gathered} Y_{i,c,t} = \alpha + \beta_1 judge\_maj\_female_{i,c,t} + \beta_2 def\_maj\_female_{i,c,t} + \\ \beta_3 judge\_maj\_female_{i,c,t} * def\_maj\_female_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \end{gathered} \tag{3}$$

where $judge\_maj\_female$ and def_maj_female are binary variables indicating whether judge panels and defendant groups, respectively, are majority female. The main outcome of interest is the interaction term, which indicates in-group bias. $\Phi_{c,t}$ is a court-year fixed effect and $X_{i,c,t}$ is a vector of additional control variables, which may include: binary variables for judge, defendant, and plaintiff plurality ethnicity; variables for the numbers of judges, plaintiffs, and defendants; a binary variable indicating whether the case is an appeal; and binary variables indicating the case type. Court-year fixed effects are used to ensure that we are comparing defendants and plaintiffs that are in the same court at the same time. Court-year periods without sufficient variation are dropped from the regressions.

The specification used to test for in-group bias towards plaintiffs is identical to (1), except a binary variable for plaintiff majority gender, pla_maj_female substitutes def_maj_female. An alternate specification includes both variables, as such:

$$\begin{gathered} Y_{i,c,t} = \alpha + \beta_1 judge\_maj\_female_{i,c,t} + \beta_2 def\_maj\_female_{i,c,t} + \\ \beta_3 pla\_maj\_female_{i,c,t} + \beta_4 judge\_maj\_female_{i,c,t} * def\_maj\_female_{i,c,t} \\ \beta_5 judge\_maj\_female_{i,c,t} * pla\_maj\_female_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \end{gathered} \tag{4}$$

4.5 Gender slant analysis

To examine the conditions under which gender bias can be expected, we examine whether judges' slant against women in opinions predicts bias against female defendants and plaintiffs. For this analysis, we use a specification that examines bias against women in general, rather than in-group bias. To do so, we build on equation (3) by adding one of the two measures of slant, described above, and an interaction between the slant measure and def_maj_female and/or pla_maj_female. The main outcomes of interest are the interactions with slant, which indicate whether female defendants and plaintiffs are less likely to win the case if the judge exhibits slant in her/his writing.

One possible critique of this approach is that that gender slant may be correlated with decisions against women because gender slant tends to appear in cases where female defendants are in fact more worthy of losing the case. However, we do not expect this to be a problem since the measure of slant is by judge rather than individual case, and we would not expect on average that slanted judges are more likely to have cases where female defendants are more worthy of losing.

4.6 Main ethnicity specifications

For the ethnicity in-group bias analysis, we use a slightly different econometric specification in order to account for the fact that there are many more categories of ethnicity. To estimate judicial ethnic bias, we

model outcome $Y_{i,c,t}$ (where Y=1 corresponds to the defendant winning the case) for case i filed in court c at time t as:

$$Y_{i,c,t} = \alpha + \beta_1 judge\_pla\_same_{i,c,t} + \beta_2 judge\_def\_same_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \tag{5}$$

where $judge\_pla\_same_{i,c,t}$ is a binary variable indicating whether the judge ethnic plurality is the same as the plaintiff ethnic plurality, and $judge\_def\_same_{i,c,t}$ is a binary variable indicating whether the judge ethnic plurality is the same as the defendant ethnic plurality. $\Phi_{c,t}$ is a court-year fixed effect and $X_{i,c,t}$ is a vector of additional control variables, which may include: binary variables for judge, defendant, and plaintiff plurality ethnicity; binary variables for judge, defendant, and plaintiff majority genders; variables for the numbers of judges, plaintiffs, and defendants; a binary variable indicating whether the case is an appeal; and binary variables indicating the case type.

4.7 Judgement text specifications

To study the relationship between in-group bias and characteristics of the judgement text for a given case, we start with variations on the following specification:

$$Y_{i,c,t} = \alpha + Y_{i,c,t} = \alpha + \beta_1 judge\_maj\_female_{i,c,t} + \beta_2 def\_maj\_female_{i,c,t} + \beta_3 judge\_maj\_female_{i,c,t} * def\_maj\_female_{i,c,t} + \beta_4 judge\_def\_same_{i,c,t} + \Phi_{c,t} + X_{i,c,t,} + \epsilon_{i,c,t} \tag{6}$$

where $Y_{i,c,t}$ represents one of the four judgement text variables (number of cases cited, number of laws/acts cited outcome, number of times the case has been cited, or length) for case i filed in court c at time *t*. $judge\_def\_same_{i,c,t}$ is a binary variable indicating whether the judge ethnic plurality is the same as the defendant ethnic plurality. $\Phi_{c,t}$ is a court-year fixed effect and $X_{i,c,t}$ is a vector of additional control variables.

In addition to running this specification on the full sample, we split the sample into two groups–cases where the defendant won and cases where the defendant lost–and run an additional series of regression for each. If in-group bias is associated with different characteristics for judgement texts, then we should see significant coefficients on $judge\_def\_same_{i,c,t}$ and/or $\beta_3 judge\_maj\_female_{i,c,t} * def\_maj\_female_{i,c,t}$ for the defendant-win sample but not the defendant-lose sample, and the coefficients in the defendant-win sample should be larger than in the full sample. For example, if erthnically biased judgements are associated with the case being cited fewer times, then we should expect to see a negative coefficient on $judge\_def\_same_{i,c,t}$ in the defendant-win sample, a (potentially null) negative coefficient of lesser magnitude in the full sample, and a null coefficient in the defendant-lose sample.

5 Results

5.1 Main gender results

Figure 4 displays defendant win proportions by judge and defendant majority gender. With win proportions higher for female majority defendants among female judge panels and win proportions higher for male majority defendants among male judge panels, the figure is suggestive of in-group bias. However, the differences are not statistically significant at $p<0.05$.

In contrast, figure 5, which displays plaintiff win proportions by judge and defendant majority ethnicity, is not suggestive of in-group bias. Rather, across both judge genders, defendants are more likely to win if plaintiffs are male–though not significantly so for female judges. We also see in this figure that female judges rule in favor of defendants less often.

These trends are consistent with the main gender regression results in table 1. The significantly positive coefficients on the interaction between judge and defendant majority gender provide evidence that there is in-group gender bias from judges towards defendants. This finding is robust to various specifications. The significant results suggest that, all else equal, defendants are between 3.7 and 4.0 percentage points more likely to win if they have the same majority gender as the judges.

The results do not provide evidence of in-group bias towards plaintiffs. The coefficients on the interaction between judge and plaintiff gender are in the direction indicative of *out-group* bias, but they are insignificant and the point estimates are small. Beyond the effects of in-group bias, we see that female judges are in general more likely to rule in favor of defendants, and male plaintiffs are in general more likely to lose.

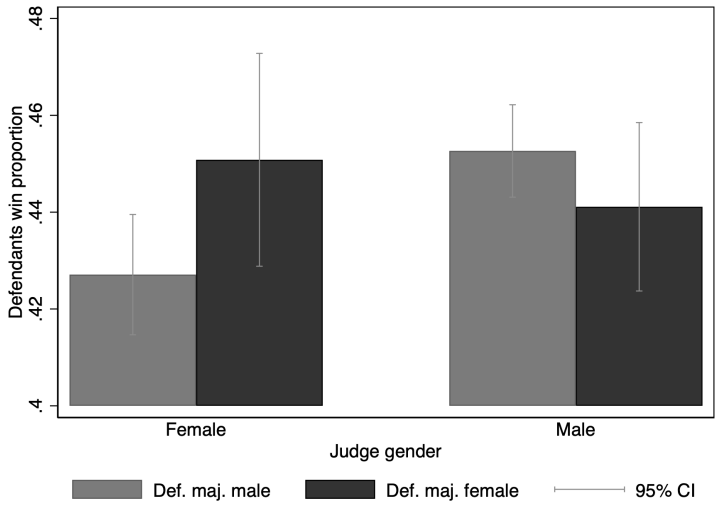
Figure 6 visualizes the in-group bias trend for defendants. Based on a series of regressions, one for each individual judge, it plots the predicted win proportion when defendants have the same majority gender as each judge in relation to the predicted win proportion when defendants have a different gender than each judge. Each bubble in the graph represents a specific judge. Bubbles above the 45-degree line indicate that the judge has in-group bias. The darker the bubble is, the more significant the relationship is. The larger the bubble, the more observations there are. Finally, the plus sign represents the predicted win proportions from a regression that includes all of the judges depicted in the graph. Since it is above the line, it shows that there is, on average, in-group bias towards defendants among the judges. As depicted by the plus sign, the predicted win proportion when judges have the same gender as defendants is 0.454. When they have a different gender, it is 0.430, 0.024 less. These results are similar to the results from table 1.

The figure introduces important nuance to the results. It is clear from the scatterplot that not all judges exhibit in-group gender bias and that many in fact have out-group bias. And although some judges display extreme bias, most are clustered around the line, suggesting mild or no bias for most.

Figure 7, which visualizes the individual judge bias coefficients as a distribution, reinforces these findings. It makes clear that most judges do not exhibit bias or exhibit mild bias. And although there are some more extreme judges, in the direction of both in- and out-group bias, the results seem to be largely driven by a clustering of mildly in-group biased judges.[5]

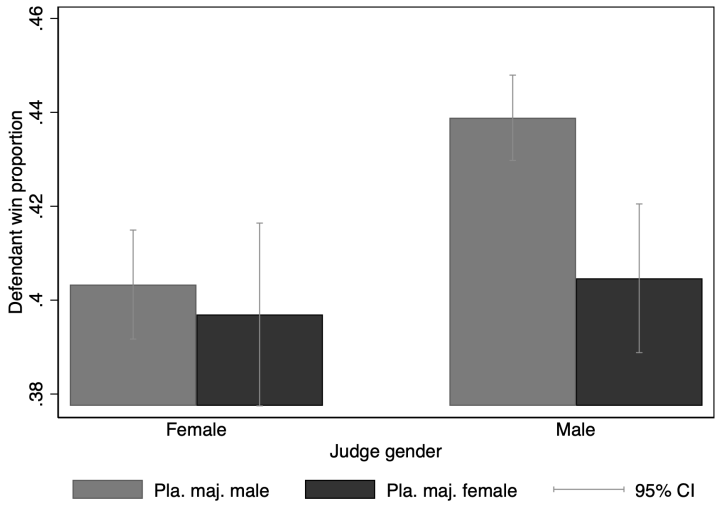
[5]Table F1 in appendix F explores the effects of putting biased judges on panels. It analyzes in-group bias among the 14 judges with significant coefficients for gender in-group bias towrds defendants in individual judge regressions. It shows that when these judges make decisions individually, the defendant is 40 percentage points more likely to win if they share the judge's gender. In contrast, when these judges rule on panels with other judges, the defendant is only 40 percentage points more likely to win if they share the judge's gender. It is important to note that thes results are not causal. Nonetheless, they suggest one potential means for policymakers to reduce bias: put biased judges on panels. There were insufficient observations to conduct this analysis for ethnicity; few of the ethnicity in-group biased judges had ruled on panels.

Figure 4: Defendant win proportion by judge and defendant majority gender



def. = defendant, maj. = majority.

Figure 5: Defendant win proportion by judge and plaintiff majority gender



pla. = plaintiff, maj. = majority.

Table 1: Gender main results

	(1) Def. win	(2) Def. win	(3) Def. win	(4) Def. win	(5) Def. win
Judge maj. female	-0.0406***	-0.0397**	-0.0483***	-0.0426***	-0.0427***
	(0.0121)	(0.0195)	(0.0133)	(0.0132)	(0.0132)
Pla. maj. female		-0.0308**	-0.0533***	-0.0450***	-0.0458***
		(0.0127)	(0.0110)	(0.0110)	(0.0110)
Def. maj. female	-0.00553		0.00112	0.00946	0.00893
	(0.0105)		(0.0108)	(0.0108)	(0.0108)
Judge maj. fem. X pla. maj. fem.		0.0173	0.00789	0.00728	0.00806
		(0.0207)	(0.0172)	(0.0169)	(0.0168)
Judge maj. fem. X def. maj. fem.	0.0369**		0.0404**	0.0384**	0.0386**
	(0.0166)		(0.0168)	(0.0168)	(0.0168)
Court-year FE	Yes	Yes	Yes	Yes	Yes
Ethnicity dummies	No	No	No	No	Yes
Other controls	No	No	No	Yes	Yes
Observations	22889	25753	20437	20437	20437

The regressions test whether defendants (plaintiffs) are more likely to win (lose) if they have the same (a different) majority gender as judges. The coefficients of interest are on the interaction terms.

Standard errors, in parentheses, are clustered at the judge level.

All columns are based on a linear regression model. For specification details, see equations 3 and 4.

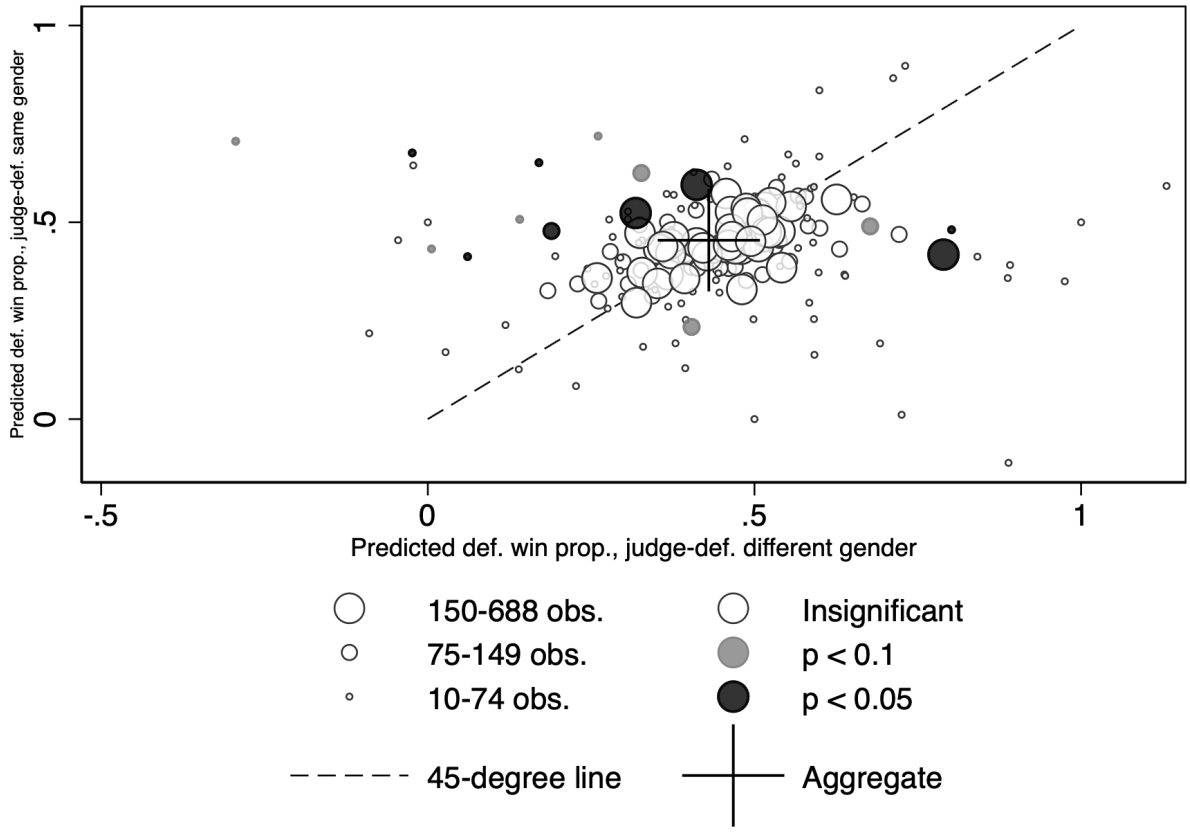
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges.

Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.

To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.

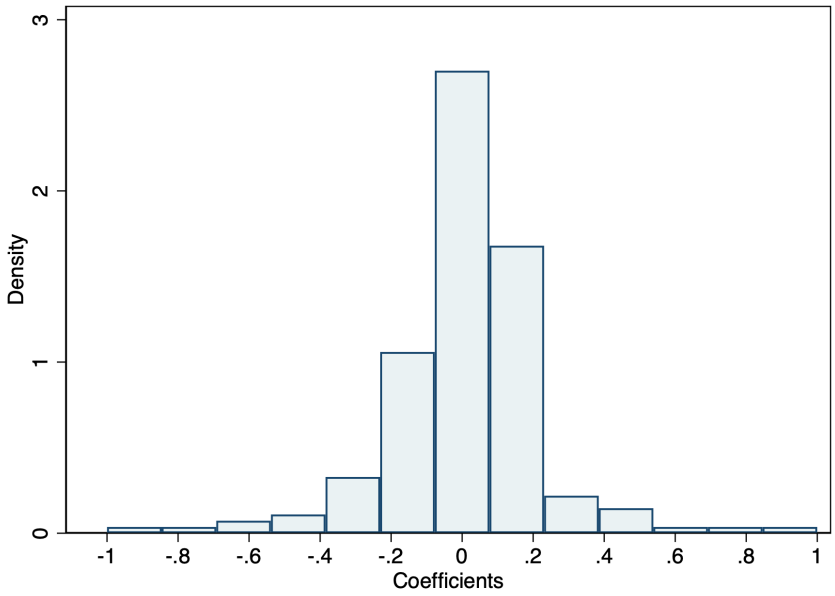
Pla. = plaintiff, def. = defendant, maj. = majority.

Figure 6: Predicted defendant win proportion, by judge and by defendant similarity with judge gender



def. = defendant, prop. = proportion. Each bubble indicates a specific judge. Only single judges are included, not judge panels. Judges without sufficient variation in outcomes were dropped. In total, 187 judges are included. The aggregate regression includes all single-judge panel observations, a total of 21,359. The outcome is significant at $p < 0.01$. Predictions are based on a regression with court-year fixed effects.

Figure 7: Distribution of coefficients estimating individual judges' in-group gender bias towards defendants



Coefficients are based on a regression with court-year fixed effects. Judges without sufficient variation in outcomes were dropped. In total, 187 judges are included.

5.2 Gender slant analysis results

Tables 2 and 3 present the results of the slant analysis. Using the career vs. family measure of slant, table 11 provides evidence of a correlation between biased writing and negative outcomes for women. It suggests that, for a 0.1 increase in slant against women (equivalent to about one standard deviation of the career vs. family measure), female defendants are about 1.5 percentage points less likely to win. The results hold across various specifications.

The results with the good vs. bad measure in table 3 are similar. They show that, for a 0.05 increase in slant against women (equivalent to about one standard deviation of the good vs. bad measure), female defendants are about 1.5 to 1.8 percentage points less likely to win. For both measures, the coefficients on the interactions for plaintiffs is in the expected direction but not significant.

Figure 8 presents the predicted win proportions for male and female defendants and various levels of judge slant, for the career vs. family measure. These predictions are based on table 2, column (3). The figure shows that male defendants are more likely to win–and female defendants are less likely to win–if judges are more slanted against women in their writing. Figure 9 presents the predicted win proportions for male and female defendants and various levels of judge slant, for the good vs. bad measure. These predictions are based on table 3, column (3). In this case, the figure shows that male defendants are essentially unaffected by a judge's slant. However, female defendants are still less likely to win if judges are more slanted against women in their writing.[6]

[6]In Appendix G, we present several other results related to textual gender slant. Tables G1 and G2 analyze the relationship between judge slant and appeals and figures G1 and G2 visualize the relatoinship. For the family vs career measure of slant, the relationship is null or even negative for some sepcifications. For the good vs bad measure, the realtionship is positive but weak and not significant for more specifications. These mostly null results are constrasted with the findings for the relationship between slant and and reversals, presented in tables G3 and G4 and figures G3 and G4. Although the results for the family vs career measure of slant are again mixed, the results for the good vs bad measure are more consistently positive. These findings suggest that 1) judge slant (according to the good vs bad measure) may be associated with lower quality judgements prine to reversals and 2) since the appeals results are null, litigants and attorneys may not able to recognize gender bias in decisions and/or are not aware that they are more likely to have decisiosn reversed if they appeal. One other noteworthy finding related to slant is that for the good vs bad measure of slant, female judges are much more likely to be slanted ($p<0.01$). For the family vs career measure, male judges are much more likely to be slanted ($p<0.01$).

Table 2: Gender results with text slant, career vs family measure

	(1)	(2)	(3)	(4)
	Def. win	Def. win	Def. win	Def. win
Judge maj. female	-0.0497***	-0.0495***	-0.0492***	-0.0449***
	(0.0138)	(0.0137)	(0.0138)	(0.0138)
Pla. maj. female	-0.0508***	-0.0489***	-0.0504***	-0.0403***
	(0.0118)	(0.0119)	(0.0118)	(0.0117)
Def. maj. female	-0.00368	-0.00649	-0.00632	0.000590
	(0.0116)	(0.0117)	(0.0117)	(0.0117)
Judge maj. fem. X pla. maj. fem.	0.00269	0.00311	0.00261	0.00180
	(0.0181)	(0.0182)	(0.0182)	(0.0178)
Judge maj. fem. X def. maj. fem.	0.0464***	0.0445***	0.0445***	0.0435**
	(0.0173)	(0.0168)	(0.0168)	(0.0169)
Slant against women, career vs family	0.00890	0.0284	0.0430	-0.00913
	(0.0824)	(0.0815)	(0.0828)	(0.0824)
Pla. maj. fem. X Slant against women	-0.0665		-0.0558	-0.0304
	(0.0853)		(0.0867)	(0.0880)
Def. maj. fem. X Slant against women		-0.154*	-0.149*	-0.141*
		(0.0852)	(0.0857)	(0.0838)
Court-year FE	Yes	Yes	Yes	Yes
Ethnicity dummies	No	No	No	Yes
Other controls	No	No	No	Yes
Observations	18236	18236	18236	18236

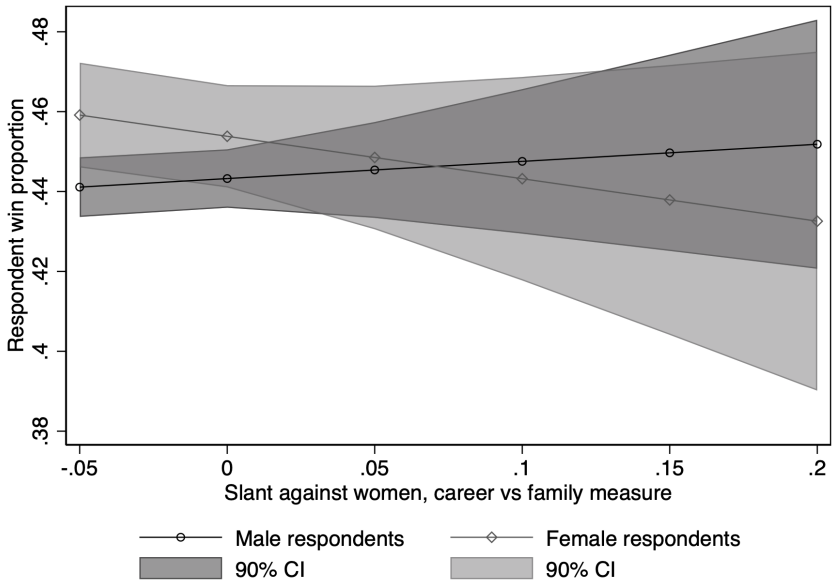
The regressions test whether defendants/plaintiffs are more likely to lose if they are female and the judge is slanted against females in their writing.
The coefficients of interest are on the interaction terms in the last two rows.
The measure of slant against women is based on the judges' stereotypical association of women with family-based qualities rather than career-based qualities.
All columns are based on a linear regression model. For specification details, see equation 3.
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiff, def. = defendant, maj. = majority.

Table 3: Gender results with text slant, good vs bad measure

	(1)	(2)	(3)	(4)
	Def. win	Def. win	Def. win	Def. win
Judge maj. female	-0.0451***	-0.0454***	-0.0462***	-0.0420***
	(0.0156)	(0.0154)	(0.0154)	(0.0157)
Pla. maj. female	-0.0332**	-0.0497***	-0.0341**	-0.0264
	(0.0166)	(0.0129)	(0.0167)	(0.0162)
Def. maj. female	-0.00561	0.0165	0.0158	0.0184
	(0.0116)	(0.0145)	(0.0145)	(0.0142)
Judge maj. fem. X pla. maj. fem.	0.00528	0.00202	0.00508	0.00267
	(0.0204)	(0.0206)	(0.0205)	(0.0201)
Judge maj. fem. X def. maj. fem.	0.0470**	0.0511***	0.0510***	0.0506***
	(0.0184)	(0.0180)	(0.0180)	(0.0182)
Slant against women, good vs bad	-0.0970	-0.0822	-0.0111	-0.0389
	(0.153)	(0.146)	(0.153)	(0.154)
Pla. maj. fem. X Slant against women	-0.286		-0.266	-0.226
	(0.182)		(0.184)	(0.180)
Def. maj. fem. X Slant against women		-0.374**	-0.358**	-0.306*
		(0.172)	(0.174)	(0.172)
Court-year FE	Yes	Yes	Yes	Yes
Ethnicity dummies	No	No	No	Yes
Other controls	No	No	No	Yes
Observations	15206	15206	15206	15206

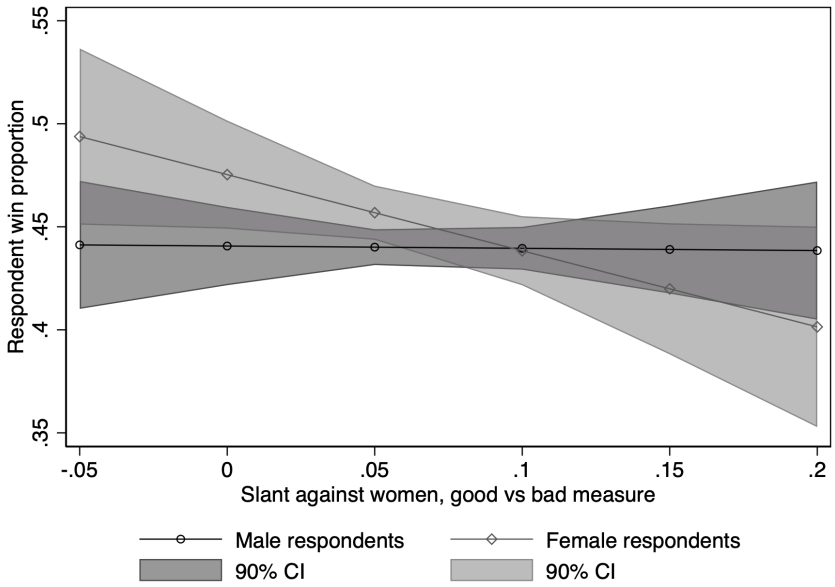
The regressions test whether defendants/plaintiffs are more likely to lose if they are female and the judge is slanted against females in their writing.
The coefficients of interest are on the interaction terms in the last two rows.
The measure of slant against women is based on the judges' association of women with negative qualities.
All columns are based on a linear regression model. For specification details, see equation 3.
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendant, plaintiffs, and judges.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiff, def. = defendant, maj. = majority.

Figure 8: Predicted defendant win proportions at various levels of judge slant (career vs family measure), by defendant gender



Based on table 2, column (3).

Figure 9: Predicted defendant win proportions at various levels of judge slant (good vs bad measure), by defendant gender



Based on table 3, column (3).

5.3 Main ethnicity results

Figure 10 displays defendant win proportions across various in-group categories relating judges, defendants, and plaintiffs. Figure 10a displays outcomes when the judge and plaintiff have the same ethnicity. Figure 10b displays outcomes when the judge and plaintiff have different ethnicities. They both show that defendants are more likely to win when judges and defendants are the same the ethnicity. The differences are not significant at $p<0.05$.

The regression results in table 4 corroborate the idea that there is in-group bias among judges towards defendants. They show that defendants are between 4.3 and 5.7 percentage points more likely to win if they share an ethnicity with the judge. The finding is robust to all of the specifications presented. As with gender, in-group bias is not observed for plaintiffs; the coefficients are both positive and negative across specifications and are not significant.

Figure 11 visualizes the in-group bias trend for defendants. Based on a series of regressions, one for each individual judge, it plots the predicted win proportion when defendants have the same majority ethnicity as each judge in relation to the predicted win proportion when defendants have a different ethnicity than each judge. Each bubble in the graph represents a specific judge. Bubbles above the 45-degree line indicate that the judge has in-group bias. The darker the bubble is, the more significant the relationship is. The larger the bubble, the more observations there are. Finally, the plus sign represents the predicted win proportions from a regression that includes all of the judges depicted in the graph. Since it is above the line, it shows that there is, on average, in-group bias towards defendants among the judges. As depicted by the plus sign, the predicted win proportion when judges have the same ethnicity as defendants is 0.486. When they have a different ethnicity, it is 0.442, 0.044 less. These results are similar to the results from table 4.

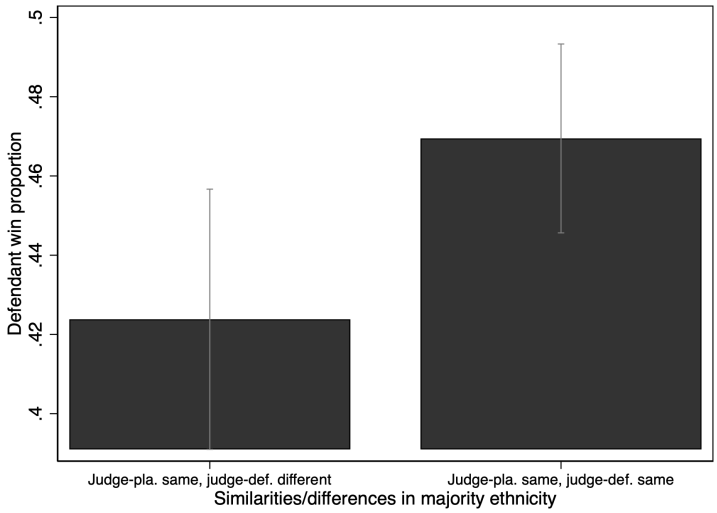
As with the gender results, these results show that individual judges exhibit both in-group and out-group bias. However, for ethnicity, out-group bias is not significant for any judges. Several judges display more extreme bias but, as with gender, most judges cluster around the line indicating unbiased judgements. However, figure 12 makes clear that there is less clustering of bias coefficients around zero for ethnicity compared to gender. With ethnicity, there appears to be slightly more extreme bias, which explains the larger coefficient.

As an additional test of bias, appendix E presents the results of a regression that combines gender and ethnicity. It examines whether there is a meaningful interaction between gender and ethnic bias. No significant effect is observed.[7]

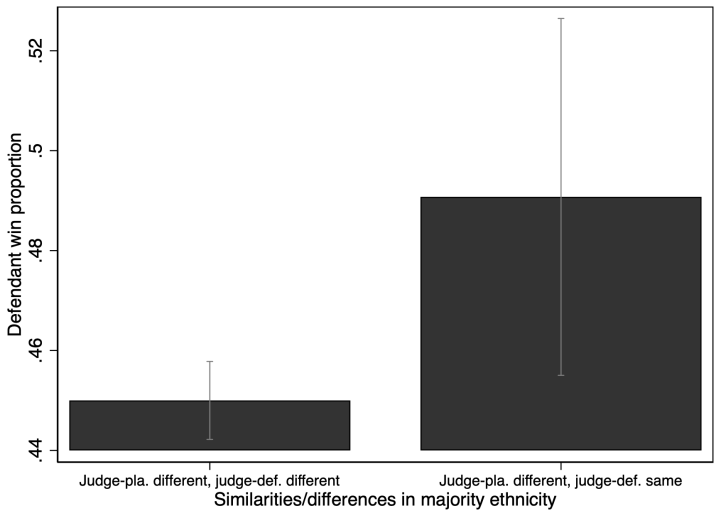
[7] Note that in column 2 there is a significant coefficient indicating out-group gender bias towards plaintiffs. However, significance is lost once defendant gender is controlled for (column 3). This could indicate the plaintiff gender is correlated with defendant gender, such that the signficant coefficient is be due to omitted variable bias.

Figure 10: defendant win proportion by similarities/differences in plurality ethnicity across judges, plaintiffs, and defendants

(a)



(b)



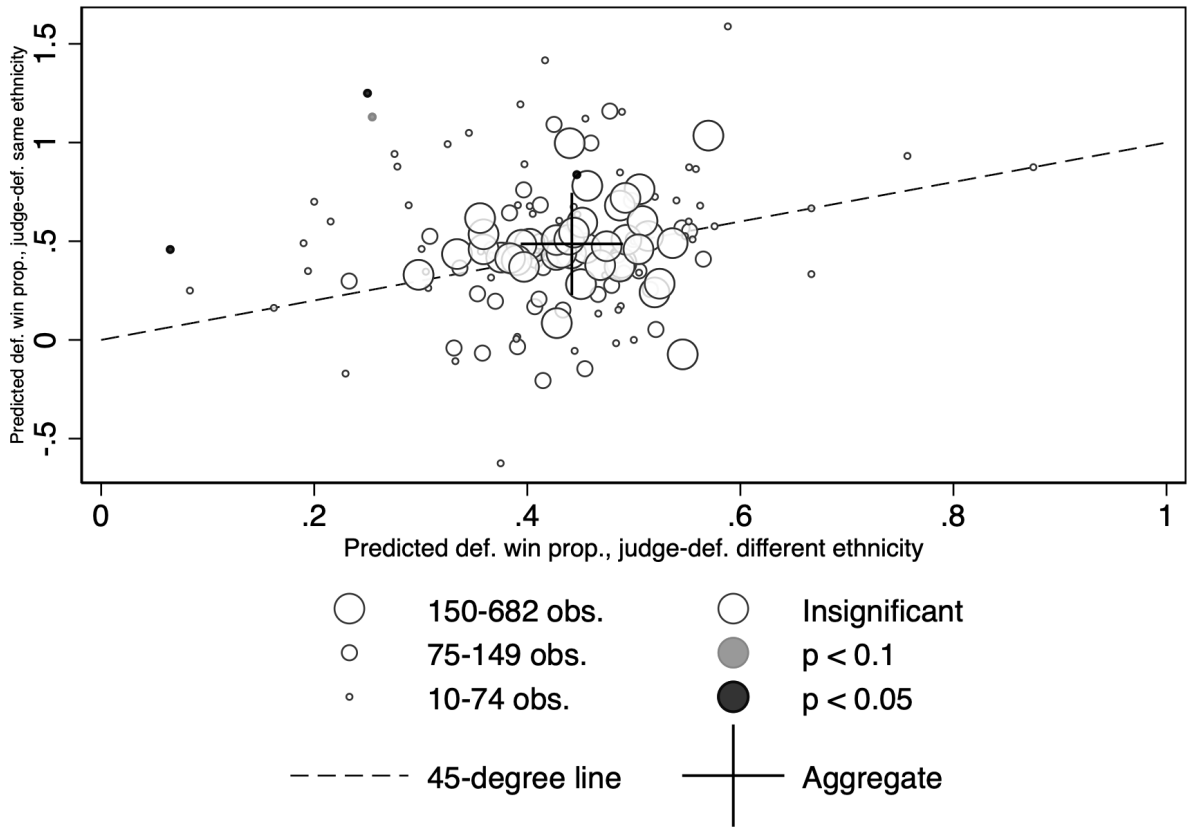
def. = defendant, pla. = plaintiff.

Table 4: Ethnicity results

	(1) Def. win	(2) Def. win	(3) Def. win	(4) Def. win	(5) Def. win
Judge-pla. same	0.00646		-0.0133	-0.00521	-0.00404
	(0.0126)		(0.0139)	(0.0153)	(0.0154)
Judge-def. same		0.0434***	0.0569***	0.0533***	0.0554***
		(0.0123)	(0.0144)	(0.0160)	(0.0157)
Court-year FE	Yes	Yes	Yes	Yes	Yes
Ethnicity dummies	No	No	No	Yes	Yes
Other controls	No	No	No	No	Yes
Observations	21964	21065	19008	19008	19008

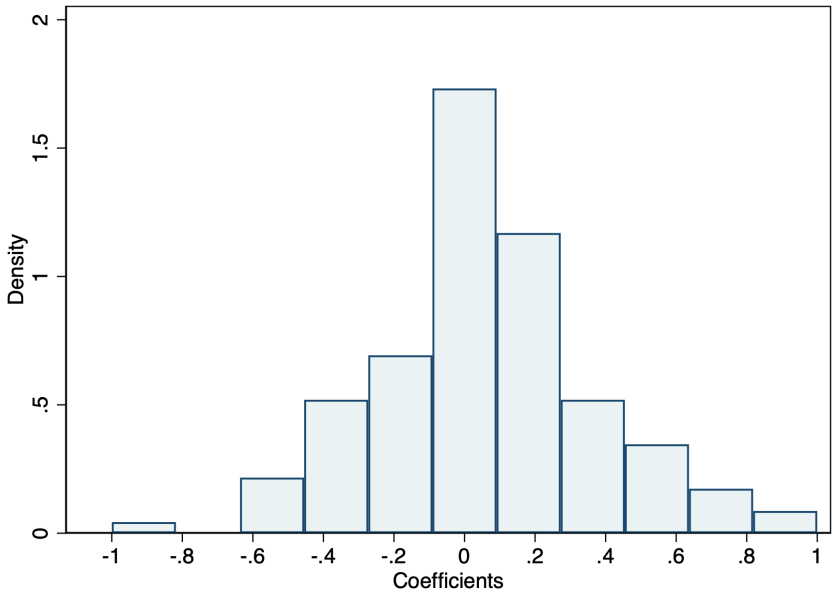
The regressions test whether defendants (plaintiffs) are more likely to win (lose) if they have the same (a different) plurality ethnicity as judges.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 5.
Judge-pla. same and Judge-def. same refer to similarity in plurality ethnicity.
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for both defendants and plaintifs.
Other controls include case type dummies; a dummy for an appeal case; variables for the numbers of defendants, plaintiffs, and judges; and dummies for defendant, plaintiff, and judge majority gender.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiff, def. = defendant.

Figure 11: Predicted defendant win proportion, by judge and by defendant similarity with judge ethnicity



def. = defendant, prop. = proportion. Each bubble indicates a specific judge. Only single judges are included, not judge panels. Judges without sufficient variation in outcomes were dropped. In total, 92 judges are included. The aggregate regression includes all single-judge panel observations, a total of 18,101. The outcome is significant at $p < 0.01$. Predictions are based on a regression with court-year fixed effects.

Figure 12: Distribution of coefficients estimating individual judges' in-group ethnic bias towards defendants



Coefficients are based on a regression with court-year fixed effects. Judges without sufficient variation in outcomes were dropped. In total, 92 judges are included.

5.4 Judgement text results

Tables 5 - 11 present the results from equation 6. Tables 5 - 7 present the results for gender in-group bias. They show that there is a significant negative correlation between the number of words in a judgement and in-group gender status for judges and defendants–but only when the defendant wins. This suggests that, when there is potential for gender in-group bias (i.e. when the judge and defendant are the same gender and the defendant wins), the judge tends to writer shorter judgements. We are not able to determine what drives this correlation. But it is possible that, when judges make biased judgements, they are less able to justify their decision based on solid legal grounds, and therefore write shorter judgements. Likewise, it is possible that biased judgements are not thought out as well, and are therefore accompanied by shorter written judgements.

The magnitude of the effect is relatively small. Column (4) of table 6 suggests that judges write about 143 fewer words when the defendant wins and they are the same gender as the judge. But as table B1 in appendix B shows, the mean and standard deviation for number of words in a judgement are 1452 and 1337, respectively.

Tables 8 - 10 present the results for ethnicity. They also provide evidence for biased decisions being associated with shorter written judgements. Again, the magnitude is relatively small. In addition, the tables show that, when the judge and defendant are the same ethnicity, the judgement is likely to be cited fewer times. Consistent with an in-group bias interpretation, the effect is strongest in the sample where the defendant wins, significant but weaker in the full sample, and null in the sample where the defendant loses. Though we cannot be certain what is driving this relationship, it may indicate that judges are less likely to cite cases with biased decisions.

Here, the effect is more substantial. The mean number of times cited is about 0.23, and column (2) of table 9 suggests that judgements are cited about 0.13 fewer times when the defendant wins and they are the same ethnicity as the judge.

Table 11 shows that most of these findings (with the exception of the relationship between in-group ethnic bias and the number of words) are robust to the inclusion of additional controls.

Table 5: Judgement text regressions, gender, full sample

	(1)	(2)	(3)	(4)
	Num. citations	Times cited	Num laws cited	Words in judg.
Judge maj. female	0.391**	-0.0931**	0.296*	85.43
	(0.193)	(0.0410)	(0.180)	(85.02)
Def. maj. female	-0.104	0.0400	-0.122	-27.51
	(0.0639)	(0.0609)	(0.0861)	(22.13)
Judge maj. fem. X def. maj. fem.	-0.0784	-0.0443	0.212	-28.45
	(0.122)	(0.0669)	(0.153)	(46.11)
Court-year FE	Yes	Yes	Yes	Yes
Observations	22889	22889	22889	22889

The regressions test whether in-group bias is associated with significantly different aspects of judges' written judgements. If in-group bias is associated with different characteristics for judgement texts, then we should see significant coefficients for the defendant-win sample but not the defendant-lose sample, and the coefficients in the defendant-win sample should be larger than in the full sample.
This table presents the full sample results.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 6.
Num. citations refers to the number of citations in the judgement.
Times cited refers to the number of times the case has been cited.
Num. laws cited refers to the number of laws and and acts cited in the judgement.
Words in judg. refers to the number of words in the written judgement.

Table 6: Judgement text regressions, gender, defendant win

	(1) Num. citations	(2) Times cited	(3) Num laws cited	(4) Words in judg.
Judge maj. female	0.328	-0.115**	0.162	92.25
	(0.203)	(0.0454)	(0.213)	(85.41)
Def. maj. female	-0.0632	-0.0638	-0.195	57.96
	(0.0856)	(0.0837)	(0.127)	(35.75)
Judge maj. fem. X def. maj. fem.	-0.184	0.0796	0.152	-143.2**
	(0.175)	(0.0969)	(0.205)	(67.68)
Court-year FE	Yes	Yes	Yes	Yes
Observations	10279	10279	10279	10279

The regressions test whether in-group bias is associated with significantly different aspects of judges' written judgements. If in-group bias is associated with different characteristics for judgement texts, then we should see significant coefficients for the defendant-win sample but not the defendant-lose sample, and the coefficients in the defendant-win sample should be larger than in the full sample.
This table presents the defendant-win sample results.
All columns are based on a linear regression model. For specification details, see equation 6.
Num. citations refers to the number of citations in the judgement.
Times cited refers to the number of times the case has been cited.
Num. laws cited refers to the number of laws and and acts cited in the judgement.
Words in judg. refers to the number of words in the written judgement.

Table 7: Judgement text regressions, gender, defendant lose

	(1) Num. citations	(2) Times cited	(3) Num laws cited	(4) Words in judg.
Judge maj. female	0.442**	-0.0938	0.398**	71.55
	(0.215)	(0.0627)	(0.196)	(96.05)
Def. maj. female	-0.136	0.100	-0.0523	-104.5***
	(0.102)	(0.0763)	(0.100)	(35.92)
Judge maj. fem. X def. maj. fem.	-0.0381	-0.110	0.211	69.77
	(0.160)	(0.0813)	(0.184)	(58.02)
Court-year FE	Yes	Yes	Yes	Yes
Observations	12469	12469	12469	12469

The regressions test whether in-group bias is associated with significantly different aspects of judges' written judgements. If in-group bias is associated with different characteristics for judgement texts, then we should see significant coefficients for the defendant-win sample but not the defendant-lose sample, and the coefficients in the defendant-win sample should be larger than in the full sample.
This table presents the defendant-lose sample results.
All columns are based on a linear regression model. For specification details, see equation 6.
Num. citations refers to the number of citations in the judgement.
Times cited refers to the number of times the case has been cited.
Num. laws cited refers to the number of laws and and acts cited in the judgement.
Words in judg. refers to the number of words in the written judgement.

Table 8: Judgement text regressions, ethnicity, full sample

	(1) Num. citations	(2) Times cited	(3) Num laws cited	(4) Words in judg.
Judge-defendant same ethnicity=1	-0.131	-0.0807***	-0.159	-66.79
	(0.109)	(0.0309)	(0.144)	(52.80)
Court-year FE	Yes	Yes	Yes	Yes
Observations	21065	21065	21065	21065

The regressions test whether in-group bias is associated with significantly different aspects of judges' written judgements. If in-group bias is associated with different characteristics for judgement texts, then we should see significant coefficients for the defendant-win sample but not the defendant-lose sample, and the coefficients in the defendant-win sample should be larger than in the full sample.
This table presents the full sample results.
All columns are based on a linear regression model. For specification details, see equation 6.
Standard errors, in parentheses, are clustered at the judge level.
Num. citations refers to the number of citations in the judgement.
Times cited refers to the number of times the case has been cited.
Num. laws cited refers to the number of laws and and acts cited in the judgement.
Words in judg. refers to the number of words in the written judgement.

Table 9: Judgement text regressions, ethnicity, defendant win

	(1) Num. citations	(2) Times cited	(3) Num laws cited	(4) Words in judg.
Judge-defendant same ethnicity=1	-0.145	-0.133**	-0.235	-95.20*
	(0.136)	(0.0582)	(0.176)	(57.25)
Court-year FE	Yes	Yes	Yes	Yes
Observations	9477	9477	9477	9477

The regressions test whether in-group bias is associated with significantly different aspects of judges' written judgements. If in-group bias is associated with different characteristics for judgement texts, then we should see significant coefficients for the defendant-win sample but not the defendant-lose sample, and the coefficients in the defendant-win sample should be larger than in the full sample.
This table presents the defendant-win sample results.
All columns are based on a linear regression model. For specification details, see equation 6.
Standard errors, in parentheses, are clustered at the judge level.
Num. citations refers to the number of citations in the judgement.
Times cited refers to the number of times the case has been cited.
Num. laws cited refers to the number of laws and and acts cited in the judgement.
Words in judg. refers to the number of words in the written judgement.

Table 10: Judgement text regressions, ethnicity, defendant lose

	(1) Num. citations	(2) Times cited	(3) Num laws cited	(4) Words in judg.
Judge-defendant same ethnicity=1	-0.0930	-0.0324	-0.151	-50.45
	(0.146)	(0.0360)	(0.182)	(61.29)
Court-year FE	Yes	Yes	Yes	Yes
Observations	11445	11445	11445	11445

The regressions test whether in-group bias is associated with significantly different aspects of judges' written judgements. If in-group bias is associated with different characteristics for judgement texts, then we should see significant coefficients for the defendant-win sample but not the defendant-lose sample, and the coefficients in the defendant-win sample should be larger than in the full sample.
This table presents the defendant-lose sample results.
All columns are based on a linear regression model. For specification details, see equation 6.
Standard errors, in parentheses, are clustered at the judge level.
Num. citations refers to the number of citations in the judgement.
Times cited refers to the number of times the case has been cited.
Num. laws cited refers to the number of laws and and acts cited in the judgement.
Words in judg. refers to the number of words in the written judgement.

Table 11: Judgement text regressions, additional controls, defendant win

	(1) Times cited	(2) Times cited	(3) Words in judg.	(4) Words in judg.
Judge maj. female	-0.0990*	-0.0776	89.39	108.3
	(0.0534)	(0.0522)	(92.08)	(91.35)
Def. maj. female	-0.0233	0.0216	56.55	80.95
	(0.0838)	(0.0958)	(46.81)	(49.43)
Judge maj. fem. X def. maj. fem.	0.0432	0.0263	-138.9*	-152.6**
	(0.0988)	(0.104)	(73.62)	(72.84)
Judge-defendant same ethnicity=1	-0.136**	-0.125*	-93.60	-60.55
	(0.0610)	(0.0683)	(59.38)	(62.64)
Court-year FE	Yes	Yes	Yes	Yes
Other controls	No	Yes	No	Yes
Observations	7991	7991	7991	7991

The regressions test whether in-group bias is associated with significantly different aspects of judges' written judgements.
Standard errors, in parentheses, are clustered at the judge level.
Num. citations refers to the number of citations in the judgement.
Times cited refers to the number of times the case has been cited.
Num. laws cited refers to the number of laws and and acts cited in the judgement.
Words in judg. refers to the number of words in the written judgement.
Other controls include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges; case type dummies; a dummy for an appeal case; and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Def. = defendant

6 Conclusion

In this paper, we examine the extent and determinants of judicial bias in Kenya, with a focus on gender and ethnic in-group bias. Our data cover Kenyan higher court cases spanning 1976-2020 and our identification strategy relies on the random assignment of judges to cases. Our analysis also looks at the relationship between bias in judge decisions and measures of slant against women in judges' written decisions, which we derive through machine learning techniques.

Our main finding is that judges in Kenya display both gender and ethnic in-group bias towards defendants. Our results suggest that defendants are about 4 percentage points more likely to win if they share the judge's gender and about 5 percentage points more likely to win if they share the judge's ethnicity. As such, bias is present but relatively mild, with most judges displaying very little in-group bias. We find no evidence of in-group bias towards plaintiffs.

We also find evidence that slant against women in written judgements is associated with lower win-rates for female defendants. The results show that a one standard deviation change in the measure of gender slant is associated with about a 2 percentage point decrease in win probability for female defendants. Finally, we show that potentially biased judgements are associated with shorter written judgements (for gender and ethnic bias) that are less likely to be cited (for ethnic bias), which suggests that biased decisions are linked to poorer quality written judgements.

These findings have important implications for the Kenyan context. Women and certain ethnic groups are underrepresented in the judiciary. As such, they are more likely to be negatively affected by in-group bias. In concrete terms–since the main cases in the dataset are civil cases, environment and land cases, and succession cases–in-group bias might imply a financial disadvantage, greater likelihood of losing disputes over land ownership, or being cut out of family inheritance or property.

Several approaches could be taken to reduce bias. Primarily, greater efforts could be made to achieve equal representation of female judges and representation of ethnic groups relative to their proportion of the total population. Second, implicit bias trainings, which have been proven effective in some settings (Jackson, Hillard, and Schneider 2014), could be implemented for judges. Third, judges could simply be provided with data on the extent of their bias in decision-making. Some research has shown that the provision of information on biases can lead to more action in favor of out-groups (Hillard, Ryan, and Gervais 2013). Importantly, the application of these approaches to the Kenyan context should be rigorously tested.

The findings also make important contributions to the literature on judicial bias. They expand the study of in-group judicial bias outside the most heavily studied contexts and provide further evidence that such bias may be prevalent across many contexts. They are also the first to show that judges may exhibit greater bias towards defendants than plaintiffs, a phenomenon which is consistent with social identity theory. Furthermore, they contribute to the broader literature on ethnic bias in Kenya and sub-Saharan Africa more broadly, showing that ethnic preferences influence decision-making in courts. They also build on the literatures related to gender discrimination and the importance of female representation in the public sector. Finally, the paper presents a novel application of machine learning techniques to help understand the determinants of bias. Future research should focus on further unveiling the determinants and scope of bias in the judiciary, as well as on how to reduce the presence of bias in the judiciary.

References

Akech, Migai (2010). *Institutional Reform in the New Constitution of Kenya*. International Center for Transitional Justice.

— (2011). "Abuse of Power and Corruption in Kenya: Will the New Constitution Enhance Government Accountability?" In: *Ind. J. of Global Legal Studies* 341, pp. 377–378.

Antoniak, Maria and David Mimno (2018). "Evaluating the stability of embedding-based word similarities". In: *Transactions of the Association for Computational Linguistics* 6, pp. 107–119.

Ash, Elliott, Sam Asher, et al. (2021). "Measuring Gender and Religious Bias in the Indian Judiciary". In: *Center for Law and Economics Working Paper Series* 3.

Ash, Elliott, Daniel Chen, and Arianna Ornaghi (2021). "Gender Attitudes in the Judiciary: Evidence from U.S. Circuit Courts". In: *Working Paper*.

Asingo, Patrick et al. (2018). *Ethnicity and Politicization in Kenya*. Kenya Human Rights Commission.

Azmat, Ghazala and Barbara Petrongolo (2014). "Gender and the labor market: What have we learned from field and lab experiments?" In: *Labour Economics* 30, pp. 32–40.

Barkan, Joel and Michael Chege (1989). "Decentralising the state: district focus and the politics of reallocation in Kenya". In: *The Journal of Modern African Studies* 27.3, pp. 431–453.

Beaman, Lori et al. (2009). "Powerful Women: Does Exposure Reduce Bias?" In: *The Quarterly Journal of Economics* 124.4, pp. 1497–1540.

Berge, Lars et al. (2015). "How strong are ethnic preferences?" In: *Working paper*.

Burgess, Robin et al. (2015). "The value of democracy: evidence from road building in Kenya". In: *American Economic Review* 105.6, pp. 1817–1851.

Carlana, Michela (2019). "Implicit Stereotypes: Evidence from Teachers' Gender Bias". In: *The Quarterly Journal of Economics* 134.3, pp. 1163–1224.

Cederman, Lars-Erik, Andreas Wimmer, and Brian Min (2010). "Why do ethnic groups rebel? New data and analysis". In: *World Politics* 62.1, pp. 87–119.

Depew, Briggs, Ozkan Eren, and Naci Mocan (2017). "Judges, juveniles, and in-group bias". In: *The Journal of Law and Economics* 60.2, pp. 209–239.

Dietz-Uhler, Beth and Audrey Murrell (1998). "Effects of social identity and threat on self-esteem and group attributions". In: *Group Dynamics: Theory, Research, and Practice* 2.1.

Friedrich-Ebert-Stiftung (2012). *Regional Disparities and Marginalisation in Kenya*.

Gainer, Maya (2015). "Transforming the Courts: Judicial Sector Reforms in Kenya". In: *Princeton University Innovations for Successful Societies* 1.7.

— (2016). "How Kenya Cleaned Up Its Courts". In: *Foreign Policy*.

Gazal-Ayal, Oren and Raanan Sulitzeanu-Kenan (2010). "Let My People Go: Ethnic In-Group Bias in Judicial Decisions–Evidence from a Randomized Natural Experiment". In: *Journal of Empirical Legal Studies* 7.3, pp. 403–428.

Harris, Andrew (2014). *Replication data for: What's in a name? A Method for Extracting Information about Ethnicity from Names*. URL: `https://doi.org/10.7910/DVN/27691`.

Hessami, Zohal and Mariana Lopes da Fonseca (2020). "Female political representation and substantive effects on policies: A literature review". In: *European Journal of Political Economy* 63.

Hillard, Amy, Carey Ryan, and Sarah J. Gervais (2013). "Reactions to the implicit association test as an educational tool: A mixed methods study". In: *Social Psychology of Education* 16, pp. 495–516.

Hochreiter, Sepp and Jurgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

IDLO (2020). *Women's Professional Participation in Kenya's Justice Sector: Barriers and Pathways*. International Development Law Organization.

Islam, Gazi (2014). "Social Identity Theory". In: *Journal of personality and Social Psychology* 67, pp. 741–763.

Jackson, Sarah, Amy Hillard, and Tamera Schneider (2014). "Reactions to the implicit association test as an educational tool: A mixed methods study". In: *Social Psychology of Education* 17, pp. 419–438.

Kenyan Judiciary (2021). *Courts: Overview*. URL: `https://www.judiciary.go.ke/courts/`.

KNBS (2019). *2019 Kenya Population and Housing Census Volume IV: Distribution of Population by Socio-Economic Characteristics*.

Knepper, Matthew (2018). "When the shadow is the substance: Judge gender and the outcomes of workplace sex discrimination cases". In: *Journal of Labor Economics* 36.3, pp. 623–664.

Kozlowski, Austin, Matt Taddy, and James Evans (2019). "The geometry of culture: Analyzing the meanings of class through word embeddings". In: *American Sociological Review* 84.5, pp. 905–949.

Miguel, Edward and Mary Kay Gugerty (2005). "Ethnic diversity, social sanctions, and public goods in Kenya". In: *Journal of Public Economics* 89.11-12, pp. 2325–2368.

Mutunga, Willy (2011). *Progress Report On The Transformation Of The Judiciary*. URL: http://kenyalaw.org/kenyalawblog/progress-report-on-the-transformation-of-the-judiciary/.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing*.

Ponticelli, Jacopo and Leonardo Alencar (2016). "Court enforcement, bank loans, and firm investment: Evidence from a bankruptcy reform in Brazil". In: *The Quarterly Journal of Economics* 131.3, pp. 1365–1413.

Rodrik, Dani (2000). "Institutions for high-quality growth: what they are and how to acquire them". In: *Studies in comparative international development* 35.3, pp. 3–31.

Shayo, Moses and Asaf Zussman (2011). "Judicial ingroup bias in the shadow of terrorism". In: *The Quarterly Journal of Economics* 126.3, pp. 1447–1484.

Sloan, CarlyWill (2020). "Racial bias by prosecutors: Evidence from random assignment". In: *Working paper*.

Spirling, Arthur and Pedro Rodriguez (2019). "Word embeddings: What works, what doesn't, and how to tell the difference for applied research". In: *Journal of Politics*.

UNDP (2020). *Gender Inequality Index*. URL: http://hdr.undp.org/en/content/gender-inequality-index-gii.

Visaria, Sujata (2009). "Legal reform and loan repayment: The microeconomic impact of debt recovery tribunals in India". In: *American Economic Journal: Applied Economics* 1.3, pp. 59–81.

Voci, Alberto (2010). "The link between identification and in-group favouritism: Effects of threat to social identity and trust-related emotions". In: *British Journal of Social Psychology* 45.2, pp. 265–284.

Wann, Daniel and Frederick Grieve (2005). "Biased Evaluations of In-Group and Out-Group Spectator Behavior at Sporting Events: The Importance of Team Identification and Threats to Social Identity". In: *The journal of social psychology* 145.5, pp. 531–546.

World Bank (2017). *World Development Report: Governance and the Law*. The World Bank Group.

— (2021). *Ease of Doing Business in Kenya*. URL: https://www.doingbusiness.org/en/data/exploreeconomies/kenya#DB_ec.

Appendix

Appendix A: Variable construction

Constructing variables with judge, defendant, and plaintiff information

The names of judges, defendants, and plaintiffs were used to remove non-humans and to extract additional information for each case, including gender, ethnicity, and the number of judges and litigants. Cases were identified as non-human and removed if either the plaintiff or defendant name included any of a long list of key words, such as "republic," "company," or "medical." A full list of the keywords can be found in the cleaning scripts posted online.[8]

Afterwards, we could determine the gender of each individual using their first name and the ethnicity of each individual using their last name. To assign gender based on first names we used the genderize.io API and Gender API, both of which use global databases of names and genders to probabilistically assign gender to names.[9] One exception was for the judges, for whom gender was assigned manually.

To assign ethnicity based on last names, we used data available on Harvard Dataverse that links names to ethnicities (Harris 2014). This data could be used to identify 12 ethnic groups (Meru, Kisii, Kalenjin, Kamba, Luo, Turkana, Mijikenda, Luhya, Kikuyu, Somali, Masai, and Pokot). This includes one ethnic sub-group, the Pokot, which is a sub-group of the Kalenjin. Throughout our analysis, Kalenjin refers to non-Pokot Kalenjin. Together, these groups account for about 91 percent of the population of Kenya. Of the other 29 major ethnic groups (i.e. non-subgroups) identified in the 2019 census, the largest group accounts for only about 0.9 percent of the population.

Gender and ethnicity could not be determined for all individuals in all cases. Gender could not be determined if the first name was either abbreviated (i.e. if only initials were given), it did not clearly match to a single gender, or it was not included in the API datasets. Ethnicity could not be determined if the last name was not included in the ethnicity dataset. For some of the cases included in analysis, information could be extracted for plaintiffs but not defendants (and vice versa) and for gender but not ethnicity (and vice versa).

It is important to note that there is the possibility of a small amount of error resulting from the automated process of removing non-humans and determining gender and ethnicity. For example, although the list of key words for non-humans is long and we have manually scanned the data for non-humans, it is still possible that some non-humans remain. It is also possible that gender and/or ethnicity has been assigned to non-humans with certain key words included in the organization name. Similarly, if names were separated in an unusual way, it is possible that the number of defendants or plaintiffs was incorrectly counted, possibly resulting in an incorrect assignment of majority/plurality gender/ethnicity. However, having thoroughly scanned the data, we are confident that the number of such errors is insignificant.

Using the Binary Classification Machine Learning Model to construct the defendant_win outcome variable

To determine the winner of each case, we created a Binary Classification Machine Learning Model using the Global Vectors for Word Representation (GloVe) algorithm (Pennington, Socher, and Manning 2014). The objective function of GloVe can be written as follows:

$$J(w) = \sum f(X_{ij})(w_i^t w_j - logX_{ij}))^2 \tag{7}$$

where X_{ij} denotes the co-occurrence count between words i and j, and $f(\cdot)$ is a weighting function that serves to down-weight particularly frequent words. The objective function $J(\cdot)$ trains the word vectors to minimize the squared difference between the dot product of the vectors representing two words and their empirical co-occurrence in the corpus. The algorithm requires two hyperparameters, *dimensionality* of the vectors and the *window* size for computing co-occurrence statistics. Prior research has found 300 to be the optimum size in many a cases and that increasing dimensionality beyond 300 has negligible improvements for downstream tasks (Pennington, Socher, and Manning 2014; Spirling and Rodriguez 2019). Following that

[8] (Provide link)

[9] See the following websites: https://genderize.io/; https://gender-api.com/.

literature, we train 300 dimensional vectors. We used a standard 10-word window size, in between a shorter window size (which tends to capture syntactic/functional relations between words) and a longer window size (which tends to capture topical relations between words). To improve accuracy, the classification model was also comprised of a Long Short-Term Memory layer in addition to the fully connected neural network layers and the initial embedding layer (Hochreiter and Schmidhuber 1997).

Applying this model to our data, we used the bottom 500 words of the case judgements, since the outcomes were found to be present towards the bottom of the judgements. As a training dataset, we applied the model to cases for which we could determine the outcome (in favor or against the defendant) directly from the case outcome variable of the metadata. There were 49,706, 6,214, and 6,213 cases in the training, testing, and validation sets, respectively. The results of the model were as follows:

Training set accuracy	92.44%
Validation set accuracy	91.92%
Test set accuracy (on previously unseen data)	92.83%
Accuracy	0.928388
Precision	0.896705
Recall	0.959647
F1 score	0.927109

Using word embeddings to determine textual slant

To determine each judge's textual gender slant (i.e. the degree to which each judge exhibits gender bias in their written judgements), we make use of word embeddings, which model the text present in the judgements in the form of low dimensional euclidean space vectors (Pennington, Socher, and Manning 2014). In other words, word embeddings are low dimensional vectors which can accommodate large vocabularies and corpora without increasing dimensionality. The representation resulting from them captures relations between the words. In order to catch semantic similarity amongst words, the positions are assigned to word vectors in the euclidean space, such that the words that appear frequently in the same context have representations close to each other in the space, while words that appear rarely together have representations that are far apart.

To train our word embeddings, we used the GloVe algorithm, described above. The embeddings we trained were then used for identification of cultural dimensions in language (Kozlowski, Taddy, and Evans 2019). That is, we identified a gender dimension by taking the difference between the average normalized vector across a set of male words and the average normalized vector across a set of female words, as such:

$$\vec{male} - \vec{female} = \sum\nolimits_n \vec{maleword}_n / \mid N_{male} \mid + \sum\nolimits_n \vec{femaleword}_n / \mid N_{female} \mid$$

where N_{male} is the number of words used to identify the male dimension. In order to determine the similarity within these dimensions, we used cosine similarity as a measure, defined as follows:

$$sim(\vec{x}, \vec{y}) = cos(\theta) = (\vec{x} \cdot \vec{x}) / (\parallel \vec{x} \parallel \parallel \vec{y} \parallel)$$

where $\vec{x}$ and $\vec{y}$ are non-zero vectors, θ is the associated angle, and $\parallel \cdot \parallel$ is the 2-norm. Therefore, we can see that words with male (female) connotations are going to be positively (negatively) correlated with the gender dimension defined by $\vec{male} - \vec{female}$.

These dimensions were then used to construct the gender slant measures. For the first, we aimed to capture the strength of the association between gender and stereotypical attitudes, which identify men more closely with careers and women with family. Specifically, we used the cosine similarity between the vector representing the gender dimension, defined by $\vec{male} - \vec{female}$, and the vector representing the career-family dimension, defined by $\vec{career} - \vec{family}$. For our second measure, we aimed to capture stereotypical attitudes that associate men with "good" and women with "bad" words. For this measure, instead of $\vec{career} - \vec{family}$, we used $\vec{good} - \vec{bad}$.

For the $\vec{male} - \vec{female}$ dimension, we used various gender-specific words which were found out to be the five most frequently occurring in our corpus. Words for $\vec{career} - \vec{family}$ and $\vec{good} - \vec{bad}$ were chosen in a similar fashion. Only five words were chosen for each because, given the relatively small size of the corpus,

Table A1: Words used for each vector dimension

Vector dimension	Words
$\vec{MaleNames}$	john, joseph, peter, james, david
$\vec{FemaleNames}$	faith, mary, rose, jane, margaret
$\vec{Male}$	his, he, him, mr, himself
$\vec{Female}$	her, she, ms, mrs, herself
$\vec{Good}$	competent, strong, power, serious, professional
$\vec{Bad}$	frivolous, vain, incompetent, unreasonable, incapable
$\vec{Career}$	company, service, pay, business, work
$\vec{Family}$	family, wife, mother, father, brother

Each dimensions includes the five most common relevant words in the corpus. Only five words were chosen for each because, given the relatively small size of the corpus, the inclusion of too many words could results in invalid measures of slant.

the inclusion of too many words could result in invalid measures of slant. The word used are displayed in table A1.

To apply this process to the data, we first preprocessed the entire Kenya Law corpus of judgements by removing punctuations (but retaining hyphenated words). To avoid case sensitivity, we transformed all our words to lower case. We then retained only the most common 50,000 words in all judicial opinions. To obtain judge-specific gender slant measures, we took the set of majority opinions authored by each judge as a separate corpus and trained separate GloVe embeddings on each judge's corpus. To ensure convergence, we trained vectors for 20 iterations with a learning rate of 0.05.

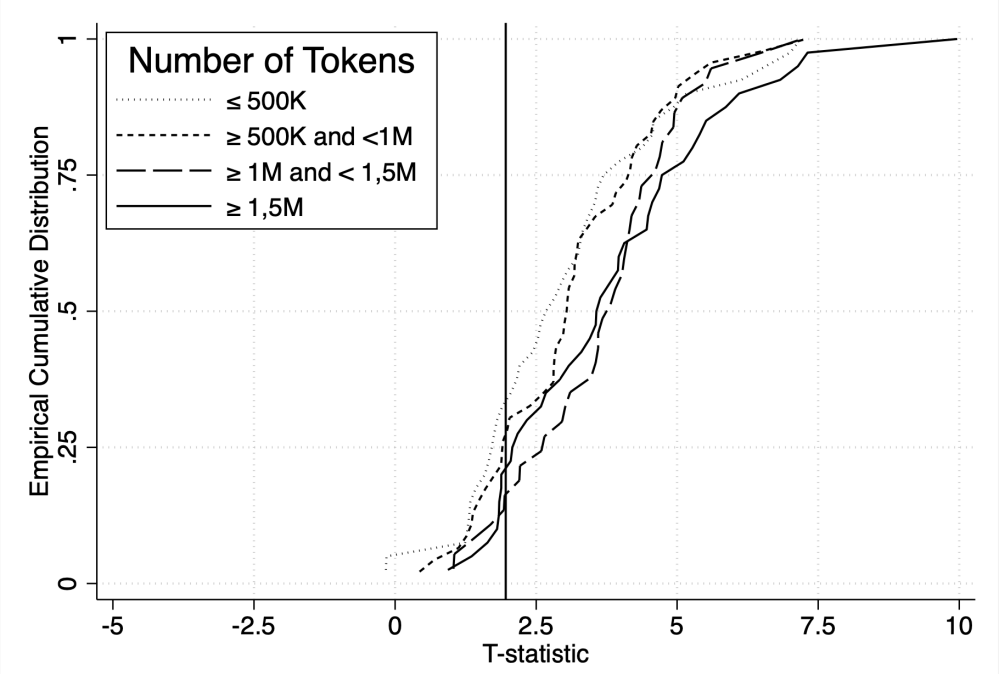
Since each judge might not have a sufficiently large number of tokens, we follow the approach suggested by Antoniak and Mimno (2018) and train embedding models on 25 bootstrap samples of each judge corpus. Specifically, we consider each sentence written by a judge as a document and then create a corpus by sampling with replacement from all sentences. The number of sentences contained in the bootstrapped sample is the same as the total number of sentences in the original judge corpus. We then calculate our slant measure for all bootstrap samples and assign to each judge the median value of the measure across the samples. Given that embeddings trained on small corpora tend to be sensitive to the inclusion of specific documents, the bootstrap procedure produces more stable results. In addition, bootstrapping ensures stability with respect to the initialization of the word vectors–a potential concern given that GloVe presents a non-convex objective function (Spirling and Rodriguez 2019). The two variables resulting from this process are *Median slant, career vs. family* and *Median slant, good vs. bad*. For both measures, positive values indicate greater slant against women.

To validate that the embeddings capture meaningful information about gender, after following the bootstrapping procedure, we compute the cosine similarity between the gender dimension and each of the vectors representing the five most common male and female names for each judge and bootstrap sample. We then regress a dummy for whether the name is male on the median cosine similarity between the vector representing the name and the gender dimension across bootstrap samples, separately for each judge. Figure A1 shows the cumulative distribution of the t-statistics resulting from these regressions for sets of judges with different numbers of tokens. It shows that most t-statistics are significant (and they are never lower than zero). This shows that the gender dimension identified in the embeddings does indeed contain meaningful gender information.

Constructing other textual variables

To create the measure of the number of cases cited in the text, we extracted a window of 10 words (5 on each side) around the words v, vs, and ndashvs (because sometimes HTML elements from the website are included in the text), which were found to be a common way of citing other judgements. This window of 10 words was then cleaned to produce the final cited judgements. A similar process was used to cite the number of laws and acts cited in a case. Once we had the information on citations in each case, we were able to also determine the number of times each case in the dataset was cited.

Figure A1: Cumulative distribution of t-statistics from regressions testing the validity of the word embeddings



The vertical line indicates T-stat=1.96, for significance at $p<0.05$.
T-statistics are from regressions between a dummy for whether the name is male on the median cosine similarity between the vector representing the name and the gender dimension across bootstrap samples, separately for each judge.

Appendix B: Variable summaries

Table B1: Summary of main variables

	count	mean	sd	min	max
Def. win	29571	.4297115	.4950433	0	1
Judge maj. female	28814	.3658291	.4816702	0	1
Pla. maj. female	26564	.2489836	.4324324	0	1
Def. maj. female	23647	.2374508	.4255298	0	1
Judge-plaintiff same ethnicity	22079	.1321165	.338625	0	1
Judge-defendant same ethnicity	21188	.1254484	.3312344	0	1
Appeal	29571	.1444997	.3516016	0	1
Number of defendants	29571	1.581786	1.436409	1	68
Number of plaintiffs	29571	1.314768	1.132962	1	65
Number of judges	29571	1.109973	.4596086	1	9
Median slant, career v family	26269	-.0281819	.098839	-.2812362	.30826
Median slant, good v bad	22172	.0616616	.0549197	-.0875989	.2815675
Case type: civil	28707	.4612116	.4985019	0	1
Case type: tax	28707	.0022991	.0478945	0	1
Case type: human rights	28707	.0011844	.034395	0	1
Case type: judicial review	28707	.0009405	.0306543	0	1
Case type: criminal	28707	.0071063	.0840002	0	1
Case type: divorce	28707	.0019856	.0445163	0	1
Case type: election	28707	.0018462	.042929	0	1
Case type: labor relations	28707	.0165813	.1276987	0	1
Case type: environment and land	28707	.3208277	.4668028	0	1
Case type: family	28707	.0067579	.0819298	0	1
Case type: industrial	28707	.0033093	.0574322	0	1
Case type: miscellaneous	28707	.081339	.2733599	0	1
Case type: succession	28707	.0946111	.2926821	0	1
Number of cases cited in judgement	29571	1.93125	3.537811	0	87
Times judgement cited	29571	.232356	1.932199	0	109
Laws cited in judgement	29571	2.221535	4.107244	0	146
Words in judgement	29571	1452.416	1337.277	0	42980

Table B2: Summary of main variables, count only

	count
Court ID	29571
Year of delivery	29571
Court-year FE	29571
Plurality ethnicity of plaintiffs	24486
Plurality ethnicity of defendants	23502
Plurality ethnicity of judges	26606

Appendix C: Additional descriptive statistics

Table C1: Frequency of court types in the dataset

	Frequency
Court of appeal	1674
Employment and labor relations	1090
Environment and land court	8616
High court	18042
Other	125
Supreme court	24
Total	29571

Other includes Election Petition in Magistrate Courts, the Judges and Magistrates Vetting Board, Kadhis Courts, and the National Environment Tribunal

Figure C1: Total number of cases, by majority gender and role in the case

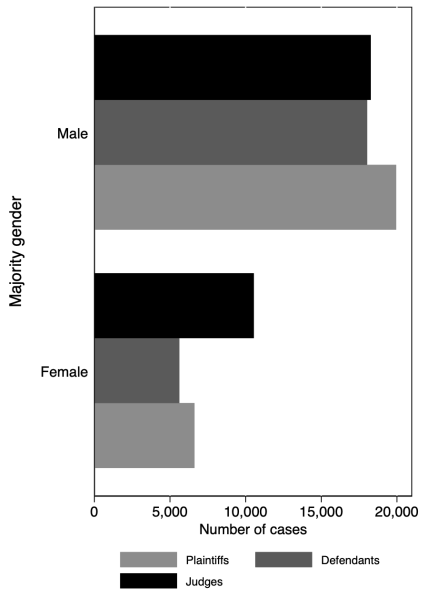


Figure C2: Proportion of cases over time with majority female judges, defendants, and plaintiffs

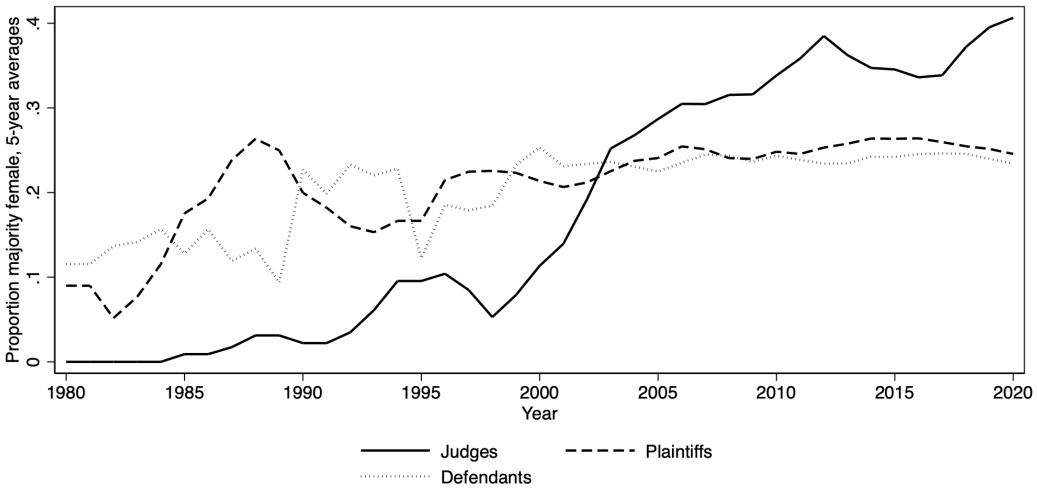
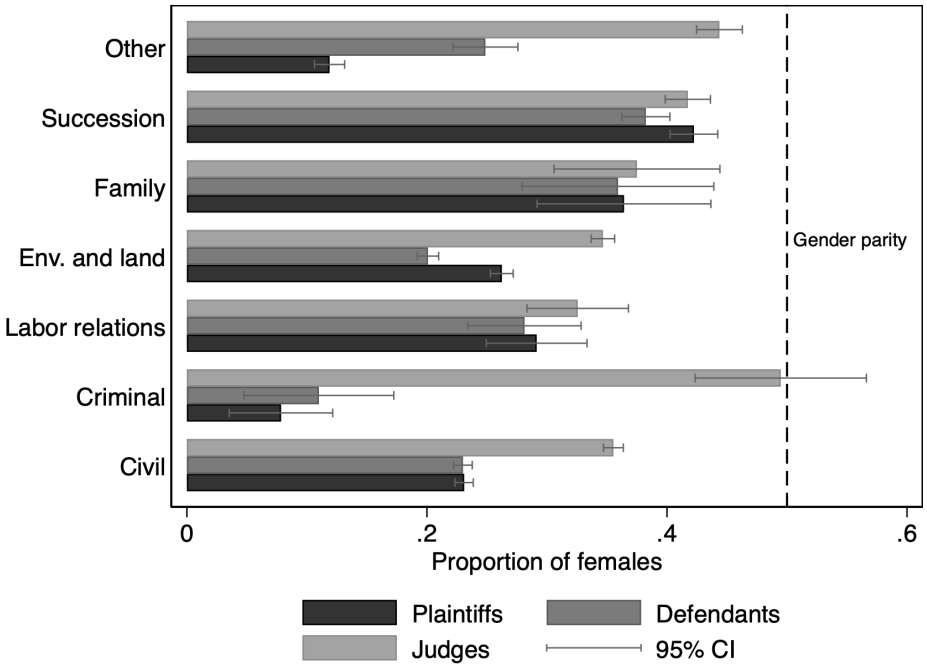
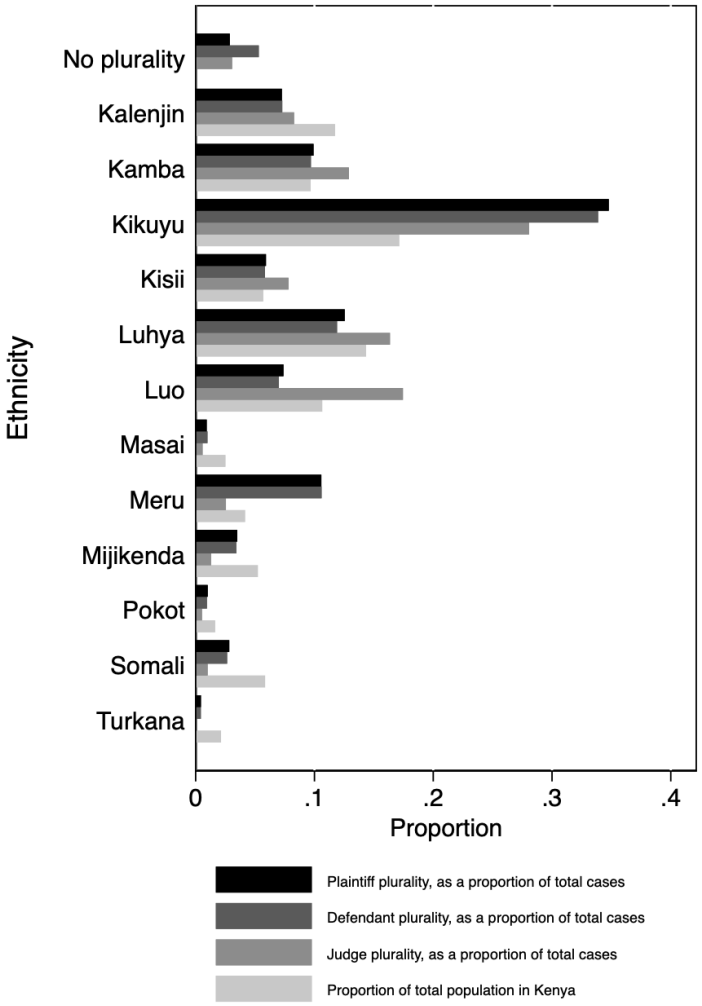


Figure C3: Proportion of female majorities, by case type and role in the case



See appendix B for list of case types included in "other."

Figure C4: Ethnicities as a proportion of total cases (by role in the case) and the total population in Kenya



Proportions of the total population are derived from the 2019 census. Kalenjin refers to non-Pokot Kalenjin.

Table C2: Various power statuses occupied by each ethnic group, 1979-2017

	Discriminated	Junior partner	Powerless	Senior partner	Total
Kalenjin	0	1	0	1	2
Kamba	0	1	0	0	1
Kikuyu	1	0	0	1	2
Kisii	0	1	1	0	2
Luhya	0	1	0	0	1
Luo	1	1	0	1	3
Masai	0	1	0	1	2
Meru	1	0	0	1	2
Mijikenda	0	1	0	0	1
Somali	1	0	0	0	1
Turkana	0	1	0	1	2
Total	4	8	1	6	19

1=status occupied at some point. 0=status not occupied at any point.
Power statuses are based on the EPR dataset (Cederman, Wimmer, and Min 2010)
Pokot is not included because it is a subgroup of Kalenjin.
Senior partner indicates that representatives from the ethnic group participate as senior partners in a power-sharing agreement for control of the executive branch of government.
Junior partner indicates that representatives from the ethnic group participate as junior partners in a power-sharing agreement for control of the executive bracnh of government.
Powerless indicates that representatives hold no political power at either the national or the regional level without being explicitly discriminated against.
Discriminated indicates that group members are subjected to active, intentional, and targeted discrimination, with the intent of excluding them from both regional and national power.

Appendix D: Balance tests, before and after 2011

Table D1: Gender randomization checks

	(1) Judge maj. female	(2) Judge maj. female	(3) Judge maj. female
Pla. maj. female	0.0116	0.0119	0.00755
	(0.00859)	(0.00857)	(0.00682)
Def. maj. female	0.00348	0.00335	-0.000734
	(0.00755)	(0.00748)	(0.00645)
Court-year FE	Yes	Yes	Yes
Ethnicity dummies	No	Yes	Yes
Other controls	No	No	Yes
Observations	20437	20437	20437

The regressions test whether female plaintiffs/defendants are more likely to be matched with female judges than male judges.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 1.
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, maj. = majority.

Table D2: Gender randomization checks, before 2011

	(1) Judge maj. female	(2) Judge maj. female	(3) Judge maj. female
Pla. maj. female	0.0227	0.0216	0.0122
	(0.0189)	(0.0188)	(0.0158)
Def. maj. female	-0.00631	-0.00555	-0.0122
	(0.0136)	(0.0132)	(0.0106)
Court-year FE	Yes	Yes	Yes
Ethnicity dummies	No	Yes	Yes
Other controls	No	No	Yes
Observations	4730	4730	4730

The regressions test whether female plaintiffs/defendants are more likely to be matched with female judges than male judges.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 1.
Sample is restricted to the years 1976-2012
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = Plaintiffs, Def. = defendants, maj. = majority.

Table D3: Gender randomization checks, 2011 and after

	(1) Judge maj. female	(2) Judge maj. female	(3) Judge maj. female
Pla. maj. female	0.00844	0.00866	0.00512
	(0.00961)	(0.00945)	(0.00733)
Def. maj. female	0.00654	0.00600	0.00301
	(0.00837)	(0.00826)	(0.00713)
Court-year FE	Yes	Yes	Yes
Ethnicity dummies	No	Yes	Yes
Other controls	No	No	Yes
Observations	15707	15707	15707

The regressions test whether female plaintiffs/defendants are more likely to be matched with female judges than male judges.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 1.
Sample is restricted to the years 2011-2020
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for defendants, plaintiffs, and judges.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = Plaintiffs, Def. = defendants, maj. = majority.

Table D4: Ethnicity randomization checks 1

	(1) Judge plur. Kalenjin	(2) Judge plur. Kamba	(3) Judge plur. Kikuyu
Pla. plur. Kalenjin	0.00715		
	(0.00878)		
Def. plur. Kalenjin	-0.00898		
	(0.0102)		
Pla. plur. Kamba		-0.00818	
		(0.00649)	
Def. plur. Kamba		0.00360	
		(0.00779)	
Pla. plur. Kikuyu			0.00592
			(0.00762)
Def. plur. Kikuyu			0.00143
			(0.00757)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	14630	14630	14980

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D5: Ethnicity randomization checks 2

	(1) Judge plur. Kisii	(2) Judge plur. Luhya	(3) Judge plur. Luo
Pla. plur. Kisii	-0.0112		
	(0.00762)		
Def. plur. Kisii	0.00180		
	(0.00794)		
Pla. plur. Luhya		0.00536	
		(0.00673)	
Def. plur. Luhya		0.0110*	
		(0.00580)	
Pla. plur. Luo			0.00137
			(0.00886)
Def. plur. Luo			0.00226
			(0.0106)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	14980	14980	14980

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D6: Ethnicity randomization checks 3

	(1) Judge plur. Masai	(2) Judge plur. Meru	(3) Judge plur. Mijikenda
Pla. plur. Masai	-0.000643		
	(0.00216)		
Def. plur. Masai	0.000342		
	(0.000784)		
Pla. plur. Meru		0.00208	
		(0.00314)	
Def. plur. Meru		-0.000833	
		(0.00272)	
Pla. plur. Mijikenda			-0.000549
			(0.00304)
Def. maj. Mijikenda			0.00378
			(0.00409)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	14980	14980	14980

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D7: Ethnicity randomization checks 4

	(1) Judge plur. Pokot	(2) Judge plur. Somali	(3) Judge plur. Turkana
Pla. plur. Pokot	-0.00447		
	(0.00454)		
Def. plur. Pokot	0.00322		
	(0.00825)		
Pla. plur. Somali		0.00406	
		(0.00536)	
Def. plur. Somali		-0.00676	
		(0.00518)	
Pla. plur. Turkana			0.000207
			(0.000204)
Def. plur. Turkana			-0.000961
			(0.00105)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	14980	14980	14980

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D8: Ethnicity randomization checks, before 2011, 1

	(1) Judge plur. Kalenjin	(2) Judge plur. Kamba	(3) Judge plur. Kikuyu
Pla. plur. Kalenjin	0.000471		
	(0.00107)		
Def. plur. Kalenjin	-0.000648		
	(0.000854)		
Pla. plur. Kamba		-0.0493***	
		(0.0187)	
Def. plur. Kamba		0.0696***	
		(0.0238)	
Pla. plur. Kikuyu			0.00559
			(0.0113)
Def. plur. Kikuyu			-0.00915
			(0.0101)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	3114	3114	3114

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 1976-2012.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D9: Ethnicity randomization checks, before 2011, 2

	(1) Judge plur. Kisii	(2) Judge plur. Luhya	(3) Judge plur. Luo
Pla. plur. Kisii	-0.0325**		
	(0.0157)		
Def. plur. Kisii	-0.0148		
	(0.0159)		
Pla. plur. Luhya		0.0143	
		(0.0285)	
Def. plur. Luhya		-0.00326	
		(0.0186)	
Pla. plur. Luo			0.0197
			(0.0265)
Def. plur. Luo			0.00369
			(0.0275)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	3114	3114	3114

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 1976-2012.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D10: Ethnicity randomization checks, before 2011, 3

	(1) Judge plur. Masai	(2) Judge plur. Meru	(3) Judge plur. Mijikenda
Pla. plur. Masai	0.000130		
	(0.000467)		
Def. plur. Masai	0.000269		
	(0.000410)		
Pla. plur. Meru		0.00369	
		(0.00918)	
Def. plur. Meru		-0.0127	
		(0.00922)	
Pla. plur. Mijikenda			-0.0129
			(0.0129)
Def. maj. Mijikenda			0.00115
			(0.00345)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	3114	3114	3114

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 1976-2012.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D11: Ethnicity randomization checks, before 2011, 4

	(1) Judge plur. Pokot	(2) Judge plur. Somali	(3) Judge plur. Turkana
Pla. plur. Pokot	-0.00991		
	(0.00691)		
Def. plur. Pokot	0.0207		
	(0.0671)		
Pla. plur. Somali		0.0139	
		(0.0212)	
Def. plur. Somali		-0.0310	
		(0.0227)	
Pla. plur. Turkana			0
			(.)
Def. plur. Turkana			0
			(.)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	3114	3114	3114

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
Results are absent for Turkana due to collinearity, driven by a lack of Turkana observations in the period.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 1976-2012.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D12: Ethnicity randomization checks, 2011 and after, 1

	(1) Judge plur. Kalenjin	(2) Judge plur. Kamba	(3) Judge plur. Kikuyu
Pla. plur. Kalenjin	0.00823		
	(0.0103)		
Def. plur. Kalenjin	-0.0111		
	(0.0123)		
Pla. plur. Kamba		-0.000919	
		(0.00812)	
Def. plur. Kamba		-0.00952	
		(0.00844)	
Pla. plur. Kikuyu			0.00659
			(0.00994)
Def. plur. Kikuyu			0.00415
			(0.00966)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	11866	11866	11866

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 2011-2020.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D13: Ethnicity randomization checks, 2011 and after, 2

	(1) Judge plur. Kisii	(2) Judge plur. Luhya	(3) Judge plur. Luo
Pla. plur. Kisii	-0.00496		
	(0.00758)		
Def. plur. Kisii	0.00834		
	(0.00983)		
Pla. plur. Luhya		0.00495	
		(0.00647)	
Def. plur. Luhya		0.0140**	
		(0.00594)	
Pla. plur. Luo			-0.00466
			(0.00935)
Def. plur. Luo			0.00267
			(0.0106)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	11866	11866	11866

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 2011-2020.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D14: Ethnicity randomization checks, 2011 and after, 3

	(1) Judge plur. Masai	(2) Judge plur. Meru	(3) Judge plur. Mijikenda
Pla. plur. Masai	-0.00108		
	(0.00259)		
Def. plur. Masai	0.000606		
	(0.000933)		
Pla. plur. Meru		0.00177	
		(0.00265)	
Def. plur. Meru		0.00219	
		(0.00291)	
Pla. plur. Mijikenda			0.00202
			(0.00255)
Def. maj. Mijikenda			0.00515
			(0.00483)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	11866	11866	11866

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 2011-2020.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Table D15: Ethnicity randomization checks, 2011 and after, 4

	(1) Judge plur. Pokot	(2) Judge plur. Somali	(3) Judge plur. Turkana
Pla. plur. Pokot	-0.00222		
	(0.00278)		
Def. plur. Pokot	-0.000538		
	(0.00171)		
Pla. plur. Somali		0.00154	
		(0.00376)	
Def. plur. Somali		-0.00122	
		(0.00241)	
Pla. plur. Turkana			0.000200
			(0.000225)
Def. plur. Turkana			-0.00112
			(0.00122)
Court-year FE	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Observations	11866	11866	11866

The regressions test whether female plaintiffs/defendants are more likely to be matched with judges of their own ethnicity than judges of other ethnicities.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 2.
Sample is restricted to the years 2011-2020.
Other controls include case type dummies, a dummy for an appeal case, and variables for the numbers of defendants, plaintiffs, and judges.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Appendix E: Results without ethnicities with significant coefficients in the balance tests

Table E1: Ethnicity results, no Kamba or Luhya

	(1) Def. win	(2) Def. win	(3) Def. win	(4) Def. win	(5) Def. win
Judge-pla. same	0.0102		-0.00508	0.00127	0.00257
	(0.0151)		(0.0171)	(0.0208)	(0.0211)
Judge-def. same		0.0434***	0.0523***	0.0483**	0.0512**
		(0.0148)	(0.0182)	(0.0238)	(0.0237)
Court-year FE	Yes	Yes	Yes	Yes	Yes
Ethnicity dummies	No	No	No	Yes	Yes
Other controls	No	No	No	No	Yes
Observations	11445	11048	9773	9773	9773

The regressions test whether defendants (plaintiffs) are more likely to win (lose) if they have the same (a different) plurality ethnicity as judges.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equation 5.
Judge-pla. same and Judge-def. same refer to similarity in plurality ethnicity.
Sample is restricted to cases without Luhya or Kamba judges or litigants.
Ethnicity dummies include binary variables indicating whether a given ethnicity is the plurality, one for each ethnicity, for both defendants and plaintifs.
Other controls include case type dummies; a dummy for an appeal case; variables for the numbers of defendants, plaintiffs, and judges; and dummies for defendant, plaintiff, and judge majority gender.
To prevent a loss of observations, all categorical controls (such as case type) include a dummy that denotes if data is missing/unknown.
Pla. = plaintiff, def. = defendant.

Appendix F: Effect o putting biased judges on panels

Table F1: Results for significantly in-group gender biased judges, off and on panels

	(1) Def. win	(2) Def. win
Judge maj. female	-0.189***	0.129
	(0.0449)	(0.159)
Def. maj. female	-0.220***	-0.0512
	(0.0520)	(0.0795)
Judge maj. fem. X def. maj. fem.	0.394***	0.0777
	(0.0605)	(0.167)
Court-year FE	Yes	Yes
Individual decisions	Yes	No
Panel decisions	No	Yes
Observations	1789	203

Sample is restricted to the 14 judges with significant gender in-group bias coefficients for defendants in individual regression Column 1 includes only cases where the judges ruled individuall.
Column 2 includes only cases where they ruled on panels.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model. For specification details, see equations 3 and 4.
Pla. = plaintiff, def. = defendant, maj. = majority.

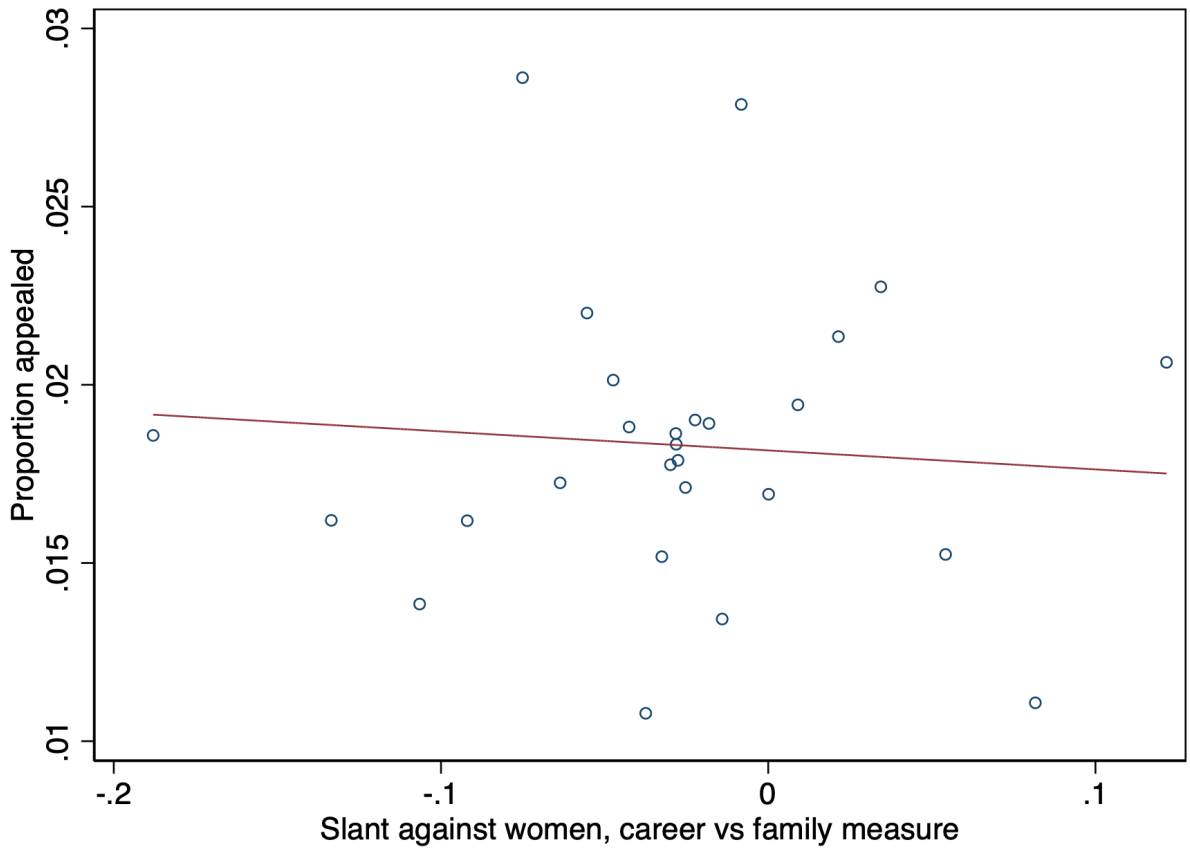
Appendix G: Relationship between slant and appeals, and slant and reverals

Table G1: Appeals and slant, family vs career

	(1) appealed	(2) appealed	(3) appealed	(4) appealed	(5) appealed
Slant against women, career vs family	-0.00534	-0.0213	0.0185	0.00408	-0.0650**
	(0.0154)	(0.0181)	(0.0249)	(0.0159)	(0.0266)
Def. maj. female		0.00190	0.00101		
		(0.00277)	(0.00430)		
Pla. maj. female				0.00316	0.00350
				(0.00278)	(0.00447)
Def. maj. fem. X Slant against women		0.0301	-0.00759		
		(0.0246)	(0.0387)		
Pla. maj. fem. X Slant against women				-0.00570	-0.0115
				(0.0221)	(0.0408)
Court-year FE	Yes	Yes	Yes	Yes	Yes
Restricted sample	No	No	Yes	No	Yes
Observations	26177	20828	11573	23587	9787

The regressions test whether slanted judges are more likely to have case decisions appealed.
The coefficients of interest are 'Slant against women, career vs family' and the interactions.
Column 3 (5) restricts the sample to cases where the defendant (plaintiff) loses,
and the interaction tests if reverals are more likely if the judges is more slanted and the
defendant (plaintiff) is female. These cases have the most potentail for gender bias.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Figure G1: Relationship between judge slant against women (career vs family measure) and appeals



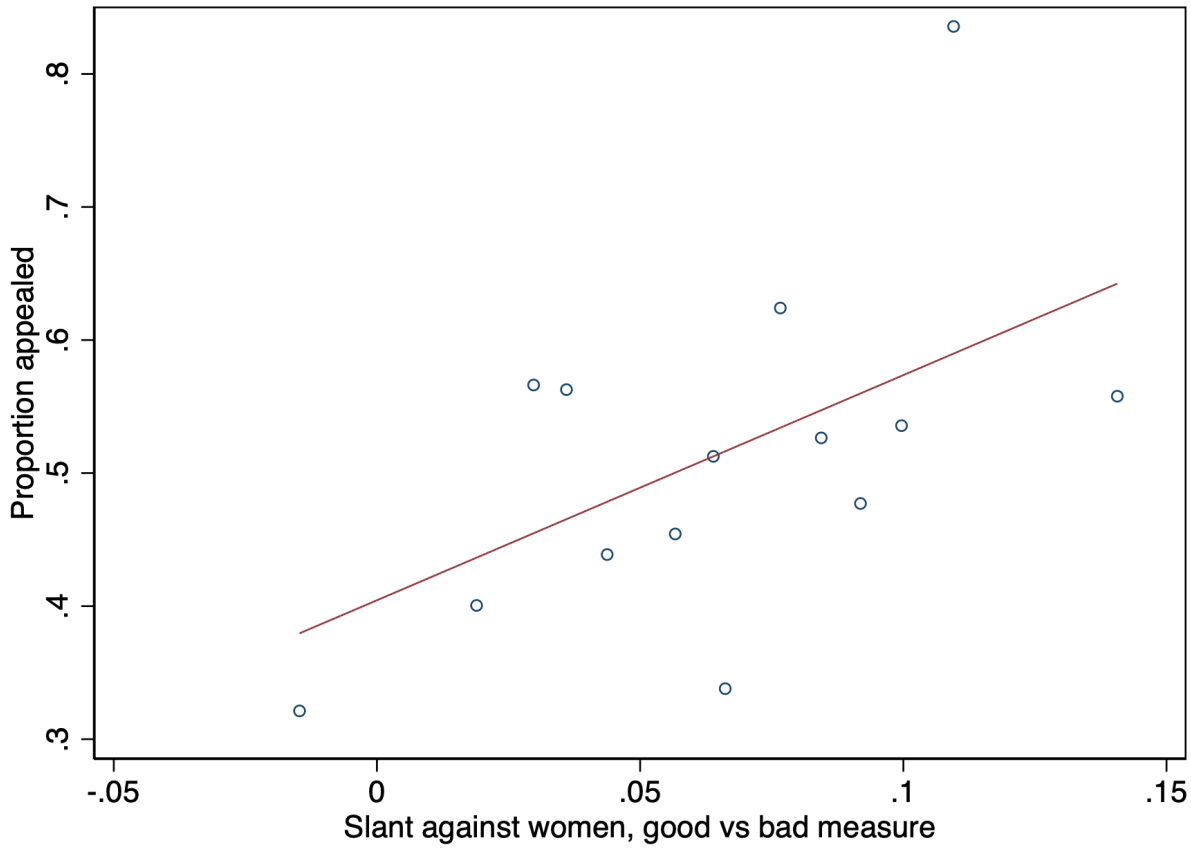
Data points are binned and account for court-year fixed effects.

Table G2: Appealed and slant, good vs bad

	(1) appealed	(2) appealed	(3) appealed	(4) appealed	(5) appealed
Slant against women, good vs bad	0.0417*	0.0396	0.0201	0.0477	0.0292
	(0.0251)	(0.0347)	(0.0417)	(0.0332)	(0.0590)
Def. maj. female		0.00509	0.00481		
		(0.00429)	(0.00714)		
Pla. maj. female				0.00457	0.00679
				(0.00535)	(0.00838)
Def. maj. fem. X Slant against women		-0.0375	-0.0340		
		(0.0513)	(0.0709)		
Pla. maj. fem. X Slant against women				-0.0178	-0.0466
				(0.0627)	(0.0901)
Court-year FE	Yes	Yes	Yes	Yes	Yes
Restricted sample	No	No	Yes	No	Yes
Observations	22100	17391	9705	19884	8149

The regressions test whether slanted judges are more likely to have case decisions appealed.
The coefficients of interest are 'Slant against women, good vs bad' and the interactions.
Column 3 (5) restricts the sample to cases where the defendant (plaintiff) loses,
and the interaction tests if reverals are more likely if the judges is more slanted and the
defendant (plaintiff) is female. These cases have the most potentail for gender bias.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Figure G2: Relationship between judge slant against women (good vs bad measure) and appeals



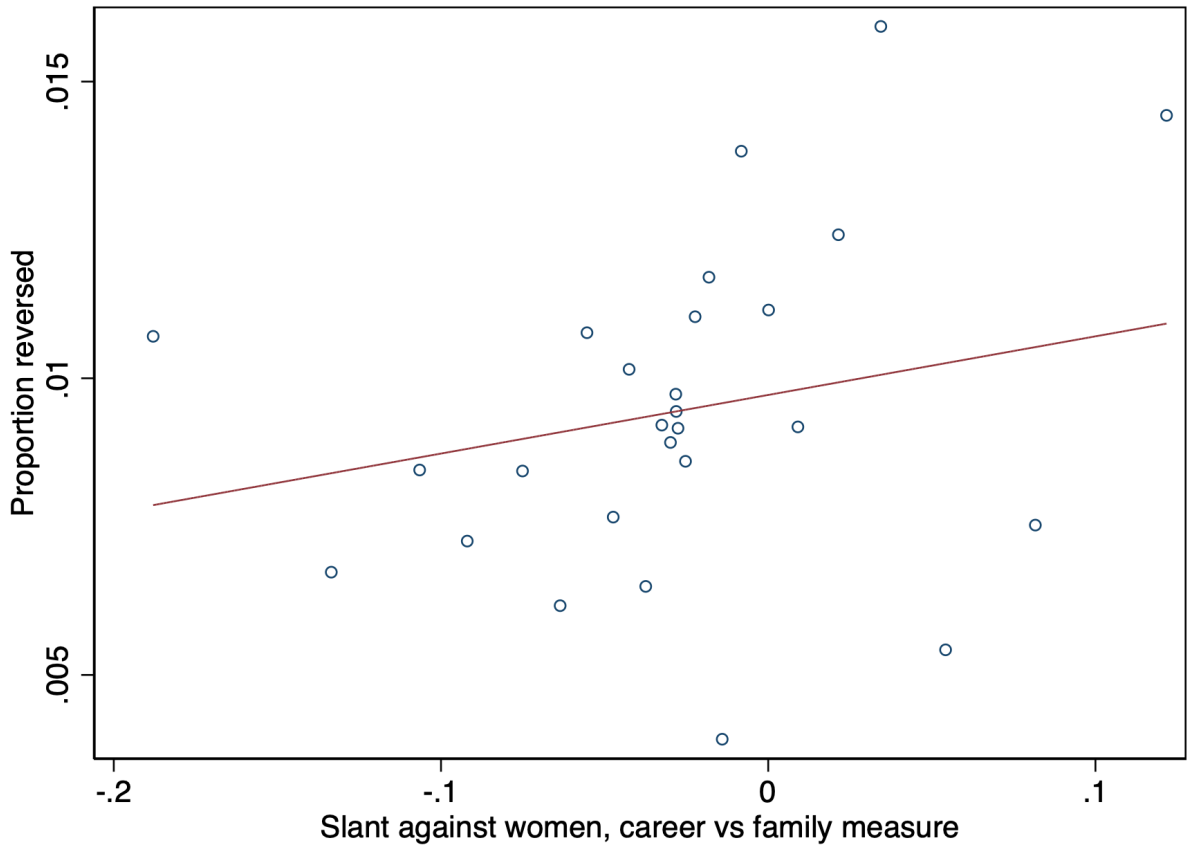
Data points are binned and account for court-year fixed effects.

Table G3: Reversals and slant, family vs career

	(1) reversed	(2) reversed	(3) reversed	(4) reversed	(5) reversed
Slant against women, career vs family	0.00988	0.00516	0.0362**	0.0145	-0.0370**
	(0.00953)	(0.0120)	(0.0166)	(0.0109)	(0.0169)
Def. maj. female		0.000552	0.000912		
		(0.00212)	(0.00342)		
Pla. maj. female				0.000915	0.00253
				(0.00185)	(0.00252)
Def. maj. fem. X Slant against women		0.0215	0.00782		
		(0.0199)	(0.0315)		
Pla. maj. fem. X Slant against women				-0.00438	0.00310
				(0.0142)	(0.0214)
Court-year FE	Yes	Yes	Yes	Yes	Yes
Restricted sample	No	No	Yes	No	Yes
Observations	26177	20828	11573	23587	9787

The regressions test whether slanted judges are more likely to have case decisions reversed.
The coefficients of interest are 'Slant against women, career vs family' and the interactions.
Column 3 (5) restricts the sample to cases where the defendant (plaintiff) loses,
and the interaction tests if reverals are more likely if the judges is more slanted and the
defendant (plaintiff) is female. These cases have the most potentail for gender bias.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Figure G3: Relationship between judge slant against women (career vs family measure) and reversals



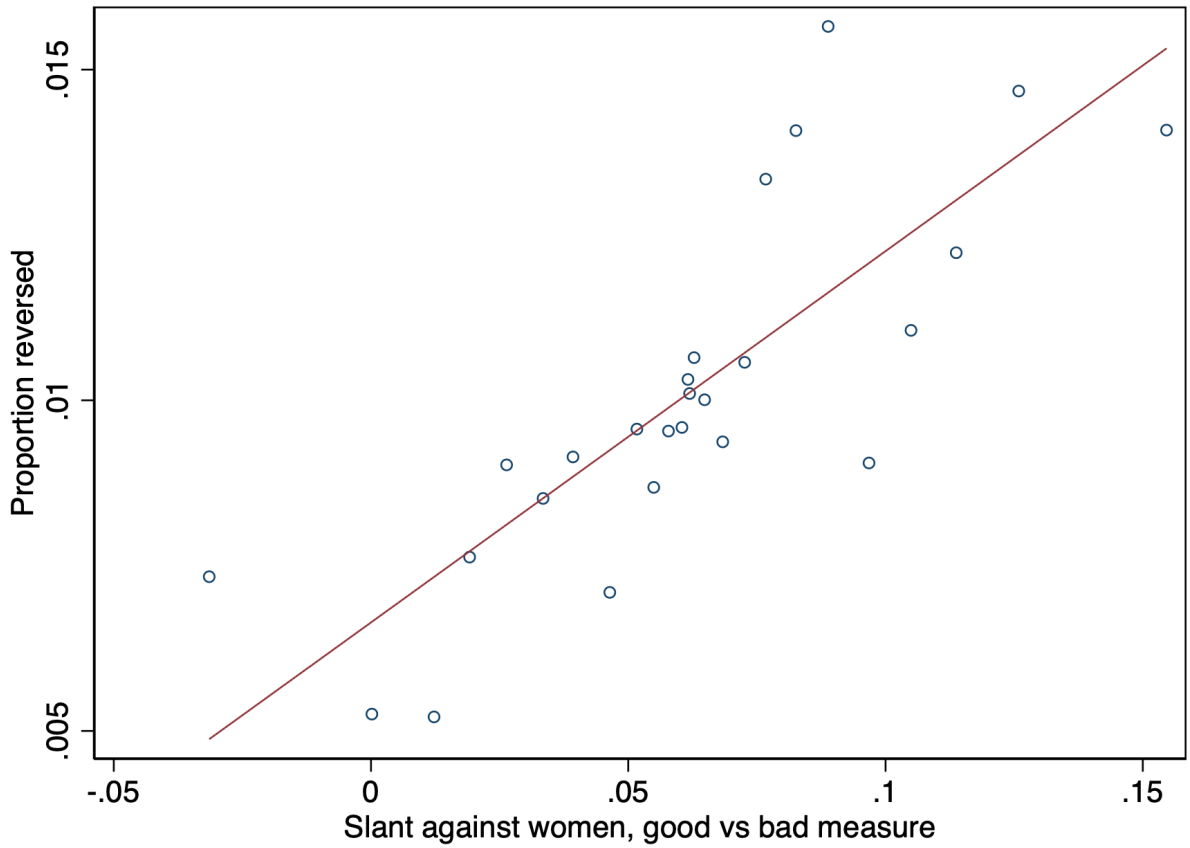
Data points are binned and account for court-year fixed effects.

Table G4: Reversals and slant, good vs bad

	(1) reversed	(2) reversed	(3) reversed	(4) reversed	(5) reversed
Slant against women, good vs bad	0.0561***	0.0498**	0.0461	0.0642***	0.0417
	(0.0167)	(0.0240)	(0.0352)	(0.0211)	(0.0262)
Def. maj. female		0.00126	0.00151		
		(0.00239)	(0.00427)		
Pla. maj. female				0.00317	0.00508
				(0.00365)	(0.00556)
Def. maj. fem. X Slant against women		-0.00398	-0.00261		
		(0.0305)	(0.0445)		
Pla. maj. fem. X Slant against women				-0.0317	-0.0394
				(0.0417)	(0.0622)
Court-year FE	Yes	Yes	Yes	Yes	Yes
Restricted sample	No	No	Yes	No	Yes
Observations	22100	17391	9705	19884	8149

The regressions test whether slanted judges are more likely to have case decisions reversed.
The coefficients of interest are 'Slant against women, good vs bad' and the interactions.
Column 3 (5) restricts the sample to cases where the defendant (plaintiff) loses,
and the interaction tests if reverals are more likely if the judges is more slanted and the
defendant (plaintiff) is female. These cases have the most potentail for gender bias.
Standard errors, in parentheses, are clustered at the judge level.
All columns are based on a linear regression model.
Pla. = plaintiffs, def. = defendants, plur. = plurality, maj. = majority.

Figure G4: Relationship between judge slant against women (good vs bad measure) and reversals



Data points are binned and account for court-year fixed effects.

Appendix H: Interaction between gender and ethnicity in-groups

Table H1: Ethnicity and gender interaction results

	(1) Def. win	(2) Def. win	(3) Def. win
Judge-def. same gender	0.0203**		0.0171
	(0.00989)		(0.0108)
Judge-def. same ethnicity	0.0452***		0.0631***
	(0.0174)		(0.0188)
Judge-def. same gender X Judge-def. same ethnicity	-0.0169		-0.0255
	(0.0224)		(0.0236)
Judge-pla. same gender		0.0196**	0.0146
		(0.00909)	(0.0119)
Judge-pla. same ethnicity		0.0264	0.00896
		(0.0170)	(0.0176)
Judge-pla. same gender X Judge-pla. same ethnicity		-0.0253	-0.0252
		(0.0241)	(0.0290)
Court-year FE	Yes	Yes	Yes
Observations	17744	19277	14613

Standard errors, in parentheses, are clustered at the judge level.

All columns are based on a linear regression model. For specification details, see equation 5.

Pla. = plaintiff, def. = defendant, maj. = majority.