Gender diversity and teams' decision-making

Maddalena Ronchi Viola Salvestrini*

January 31, 2025

Abstract

We investigate how gender composition of teams affects the decision-making process and the quality of decisions, using a novel database on the universe of collegial rulings from a sample of first- and second-instance criminal courts in Italy. Leveraging the quasi-random allocation of both judges and cases to judicial teams, we document three main findings. First, all-women teams have higher conviction rates than both mixed-gender and all-men teams. Second, all-men teams converge faster to a decision than any other team composition. Third, decisions made by all-men teams are more likely to be incorrect, as indicated by a higher likelihood of being appealed and subsequently overturned, whereas no significant difference in decision quality is observed between mixed-gender and all-women teams. To explain these results, we propose a model of diagnostic skills in which individuals differ in both preferences and ability. The observed relationship between conviction and overturn rates, together with the analysis of written judicial opinions, supports the conclusion that variations in effort and skills, rather than preferences, drive our results.

JEL codes: J16, J82, M5, K4

*Ronchi: Northwestern University, IZA, and CESifo. Salvestrini: Bocconi University, Queen Mary University of London and IZA. We thank François Gerard and Anna Raute for their support and guidance. For very helpful comments and suggestions we thank Jan Bakker, Barbara Petrongolo, Wouter Dessein, Adriano De Falco, Alfonso Gambardella, Elia Moracci, Andrea Ichino, Mikkel Mertz, Martina Zanella, as well as seminar participants at Bocconi University, European University Institute, Queen Mary University, and conference participants at the The Workplace of the Future Conference; CSEF- IGIER Symposium on Economics and Institutions; Organizational Economics Summer Symposium (OESS); BSE Summer Forum 2024; IEB Workshop on Public Policies; EAYE Annual Meeting 2024; EEA Congress 2023. We gratefully acknowledge support from Einaidi Institute for Economic and Finance and Institute for Human Studies.

1 Introduction

Despite the narrowing gap in labor force participation between men and women, occupational segregation by gender persists (Bertrand, 2020). This has significant implications. At the individual level, occupational segregation contributes to gender disparities in earnings and career trajectories (Blau and Kahn, 2017; Goldin, 2014). At the aggregate level, barriers preventing gender-balanced entry into professions can distort the allocation of talent in the economy and impose substantial welfare costs, especially when individuals have occupation-specific skills (Hsieh et al., 2019; Ashraf et al., 2022). An additional and unexplored consequence of gender-segregated occupations relates to the potential impact of the gender composition of the workforce on team dynamics and team decision-making. This is an important margin, as teams have become increasingly central to decision-making processes in many organizations (Hamilton et al., 2003; Lazear and Shaw, 2007).[1]

This paper investigates whether and how the gender composition of a team affects its decision-making process and the quality of its decisions. From a theoretical perspective, the impact of diversity on team dynamics is ambiguous. On the one hand, it can improve the team performance (Hamilton et al., 2003, 2012), especially when there are complementarities in skills, knowledge or perspectives (Lazear, 1999; Prat, 2002). Moreover, team diversity can in principle change the way teams process the information needed to make decisions, ultimately altering the behavior of a group in ways that can lead to improved and more accurate thinking and to reduced social conformity (Halberstam and Knight, 2016; Carrell et al., 2019). On the other hand, team diversity can raise coordination and communication costs (Lazear, 1999; Dessein and Santos, 2006; Hjort, 2014). Since women and men have been shown to differ in terms of preferences, attitudes, and social values (Andreoni and Vesterlund, 2001; Chattopadhyay and Duflo, 2004; Croson and Gneezy, 2009), teams might have different work dynamics depending on their gender composition, and the effect might not be symmetrical (Apesteguia et al., 2012). Understanding the interaction between team composition and team effectiveness is crucial, as the persistence of gender segregation not only raises normative concerns but also potential implications for team productivity and performance. Yet, empirical evidence on the effect of gender diversity on

[1] In the U.S., 80 percent of employment is in occupations where group work is *very* or *extremely* relevant (O*NET, 2022).

decision-making in teams remains scarce.

Despite the crucial role gender diversity can play in group dynamics, two main factors have limited our ability to empirically assess whether and how the gender composition of a team affects its performance. First, team composition is potentially endogenous to the characteristics of both the organization the team belongs to and its members. Second, defining and measuring the intrinsic quality of team decisions is challenging in most real-life settings. We overcome these challenges by focusing on the Italian criminal justice system, which allows us to exploit the quasi-random allocation of judges and cases to judicial panels – teams of three judges making decisions collectively. We leverage exogenous variation in gender diversity across judicial panels within courts to answer three main questions. First, by investigating whether similar offenses are ruled against differently depending on the gender composition of the judicial panel, we can assess whether teams make different decisions according to their gender composition, and whether the effects are symmetric between homogeneously female or male teams. Second, we investigate whether a team's gender composition affects the decision-making time as proxied by trial duration. Finally, the judicial setting presents us with the unique opportunity to measure decision quality by studying whether the final ruling of a judicial panel is a) appealed against and b) confirmed or overturned in subsequent courts.[2]

We collect a novel database containing the universe of collegial ruling sentences from a sample of first-instance criminal courts in Italy, which we then merge with the corresponding second-instance sentences. Specifically, we obtain data from four courts in Central and Southern Italy and the corresponding Courts of Appeal.[3] Sentences have been made available by the presidents of each court in paper format. We digitized and processed them using a series of OCR and text-parsing algorithms. To our knowledge, such a comprehensive database has never been assembled before due to the lack of digitized and accessible data at the court level, making this a significant contribution of this paper.

[2]This strategy has been used by Alesina and La Ferrara (2014) to assess the presence of racial bias of judges in capital sentencing in the US, and by Coviello et al. (2019) to study the productivity effect of labor specialization.

[3]We have data on the court of Naples, Naples Nord, and Torre Annunziata (under the Court of Appeal of Naples) and the court of Florence (under the Court of Appeal of Florence). These courts are different both in terms of size (with Naples being the largest court in Italy, Florence being a medium-sized court, and Naples Nord and Torre Annunziata being smaller courts) and geographical location (with Florence being in the center-north and the other three in the South). This diversity reassures us that our results are not specific to a particular type of court or geographical area.

To isolate the causal effect of the gender composition of the judicial panels on their decision-making, we leverage quasi-random variation in gender composition *across* judicial panels *within* a court. Judges within a court are assigned to panels and cases in an effectively random manner, allowing us to compare teams who work in the same environment and on similar crimes but differ in gender composition. The identifying assumption is that judges cannot decide which panel to sit in and what cases to work on. While the structure of the Italian judiciary allows us to confidently rule out both that judges' allocation across panels is a function of their gender, and that the cases assigned to each collegial panel depend on the gender composition of the group, we present formal tests supporting the assumption of exogeneity of assignment of both judges and cases to panels. Moreover, we perform several checks to ensure that our results reflect the impact of gender differences rather than the influence of other judicial characteristics that might correlate with gender, such as age and experience or geographical mobility.

We present three key sets of results. First, we document that judicial panels ruling on similar cases reach different sentencing outcomes depending on their gender composition. Specifically, mixed-gender panels tend to be more lenient than same-gender panels. The effect appears to be driven by all-women panels, which are 10.2 percentage points and 7.7 percentage points (21.7 percent and 16.3 percent) more likely to convict the defendant than mixed-gender teams and all-men teams, respectively. However, conditional on conviction, all-women teams impose sentences that are, on average, 12 months shorter than those given by all-men teams.

Second, we show that all-men teams reach decisions more quickly than mixed-gender and all-women teams. Specifically, trials overseen by all-men teams last approximately four months less, corresponding to a 15 percent decrease relative to the mean trial duration. Given that shorter trials are desirable for reducing court backlogs (Coviello et al., 2019), fast decision-making would be beneficial if these rulings are not appealed at higher rates. Otherwise, there would be a trade-off between quicker decisions and increased costs associated with appeal trials.

Finally, we exploit the hierarchical structure of criminal courts to estimate whether a team's gender composition influences the quality of its decisions. We start by examining the *perceived quality* of a ruling, proxied by its likelihood of being appealed. Our analysis shows that rulings by all-men panels are more likely to be appealed, particularly those involving convictions. We then measure *effective quality* by determining

whether a ruling is confirmed or reversed by the appellate court. While all-women teams do not differ significantly from mixed-gender teams, all-men teams are 20 percentage points (26 percent) less likely to make decisions that withstand appeal.[4]

These patterns can be rationalized with a stylized framework of diagnostic decisions, following the model proposed by Chan et al. (2022). Teams evaluate cases and form a signal of the defendant's true guilt, and decide whether to convict or acquit based on this signal. Mistakes are costly and asymmetric, with the cost of false positives (convicting an innocent person) being higher than false negatives (acquitting a guilty person).[5] More precise signals lead to higher conviction rates and fewer errors, as the expected cost of a false positive decreases, while imprecise signals result in lower conviction rates and more errors. Signal precision depends on the team's diagnostic skill, which is a function of effort and relates to both the time spent analyzing information and the time needed for consensus-building. When effort is low or consensus is difficult to achieve, conviction rates decrease, as teams rule guilty only when convinced beyond a reasonable doubt. Mixed-gender teams are slower and have lower conviction rates, possibly due to the challenges of reaching an agreement amidst diverse viewpoints. However, they are not more frequently appealed or overturned, suggesting that the internal debates and diversity of perspectives lead to robust decisions when consensus is achieved. All-men teams are faster but face higher appeal and overturn rates, indicating quicker but more error-prone decisions. In contrast, all-women teams, while slower, have higher conviction rates and are not more frequently appealed or overturned, suggesting thorough and careful decision-making with fewer mistakes.

In the final part of the paper, we investigate the plausible mechanisms through which gender composition might affect team dynamics. The lack of a positive relationship between conviction rates and false positives suggests that differences in effort, rather than preferences, drive our findings. In line with this, we find that gender composition matters in complex decision-making only, suggesting that gender-based differences become prominent precisely when extracting a precise signal is more chal-

[4]In Section 5.4 we show that our findings are not driven by differences in other dimensions that correlate with gender. nor by particular types of crimes where we might expect gender to play a role. We also show that our results are not driven by the share of women in the teams, as we find no difference between teams with one or two women.

[5]In criminal law, the *Blackstone's ratio* principle (also known as Blackstone's formulation) is the idea that *it is better that ten guilty persons escape than that one innocent suffer*, as expressed by the English jurist William Blackstone in his seminal work Commentaries on the Laws of England, published in the 1760s.

lenging. Moreover, we document how writing styles differ among different teams. All-women and mixed-gender teams tend to issue shorter opinions but spend more time writing them, suggesting greater effort in ensuring clarity and effective communication. In contrast, all-male teams produce significantly longer judgments – more than twice the length of others – and heavily rely on Court of Cassation maxims – more than twice as often as both mixed-gender and all-female teams. Finally, previous interaction with colleagues of the opposite gender appears to improve decision-making in all-men teams, indicating that diversity in workforce composition can enhance the performance of teams who are underperforming at baseline.

This paper contributes to three stands of the literature. First, it relates to the growing body of evidence on the effects of gender diversity in decision-making bodies (Dufwenberg and Muren, 2006; Hoogendoorn et al., 2013; Born et al., 2020; Beaman et al., 2009; Ahern and Dittmar, 2012). We add to this literature by focusing on a non-experimental setting with high-stakes decisions and where the women's presence was not mandated by law (e.g., through quotas). Given that it is still debated whether mandated participation of women in decision-making bodies leads to better group dynamics and higher team effectiveness, providing evidence of the effect of gender composition on team decision-making and performance in a setting where gender diversity arises naturally can shed further light on this topic. The studies more closely related to ours are Bagues and Esteve-Volart (2010) and Bagues et al. (2017), which leverage quasi-random variation in the gender composition of recruiting committees inside universities and the judiciary to study whether female evaluators are more favorable towards female candidates in hiring decisions. Differently from these studies, we do not focus on the effects of same-gender interactions in an employer-employee setting on women's career trajectories. Instead, we examine whether the gender composition of teams affects the decision-making process itself. Specifically, our analysis pertains to the effects of gender composition on the type and quality of decisions, rather than on its effects on under-represented groups. This allows us to provide new insights into how gender diversity influences high-stakes decision-making dynamics, highlighting the broader implications of gender composition for team performance beyond hiring and career advancement.

Second, we contribute to the literature on gender segregation across occupations (Goldin, 2014; Blau and Kahn, 2017). Building on recent work by Hsieh et al. (2019), which demonstrates that convergence in occupational distribution across gender and

ethnicity positively impacts economic growth, our paper provides novel evidence on the effect of gender diversity within a truly gender-mixed occupation. Indeed, while the Italian judiciary was a male-only occupation until 1965, the number of female judges has increased substantially and represent now 55 percent of the workforce.[6] This is particularly important because it sheds light on the potential efficiency losses associated with occupational segregation, not only in instances where women are underrepresented but also in those where men are underrepresented (Schaede and Mankki, 2022; Delfino, 2024).

Finally, this paper relates to the literature on decision-making in the legal system. Most prior research has focused on documenting biases in decision-making based on the demographic characteristics of judges (or juries) and defendants, such as gender, ethnicity, or religion (see among others Anwar et al., 2012; Alesina and La Ferrara, 2014; Knepper, 2018; Arnold et al., 2018; Anwar et al., 2019; Ash et al., 2023). Rather than examining the behavior of individual judges, our study investigates how the gender composition of judicial panels influences their collective decision-making processes and outcomes. Furthermore, we go beyond exploring the effects of similarities between judges and defendants in observable characteristics, to uncover the broader impacts of increasing diversity within the judiciary on both the nature and quality of their rulings.

The remainder of this paper is organized as follows. Section 2 describes the institutional setting. Section 3 outlines the data and the main outcomes of interest. Section 4 presents the empirical strategy and discusses its validity. Section 5 presents the results. Section 6 propose a theoretical framework to interpret our results and examines the plausible mechanisms underlying our findings. Finally, section 7 concludes.

2 Institutional setting

The Italian judicial system is organized in a hierarchy of courts (see Figure A1). Our analysis focuses on the criminal courts (*Tribunali*), therefore excluding civil cases and criminal cases handled by lower courts (*Giudici di pace*, dealing with minor civil and criminal matters) and the court of assize (*Corte di Assise*, handling the most serious crimes such as terrorism and aggravated murder). The criminal courts are presided

[6]Donne in Magistratura, CSM `https://www.csm.it/web/csm-internet/statistiche/analisi-e-studi`

over by ordinary first-instance judges, and consist of either a single judge for less complex matters or a panel of three judges for more serious cases.[7] Our study focuses on decisions made by these judicial panels.

The judgments delivered by the courts of first instance can be appealed against and be reviewed on factual grounds and the interpretation of the law by courts of second-instance, also known as courts of appeal (*Corti di Appello*). These are composed of panels of three judges who are different from and do not work with those who ruled in the first instance. The highest court in the judicial hierarchy is the Supreme Court of Cassation (*Corte Suprema di Cassazione*), which serves as the highest appellate court and reviews the judgments from the courts of appeal. The Supreme Court of Cassation is composed of a panel of five judges, who can uphold, modify, or quash the appealed sentence. Reviews by the Supreme Court must concern only the correct application of law, not factual matters. While the first and second-instance courts are located across various districts in Italy and operate under territorial jurisdiction rules, the Supreme Court is based in Rome. In this paper, we will focus on the first round of appeals only.

There are two factors that make the Italian judicial context a particularly interesting setting to study. First, Italy has witnessed a dramatic transformation in the gender composition of its workforce. Figure A2 illustrates the shift in gender composition among new judges from 1965 to 2020. Until 1965, the profession was legally reserved for men, but since then, there has been a steep rise in the share of female judges. Today, women constitute almost 60 percent of ordinary judges. This evolution allows us to observe the effects of both all-female and all-male teams within the same profession. Second, Italy's judicial system is characterized by a quasi-random allocation of both judges and cases into judicial panels, providing a natural experiment to study the effects of gender composition on decision-making. In what follows, we review the rules regulating the assignment of judges and cases to panels in the Italian judiciary.

[7]The offenses under the jurisdiction of collegial panels are regulated by law (Art. 33 bis c.p.p.) and include crimes with imprisonment terms between 10 and 24 years, as well as crimes of significant criminal relevance or social alarm not under the competence of the Court of Assize. These include, but are not limited to, mafia-type crimes and associations; terrorism; arms trafficking; criminal associations; usury; sexual violence; prostitution; racial, ethnic, and religious discrimination; and crimes against public safety and order.

2.1 Random assignment of judges and cases

Assignment of judges to courts and panels. All Italian judges hold a law degree and are appointed through a public examination covering all legal subjects and procedural rules, resulting in a ranking. When court vacancies are published, newly appointed judges choose their destination court based on their exam scores: the first-ranked chooses first, the second-ranked chooses second, and so on.[8]

Within each court, judges are randomly assigned to judicial panels in an effectively random manner. This presumption follows from an analysis of the courts' institutional practices. While assignment practices vary across courts – some employing explicitly random methods, such as lotteries or random number generators, and others incorporating additional practical factors like judge availability – they must adhere to the requirement for objectivity, predetermination, and randomness as stipulated by law (Art. 25 of the Italian Constitution). Such criteria for assigning judges to panels are approved and reviewed every three years by the Superior Council of Magistrates, an independent self-governing institution that oversees the operation of the judiciary. Importantly for our study, judges have no influence over their panel assignments.

Assignment of cases to panels. Cases are then assigned to panels as follows. First, each court has territorial competencies over crimes committed and filed within its geographical limits. Then, cases are randomly assigned to panels. Random assignment of cases is required by law (Art. 25 of the Italian Constitution) to ensure the absence of any relationship between the identity of judges and the characteristics of the cases assigned to them, including the identity of lawyers and the complexity of cases. Again, each court can decide its own allocation rule, as long as it is random.[9] Judges are typically unable to decline assignments unless they are on extended leave (lasting more than one week) or encounter conflicts of interest that compromise their impartiality. If a member of the panel is unavailable or unfit to rule on a case, they are replaced by another judge according to predetermined and objective rules that are published

[8] After three years, judges can transfer to another court. Court vacancies are published twice a year, and judges can apply for reassignment based on a new score ranking. This ranking differs from the initial one and depends on *aptitudes*, *merit*, and *seniority*.

[9] The main allocation method used in Italian courts is a case management software called GIADA (Gestione Integrata delle Attività e dei Dati Amministrativi - Integrated Management of Activities and Administrative Data). This system ensures a fair distribution of workloads among judges using automatic and predetermined criteria. Cases are assigned to panels on a rotating basis according to their complexity, as evaluated by GIADA. Once a case is assigned to a panel, that panel will not receive another case of the same complexity until all other panels have been assigned a similar case.

along with the assignment criteria.

Thus, the structure of the Italian judiciary ensures that the allocation of judges to panels is not a function of their gender, nor are cases assigned based on the gender composition of the panels. Nevertheless, we present formal tests of the exogeneity of assignment of both judges and cases to panels in Section 4.2.

3 Data

To answer our research questions, we require data on the gender composition of judicial panels, the crimes they rule on, the outcomes of the rulings, and a metric for sentence quality. We obtain this information by manually assembling a novel database containing the universe of collegial ruling sentences from a sample of first-instance criminal courts in Italy, which we then merged with the corresponding second-instance sentences. Specifically, we obtain data from four courts in Central and Southern Italy and the corresponding Courts of Appeal: Naples Nord, Torre Annunziata, Naples (under the Court of Appeal of Naples), and Florence (under the Court of Appeal of Florence). These courts vary in size and location. Naples is the largest court in our sample (and second largest in Italy), producing nearly 600 sentences per year in collegial ruling. Florence is a mid-size court, with an average of 172 collegial ruling sentences per year, while Torre Annunziata and Naples Nord are small courts, with 80-90 sentences in collegial ruling per year. Moreover, while Florence is located in the center-north of the country, the other three courts are located in the south. This diversity reassures us that our results are not specific to a particular type of court or geographical area. To our knowledge, such a database has never been assembled before due to the absence of a digitized and accessible database at the court level and thus represents an important contribution of this paper.

Sentences have been made available by the presidents of each court in paper format. We digitized the sentences in collegial ruling and processed them using a series of OCR and text-parsing algorithms to collect the following information: the name of the court and the city it is located in, the start and end date of the trial, the identities of members of the judicial panel, the crime being processed, and the final decision. As Italian names unambiguously identify an individual's gender, we impute the gender of judges using their first name as collected during the scraping. We obtained other personal information on the judges, such as age and municipality and region of provenience, by

web scraping. Figure A3 provides an overview of the structure of a sentence with the relevant parts containing the data of interest highlighted in red.

3.1 Descriptive statistics

Table 1 provides descriptive statistics. Our sample covers almost 20 years of data, from 2004 to 2021.[10] In total, we digitized nearly 10,000 cases. Since a single case may involve multiple defendants and multiple charges, this corresponds to a total of 61,302 decisions. Each of these decisions serves as our unit of observation.

Panel A of Table 1 provides an overview of the composition of the first-instance judiciary in our courts. There are a total of 370 unique judges operating in our sample's courts in the years considered. In line with the trend shown in Figure A2, women in our sample are 56 percent of all judges. This is fairly constant across courts, where the share of women ranges from 50 to 59 percent (not reported in the Table). The average judge age is 49 years old and nearly 20 percent of the judges were born in a region different from the one they work in. As expected, women are younger (47 years old vs 51 years old, not reported in the Table) and less likely to come from a different region (16 percent vs 24 percent, not reported in the Table). Among the 1093 unique panels we cover in our dataset, the majority are mixed-gender (panel B): in particular, 33.9 percent of the panels are composed by two men and one woman, while 41.3 percent are made by one man and two women. All-female panels are more common than all-male panels (18.3 percent compared to 6.5 percent), again in line with the trends shown in Figure A2. Panel C reports the case characteristics in our sample. Cases on average have slightly more than two defendants and five different crimes. The large majority of defendants are male (89 percent) and only 16 percent are foreigners. Finally, Panel D of Table 1 describes the type of decisions taken by the judicial panels. A guilty verdict is issued in 47 percent of the decisions, and conditional on conviction, the average jail time is five years. Trials are long, lasting on average almost three years. Just below 20 percent of the rulings in our sample are appealed against, almost all of which are convictions (97 percent, not reported in the Table).

The types of crimes we observe in our sample are summarized in Appendix Table A1. The three most common categories of offenses, accounting for nearly 70 percent of the total, are crimes against property, against public order, and against the person.

[10] Data from the Court of Florence date back to 2004, while data from Naples Nord start in 2013, when the court was established. Data from Torre Annunziata and Naples start in 2010.

The first group includes offenses against property by violence to things or persons (such as theft, robbery, and extortion), and offenses against property by fraud (such as scam, fraudulent bankruptcy, usury, and laundering). The second group includes offenses such as (mafia-type) criminal associations, drug and arms trafficking, devastation, and looting. The third group instead includes crimes such as personal injuries, manslaughter, threats, and crimes against women and children.

4 Empirical strategy

The main challenge in identifying the effect of team gender composition on decision-making is that the sorting of individuals into teams is unlikely to be random. The ideal experiment to determine the impact of gender composition on the type and quality of team decisions would involve randomizing individuals into teams, and decisions to teams, and comparing the outcomes across teams with different gender compositions.

The setting of the Italian judiciary – characterized by the random allocations of judges to judicial panels and the random assignment of cases to panels within courts – allows us to do precisely that. Intuitively, we compare the probability of ruling guilty for two panels that operate in the same court and rule on the same crime but that are characterized by a different gender composition. This strategy, which allows for endogenous selection of judges into courts, relies on the identifying assumption that both the variation in the gender composition of panels within a particular court and the variation of cases assigned to these panels are random. While judges can presumably choose a court based on its overall characteristics (including the gender composition of its workforce), they are not allowed to choose who to work with and what to work on. We provide evidence in favor of the validity of our identifying assumption in Section 4.2.

4.1 Regression model

Leveraging the random allocations of judges to judicial panels and the random assignment of cases to panels *within courts*, we test whether (i) different judicial panels ruling on similar cases reach different sentencing outcomes depending on their gender composition; and (ii) whether the gender composition of the panel influences the quality of the sentencing, as measured by a) the probability of appeal and b) whether

the appellate court upholds or rejects the panel's ruling. Specifically, we estimate the following models:

$$Y_{isct} = \beta_0 + \beta_1 Mixed_{isct} + \delta X_{ict} + \eta Z_{isct} + \gamma_c + \gamma_{s,t} + \epsilon_{isct} \quad (1)$$

and

$$Y_{isct} = \alpha_0 + \alpha_1 F0_{isct} + \alpha_2 F3_{isct} + \gamma_c + \gamma_{s,t} + \theta X_{ict} + \eta Z_{isct} + \epsilon_{isct} \quad (2)$$

where Y_{isct} is the outcome of interest for panel i in court c ruling on crime s in year t: an indicator equal to 1 if the panel reaches a guilty verdict; the duration of the trial; the probability of appeal; and the probability of having a decision confirmed by the court of appeal. The main variables of interest are $Mixed_{isct}$ in equation (1), which is an indicator equal to 1 if the panel is mixed-gender, and $F0_{isct}$ and $F3_{isct}$ in equation (2), where Fj_{isct} is an indicator equal to 1 if panel i has j women and $(3-j)$ men.[11] Because of random assignment, β_1 identifies the causal effect of having a gender-mixed panel on our outcome of interest compared to a homogeneous panel, while α_1 and α_2 identify the causal effects of having an all-men or an all-women panel on our outcome of interest compared to a mixed-gender panel, respectively.

The coefficients γ_c and $\gamma_{s,t}$ denote court, and type of crime-by-year fixed effects, respectively. The court fixed effects account for unobserved differences in average judicial panel characteristics across courts as well as for courts' characteristics that are constant across panels within a court. We also include crime type-by-year fixed effects to ensure that we are comparing judicial panels ruling on similar crimes at the same time. This allows us to control for crime-specific time trends, such as changes in the interpretation of the law or in the public sensitivity around specific crimes over time. Finally, we also include a vector of controls for panel characteristics, X_{ict}, and a vector of case characteristics, Z_{isct} to increase precision, and we cluster the standard errors at the panel level to account for potential correlation in decision outcomes within panels.[12]

[11] The decision to model mixed-gender teams as a single category, rather than distinguishing between teams with one man and two women versus those with two men and one woman, is based on the absence of statistically significant differences in their decision-making patterns, as we will discuss in more detail in Section 5.4.

[12] The vector X_{ict} includes indicators for whether judges in the panel are all beyond or above the median judge age; indicators for whether judges in the panel are all from the same region or all from different regions. The vector Z_{isct} includes the number of crimes; the number of defendants; an indicator for whether the defendant is female; and an indicator for whether the defendant is foreign.

4.2 Support to the identification strategy

In this Section, we discuss the empirical tests we perform to lend support to the validity of our identification strategy and the proposed interpretation of the results.

Residual Variation: A first-order concern for our ability to obtain precise estimates using our identification strategy is whether we have sufficient variation in the gender composition of the panels once we condition on court and crime type-by-year fixed effects. Appendix Table A2 reports the raw and residual variation in the gender composition of the panels. Conditioning on court and crime type-by-year fixed effects reduces the standard deviation in the share of mixed-gender panels from 43 percent to 41 percent (Panel A): that is, we are left with 95 percent of the original variation unexplained. Similar results hold when we look at all-women and all-men panels (Panel B and C, respectively). This reassures us that our empirical strategy leaves sufficient residual variation in our key variables.

Quasi-random variation: As discussed at the beginning of this section, a causal interpretation of our estimates rests on the assumption that variation in the gender composition of panels within courts results from random fluctuations rather than from systematic selection.

We provide evidence in favor of this assumption in two ways. First, we formally test for selection on observables by examining whether the variation in the gender composition of the panels is correlated with the variation in case characteristics.[13] Table 2 displays the results of separate regressions of trial characteristics on the mixed-gender panel indicator, after including court and crime type times year fixed effects. All variables have been standardized to have mean zero and standard deviation one. None of the coefficients is statistically significant, and they are all negligible in magnitude. This mitigates concerns regarding the possibility that case characteristics are correlated with the gender composition of the panel.

Second, we show that judges assigned to mixed-gender panels and same-gender panels do not differ in terms of observable characteristics. Table 3 displays the results of separate regressions of judges' observable characteristics on the mixed-gender panel indicator, after including court and year fixed effects. All variables have been standardized to have mean zero and standard deviation one. None of the coefficients is statistically significant at the 5 percent significance level, and they are all small in

[13]Tests such as this are common in the literature (See Lavy and Schlosser, 2011, among others).

magnitude.

5 Results

5.1 Type of decisions

We start by investigating whether similar offenses are ruled against differently depending on the gender composition of the judicial panel. Table 4 reports the results from regression (1), where the dependent variable is an indicator equal to one if the panel is mixed-gender. In columns (1) to (3), we progressively add court and crime type-by-year fixed effects and a set of controls at the panel and case level. Across all specifications, mixed-gender teams are less likely to convict the defendant relative to homogeneous teams. In terms of magnitude, mixed-gender panels are 8.4 percentage points less likely to find the defendant guilty (Table 4, column (3)), corresponding to a decrease of 17.9 percent relative to the average conviction rate.

In order to understand whether the effects are different between homogeneously female or homogenously male teams, columns (4) to (6) report the estimated coefficients from equation (2). As above, we progressively add court and crime type-by-year fixed effect and a set of controls at the panel and case level. We find that the effect is driven by all-women teams who rule more severely: they are 10.2 percentage points (21.7 percent) more likely to convict the defendant than mixed-gender panels (column (6)). In contrast, all-men panels do not show a significant difference in conviction rates compared to mixed-gender panels. All-women teams are also 7.7 percentage points (16.3 percent) more likely to convict the defendant than all-men panels.

As we are interested in examining not only differences in conviction rates but also the severity of sentencing, we turn to investigate whether teams make different decisions along the sentencing margin according to their gender composition. We therefore estimate the effect of team gender composition on the number of years of jail time. Columns (1) and (2) of Table 5 report the results of an unconditional regression of team's gender composition on jail time, and show that all-women teams sentence defendants to 5.5 months more, on average, than mixed-gender teams.[14] Because this unconditional measure reflects both the extensive and the intensive margins, in

[14]Note that the number of observations in these regressions is lower than our total number of observations as there are a small fractions of decisions for which the jail time is missing. Our results are robust to excluding these observations.

Columns (3) and (4) we focus on conviction decisions only. Conditional on a conviction, all-women teams impose sentences that are, on average, 12 months shorter than those given by all-men teams. Thus, although all-women panels have higher conviction rates than both mixed-gender and all-men panels, they do not impose on average longer jail sentences.[15]

The findings in Table 5 represent average effects that, as such, might mask important differences along the sentencing distribution. To investigate this potential heterogeneity, we extend the analysis from columns (3) and (4) of the table to examine differences across percentiles of jail time.[16] Figure 1 reports the estimated coefficients. At the lower end of the sentencing distribution, there is little to no significant difference in jail times between the three types of teams. As sentencing severity increases, all-men teams impose significantly longer sentences than both mixed-gender and all-women teams, with the effects gradually decreasing to disappear at the top percentiles of the sentencing distribution. By contrast, all-women teams exhibit minimal differences relative to mixed-gender teams across most of the distribution, with the exception of the 90th percentile, where they impose shorter sentences, albeit the coefficient is imprecisely estimated.

5.2 Duration of the decision-making process

In this subsection, we investigate whether the gender composition of a decision-making team affects the time required to reach a decision. This is an important margin to consider, as faster decision-making is often desirable, especially in the judiciary where case backlog is substantial and costly (Coviello et al., 2019). However, longer decision times can be associated with increased accuracy, as teams might spend more time deliberating or gathering additional information. At the same time, extended decision times can also reflect the presence of friction, such as prolonged discussions due to differing viewpoints. In the first scenario, there is a clear trade-off between speed and quality of decisions. In the second scenario, the impact of friction is ambiguous: they can either hinder or enhance the decision-making process depending on whether the discussions are constructive.

[15] This suggests that the higher conviction rates observed for all-women teams are driven by a higher propensity to find defendants guilty, rather than a tendency to impose more severe punishments per se – an interpretation that will be revisited in Section 6.

[16] The 10th, 25th, 50th, 75th and 90th percentiles of jail time are 1.3, 2, 4, 7, and 12 years, respectively.

We use the trial duration as a proxy for the duration of the decision-making process. We calculate the difference between the year when the case is assigned to a panel and the year in which they take their final ruling and estimate the effect of gender composition on this measure. Given that cases are randomly assigned to panels, any difference in trial duration conditional on court-fixed effects can be attributed to the different composition of the panel itself. Table 6 shows the results. All-men teams are faster than any other team composition, with their trials lasting 5.4 months less than those of mixed-gender teams, and 4.7 months less than those of all-women teams (Columns (1)-(2)). Compared to the average trial duration, these effects correspond to a 15.6 percent and 13.4 percent decrease, respectively. This pattern holds also when focusing on conviction decisions only: all-men teams trials last approximately 4.3 months (12.2 percent) less than mixed-gender teams and all-women teams (Columns (3)-(4)).

Since having a fast trial can be desirable for addressing court backlogs (Coviello et al., 2019), our results suggest that all-men teams could be more efficient provided that their rulings are not appealed – and subsequently reversed – at a higher rate. Otherwise, there is a trade-off between faster decisions and increased costs due to more appeal trials. In the next section, we will investigate whether the patterns we uncovered have implications for decision quality.

5.3 Quality of decisions

We begin by investigating whether the gender composition of a decision-making team affects the *perceived* accuracy of their decisions. To do this, we exploit the hierarchical structure of the judiciary and use the probability of a decision being appealed as a proxy for its perceived quality. The underlying assumption is that if a decision is not appealed against, it is perceived to be a good one. While in principle all types of decisions can be appealed against (by defendants and prosecutor), appeals on convictions are more likely.[17] Indeed, conviction decisions are 38 percentage points more likely to be appealed against for the same type of crime, and one more year of jail time increases the probability of appeal by 3.6 percentage points (Table 7).

[17]When only the defendant appeals, the Court of Appeal cannot worsen the outcome of the initial judgment (art.579 Codice di Procedura Penale). In other words, the appellate judge can either uphold the initial sentence or reduce it. This creates an incentive for defendants to appeal as long as the probability of reducing her sentence is larger than the fixed cost of appealing.

Table 8 reports the results of the effect of gender composition on appeal rates. We find that decisions made by mixed-gender teams are 4 percentage points less likely to be appealed compared to those made by same-gender teams (column (1)). This effect appears to be entirely driven by all-men teams, whose decisions are 13 percentage points more likely to be appealed (column (2)). However, since there are differences in conviction rates, part of the observed effect could be mechanical. To address this, in columns (3) and (4) we replicate the analysis conditioning on conviction decisions. Our findings indicate that decisions made by all-men panels are 22.7 percentage points more likely to be appealed, while there is no difference in appeal rates for conviction decisions made by mixed-gender and all-women panels. This suggests that although all-men teams are faster in reaching decisions, their verdicts are more likely to be perceived as incorrect.

Next, we investigate the effect of team gender composition on the *effective* quality of judicial decisions. We leverage the appellate process, which reassesses the rulings made by the judicial panels in the first-instance court. The underlying assumption is that judges in the appellate courts are more skilled and their decisions are more likely to reflect the true state of the defendant's guilt or innocence. Crucially, when a ruling is appealed, appellate courts do not review the original case by retrying it or hearing new evidence. Instead, they do so by examining the procedures and decisions of the first instance court to ensure the law was both correctly interpreted and applied. This is important, as it implies that appellate courts operate with the same set of information as the judges who made the initial ruling. We define a panel's ruling as *correct* if:

(a) the ruling has not been appealed against; or

(b1) the ruling has been appealed against and the conviction/acquittal decision is confirmed; and

(b2) the defendant was convicted and the difference in imprisonment time between the first instance ruling and the appeal ruling is lower than the average difference (in absolute value).[18]

[18]Including (b2) is essential to introduce meaningful variation in the outcome. Without considering the intensive margin of conviction and focusing solely on whether the conviction or acquittal decision was upheld, 99 percent of decisions (and 98 percent of convictions) would be classified as correct decisions. By contrast, our more nuanced definition reduces this figure to 88 percent (75.5 percent of convictions), allowing for a more detailed evaluation of decision quality.

Table 9 reports the results. The effects of gender composition on *effective* quality mirror those on *perceived* quality. All-men teams are not only more likely to be perceived as wrong, but they are also 10 percentage points more likely to see their decisions reverted, while all-women teams and mixed-gender teams are not (columns (1)-(2)). The effects are even more pronounced when we focus on the subsample of convictions decisions. Mixed-gender panels are 3.5 percentage points more likely to make a correct decision compared to same-gender panels (column (3)). The effect is driven by all-men panels, whose decisions are 20 percentage points (26 percent) less likely to be classified as correct.

5.4 Robustness checks and heterogeneity

In this section, we assess the robustness of our results and perform additional heterogeneity analysis. To prevent duplication in the presentation of robustness tables, we focus on all-men and all-women teams only.[19]

Selection on unobservables: The balancing tests in Tables 2 and 3 suggest that the variation in the gender composition of the panels is not correlated with the variation in case or judge characteristics. Consequently, it is unlikely that unobserved characteristics correlated both with our outcomes and with variation in gender composition are the main drivers of our results. To provide additional support to this, we show that our results are not sensitive to the inclusion/exclusion of controls and type of crime-by-year fixed effects. This is important for two reasons. First, it provides additional evidence that the allocation of judges to panels and of cases to judges is indeed random, conditional on court. Second, it helps us argue that our effects are not driven by other dimensions of diversity, as we will discuss more in depth in the next subsection. Appendix Table A3 reports the results on conviction rates (panel A), trial duration (panel B), and on our decision-quality indicator (panel C). Column (1) shows the results of a regression with court fixed effects only. Column (2) includes the crime type-by-year fixed effects. In columns (3) and (4), we progressively add judge-level controls and case-level controls, respectively. The point estimates are stable across all columns.

Correlated characteristics: One concern we might have is that, rather than

[19]Results on the effects associated with mixed-gender teams are also robust to the tests discussed below.

capturing the effect of gender, we are picking up that of other characteristics that correlated with gender. For instance, in many male-dominated occupations, women tend to be younger and more educated than men. Women, moreover, tend also to be less geographically mobile than men. While we are confident that educational differences do not play a role in our setting – given that the path to becoming a judge is highly standardized and identical for both men and women – we may still be concerned about women being on average younger. Additionally, female judges are less likely to work in a region different from where they were born.

We rule out that our results are driven by these other dimensions of diversity in three ways. First, in all our specifications, we control for judges' age and region of origin, ensuring that our results reflect the effect of gender composition beyond these other characteristics. Moreover, the inclusion of these controls does not substantially alter our estimates, indicating that they are not significant contributors to the observed effects (see Table A3). Second, we examine the effects of team age and regional composition on conviction rates, trial duration, and decision quality, regardless of gender composition. Results are reported in Table A4. In Panel A, we analyze the effect of age composition, dividing panels into all-young, all-old and mixed-age groups based on the proportion of judges above the median age of the court.[20] In Panel B, we focus on regional composition, classifying panels into all-south, all-north and mixed-region groups based on judges' region of birth. If our earlier analyses were capturing the effects of age or regional composition rather than gender composition, we would expect similar estimated coefficients to those reported in Tables 4, 6 and 9.[21] This does not seem to be the case. We find no significant effects of age or regional composition on conviction rates, trial duration and appeal rate, suggesting that our effects are indeed driven by differences in the gender composition of panels as opposed to differences in characteristics that correlate with judges' gender. Moreover, these results suggest that other dimensions of diversity, such as gender or regional differences, do not matter in our context. Finally, to further reassure that our results are not driven by judges' characteristics, we estimate a more restrictive specification with judge fixed effects. That is, we estimate the following model separately for men and women:

[20]The median age is 39 years in the court of Naples Nord; 42 in the court of Torre Annunziata; 48 years in the court of Florence; 51 years in the court of Naples.

[21]Specifically, we would expect the coefficient of all-young and all-south to be similar to that of all-women teams given that women tend to be younger than the median age and are more likely to work in the same region in which they are born.

$$Y_{j,isct} = \beta_0 + \beta_1 Samegender_{j,isct} + \delta Z_{ict} + \gamma_c + \gamma_{s,t} + \gamma_j + \epsilon_{isct}$$

where Y_{isct} is the outcome of interest for judge j in panel i in court c ruling on crime s in year t; $Samegender_{isct}$ is an indicator equal to 1 if the judge sits in a gender homogenous team; Z_{isct} is a vector of case characteristics; and γ_c, $\gamma_{s,t}$ and γ_j are court, crime type-by-year and judge fixed effects, respectively. Table A5 reports the results, which are largely consistent with our main findings.

Different classifications of crimes: Our crime classification follows the broad categories defined by the Penal Code, which group together different crimes within the same general area. However, this approach may mask important heterogeneity among offenses. To address this, we construct a more granular classification that separates the most common offenses into distinct categories: mafia-type crimes, robbery, money laundering, extortion, personal injury, fraud, sexual violence, and fraudulent bankruptcy. Appendix Table A6 shows that our results remain robust under this alternative and more detailed classification.

Gender-specific attitudes on different topics: One potential driver of the observed differences in team behavior is that men and women have different attitudes, knowledge, and sensitivity toward various topics. To explore this possibility, we investigate whether specific types of crimes drive the effects documented in the previous sections. If this mechanism was driving our results, we might expect the presence of women to have a more significant impact on gender-specific offenses, such as violence against women or children (included in the "crimes against the person" category), compared to crimes like theft (included in the "crimes against the property"). We investigate this possibility in Appendix Table A7. In each column (1)-(20), we exclude one category of crimes at a time and estimate the effect of gender composition on conviction rates (Panel A), trial duration (Panel B), and decision quality conditional on conviction (Panel C). Our findings show that the coefficients remain consistent in both sign and magnitude across all specifications and are not statistically different from our baseline estimates. This suggests that no specific crime is driving our results, and that our effects are robust and pervasive across various types of decisions.

Gender and ethnicity of the defendant: One might worry that our results might be partially driven by different attitudes that men and women might have when interacting with defendants of the same or opposite gender. We replicate our analysis

by distinguishing between the defendant's gender and nationality. Appendix Table A8 presents the results. We estimate the effect of gender composition on conviction rates (Panel A), trial duration (Panel B), and quality conditional on conviction (Panel C) for male vs. female defendants (columns (1) and (2)) and Italian vs. foreign defendants (columns (3) and (4)). The patterns observed in the previous sections remain consistent across most outcomes.

Unpacking the mixed gender teams: So far, we have considered the mixed-gender teams as a single category. However, one might think that the effects are driven by the number of women in the team. To explore this, we replicate the analysis in Tables 4, 6 and 9 by distinguishing between teams made by one man and two women, and teams made by two men and one woman. Specifically, we estimate the following specification:

$$Y_{isct} = \theta_0 + \theta_1 F0_{isct} + \theta_2 F2_{isct} + \theta_3 F3_{isct} + \delta X_{ict} + \eta Z_{isct} + \gamma_c + \gamma_{s,t} + \epsilon_{isct}$$

where Fj_{isct} takes value 1 if panel i has j women and $(3-j)$ men and 0 otherwise, and all the other variables are defined as in equations (1) and (2).[22] Then, θ_1, θ_2, and θ_3 describe the causal effect of having a judicial panel with no woman, one woman, and two women on our outcomes of interests compared to a panel with three women. Appendix Table A9 presents the results. Across all specifications, the estimates for the two types of mixed-gender teams are not statistically different. In other words, mixed-gender teams exhibit similar decision-making behavior regardless of the ratio of men to women, as long as both genders are represented.

6 Interpretation of results

Overall, our findings indicate that gender composition significantly influences decision-making, with distinct behaviors observed in women-only and men-only teams. Specifically, mixed-gender teams tend to be more lenient compared to same-gender teams. This effect is driven by all-women teams convicting more often for similar crimes. Conversely, all-men teams are faster than any other team composition, but this speed comes at the cost of higher appeal and reversal rates. In this section, we investigate the plausible mechanisms explaining why gender composition matters for the decision-

[22] The omitted category is a women-only team $F3_{isct}$.

making process. Our identification strategy allows us to exclude explanations based on factors such as general trends in conviction rates or the characteristics of the courts, as we compare judges working in the same court and ruling on the same type of crime.[23] Moreover, in Section 5.4 we showed that our results are not driven by differences in other characteristics that correlate with gender, nor by specific types of crimes where we might expect gender to be a salient dimension.

In the following section, we argue that these patterns can be best understood through a stylized framework of diagnostic decisions, following the model proposed by Chan et al. (2022), which decomposes the roots of practice variation in decisions across radiologists into dimensions of skills/effort and preferences. In particular, while the standard view in much of the literature on both gender and judges is to assume that systematic variation in decision-making results from variation in agent's preferences, in Section 6.2 we present evidence supporting the notion that our results are driven by a skill and effort channel.

6.1 Conceptual framework

Our findings can be best interpreted through a stylized framework of diagnostic decisions, following the model proposed by Chan et al. (2022). Within this framework, we can break down judges' decision-making processes into two steps. First, there is an evaluation phase during which judges aim to establish the facts of a crime and determine the defendant's true state of guilt. Following the terminology used by Chan et al. (2022), we refer to the accuracy of these assessments as judges' *diagnostic skills*, which might reflect not only raw talent but also effort. The second phase involves a selection process, where judges, based on the signal of the defendant's true guilt extracted in the first phase, determine the preference weights to assign to the various costs and benefits influencing the final decision. We refer to the weights used in this phase as judges' preferences. For example, in the context of a binary decision such as whether to convict, the first step involves ranking cases by their suitability for conviction, while the second step involves selecting the cutoff point for conviction within that ranking.

In their classification decisions, judges can make mistakes. Importantly, in our context mistakes are costly and asymmetric, with the cost of false positives (convicting an innocent person) being higher than that of false negatives (acquitting a guilty

[23]Moreover, we may reasonably expect these factors to affect both men and women similarly.

person).[24] However, it is empirically difficult to distinguish false negatives from true negatives (where an innocent defendant is acquitted) because acquittal decisions are rarely appealed, therefore eliminating the possibility of further scrutiny to identify potential errors. In contrast, defendants' incentive to appeal a conviction (as by law their sentence cannot be worsened) implies that false positives are more likely to come to light during the appeals process, where the conviction can be reviewed and potentially overturned.

Judges can differ both in how they prioritize the trade-off between false positives and false negatives (i.e., their *preferences*) and in their ability to accurately assess a defendant's true guilt (i.e., their *diagnostic skills*). Intuitively, the greater a judge's aversion to false positives relative to false negatives, the lower the conviction rate will be. At the same time, higher signal precision leads to higher conviction rates, as a more accurate assessment reduces the likelihood of false positives. Moreover, because precise signals enable judges to better distinguish between guilty and innocent defendants, greater diagnostic skills are expected to result in fewer errors of both types.

6.2 Role of skills and effort

Two important considerations emerge within this framework. First, the gender composition of a panel can induce systematic variation in decision-making either through a *skill and effort channel*, affecting a panel's assessment of a case, or through a *preference channel*, affecting a panel's propensity to convict given the assessment of a case. Second, a model à la Chan et al. (2022) implies that under quasi-random assignment of cases to panels, the joint distribution of conviction rates and error rates across panels with different gender compositions (as measured by appeal rates and reversal rates) can be used to assess the role of these two channels in explaining our results.

Our findings rule out the possibility that variations in preferences alone drive the observed patterns. Specifically, the fact that teams prone to committing more false positives also fail to convict at a higher rate (see Tables 4 and 9) allows us to reject the hypothesis that all teams operate on the same production possibilities frontier.

[24] In criminal law, the *Blackstone's ratio* principle states that "it is better that ten guilty persons escape than that one innocent suffer", as articulated by the English jurist William Blackstone in the 1760s. This principle underpins the standard of *guilt beyond a reasonable doubt*, which ensures that a defendant is presumed innocent unless evidence proves them guilty beyond a reasonable doubt. This standard reflects the principle's caution against convicting the innocent, and implies that any doubt about a defendant's guilt should result in acquittal.

Instead, this points to differences in skill and effort across teams playing a significant role.[25]. Intuitively, when teams observe an imprecise signal, they convict less to avoid the high cost of false positives, but their error rates upon conviction increase precisely because of the lower quality of the signal. This gives rise to a negative relationship between conviction rates and error rates, which is inconsistent with the predictions from standard models with no skill variation. In what follows, we provide two additional pieces of evidence supporting the role of variations in skill and effort in driving our results.

Case complexity: The observed relationship between conviction rates and false positive rates is consistent with the idea that effort varies across teams. We further explore this hypothesis by examining how our findings differ based on the complexity of the cases teams adjudicate. Assuming that simpler cases require less effort to extract a precise signal, we would expect the effects to be concentrated in complex cases. Table 10 presents estimates separately for *more complex cases* (involving multiple crimes and defendants) and *less complex cases* (involving a single crime or defendant). Consistent with the idea that effort plays a key role, we find that our baseline effects are primarily driven by complex cases. In simpler cases, all teams behave similarly, regardless of their gender composition. In contrast, in complex cases the patterns identified in Section 5 emerge: all-women teams exhibit higher conviction rates than any other team, while all-men teams tend to decide more quickly but make more mistakes in their convictions.

Written opinions: Disentangling the role of innate skill from effort in decision-making is inherently challenging. To attempt this distinction, we leverage the requirement that judicial teams must provide written reasoning for their decisions. These "Reasons for Judgment" resemble U.S. judicial opinions but differ in key ways. Italian judges must produce a detailed written explanation for every verdict, whether guilty or not guilty, analyzing all evidence, witness testimonies, expert opinions, and documents presented at trial. Unlike in the U.S., where reasoning often precedes the verdict, Italian judges issue the verdict first and then have a legal deadline—typically 30 to 90 days, though longer for complex cases—to file the written justification. Instead of relying on case law, they base their reasoning primarily on legal codes and, occasion-

[25]In other words, the results shown in Section 5 not only allow us to reject the assumption of uniform skill and effort across teams, but also imply that skill and effort must be systematically correlated with conviction propensities

ally, Supreme Court maxims. While these texts do not create binding precedent, their reasoning may be scrutinized on appeal.

We focus on three key aspects of these documents: the time taken to draft them, their length (in words), and the frequency of Court of Cassation maxims cited. Results are reported in Table 11.[26] Mixed-gender teams take longer to write their opinions, but their judgments are shorter, suggesting greater effort in ensuring clarity and effective communication. In contrast, all-male teams produce significantly longer judgments – more than twice the length of others – and cite Court of Cassation maxims more than twice as often as both mixed-gender and all-female teams. While referencing past maxims is not inherently discouraged, an overly long judgment and excessive reliance on maxims can weaken judicial reasoning. Lengthy opinions often lack clarity and conciseness, making it harder for parties, lawyers, and appellate courts to follow the legal reasoning. Furthermore, heavy citation of maxims can lead to the arguments appearing mechanically constructed, rather than reflecting a thoughtful and case-specific analysis.

Learning from the other gender: Finally, we examine how a team's behavior changes when interactions with co-workers of the opposite gender are more frequent. We measure this by calculating the proportion of team members who have worked in mixed-gender panels in the previous 15 trials.[27] A team is considered to have more experience working with the opposite gender if at least two of its members have a higher-than-average share of experience in mixed-gender teams. We estimate the following regression:

$$Y_{isct} = \beta_0 + \beta_1 F0_{isct} + \beta_2 F3_{isct} + \beta_3 MG_exp_{isct} + \beta_4 F0 \times MG_exp_{isct} + \\ + \beta_5 F3 \times MG_exp_{isct} + \gamma_c + \gamma_{s,t} + \theta X_{ict} + \eta Z_{isct} + \epsilon_{isct}$$

where MG_exp_{isct} is an indicator equal to one if the panel has at least two judges with a higher than average share of cases ruled in mixed-gender panels (among the last 15 cases).[28] Table 12 shows the results. We report the coefficients $\beta_3 + \beta_4$ and $\beta_3 + \beta_5$ at the bottom of the table. These coefficients represent the effect of having a higher experience working in mixed-gender panels for all-men teams and for all-

[26]The analysis excludes the Court of Naples, for which data on written opinions was not available.

[27]Replicating this analysis using different time horizons (3, 5, or 10 cases) does not significantly alter our results.

[28]The average share of cases worked in ruled in mixed-gender panels is 74 percent, corresponding to 11 cases.

women teams, respectively. Our results indicate that all-men teams whose members have ruled more frequently in mixed-gender panels have longer decision-making times and higher decision quality compared to all-men teams with less mixed-gender experience. These patterns suggest that all-men teams with greater mixed-gender experience engage in longer deliberations and ultimately make sound decisions when consensus is reached. Overall, these results suggest that frequent interactions with colleagues of the opposite gender might improve decision-making in homogeneous teams that were underperforming at baseline.

7 Conclusions

This paper aims to investigate the effect of a team's gender composition on decision-making and decision quality. We focus on the Italian judiciary setting, as it allows us to i) exploit the quasi-random allocation of judges and cases to judicial panels, in which ruling takes place collegially; ii) empirically assess the *quality* of the decisions taken by the teams as proxied by the probability that the sentence is confirmed or overturned in appeal.

Our findings indicate that gender composition significantly influences decision-making, with distinct behaviors observed in women-only and men-only teams. Specifically, mixed-gender teams tend to be more lenient compared to same-gender teams, with the effect driven by all-women teams issuing stricter rulings for similar crimes. Conversely, all-men teams are faster than any other team composition, but this speed comes at the cost of higher appeal and reversal rates. We do not find evidence in support of men and women having different sensitivities on specific topics, as our effects are not driven by specific types of crimes where we might expect gender to play a role. Similarly, our results are not driven by other dimensions of diversity, such as age and regional provenience.

These patterns can be understood through a stylized framework of diagnostic decisions, as proposed by Chan et al. (2022). In this model, the accuracy of the signal that teams extract from cases determines conviction rates and error rates. More precise signals lead to higher conviction rates and fewer errors, while imprecise signals result in lower conviction rates and higher error rates due to the increased cost of false positives. As we empirically observe a lack of a positive relationship between conviction rates and false positives, we argue that differences in skills and effort, rather than

in preferences alone, drive our findings.

Our analysis of underlying mechanisms further supports this hypothesis. Specifically, we find that gender composition matters in complex decision-making only, suggesting that gender composition plays a key role precisely when the case requires deeper analysis and extracting an informative signal is more challenging. Furthermore, we find that teams with longer decision-making times and higher quality outcomes are also teams who put more effort in writing the judicial opinions. Finally, we find that interaction with colleagues of the opposite gender improves decision-making for all-men teams, indicating that diversity in team composition can improve the quality of judicial decisions of teams that were underperforming at baseline.

References

Ahern, K. R. and Dittmar, A. K. (2012). The changing of the boards: The impact on firm valuation of mandated female board representation. The Quarterly Journal of Economics, 127(1):137–197.

Alesina, A. and La Ferrara, E. (2014). A test of racial bias in capital sentencing. American Economic Review, 104(11):3397–3433.

Andreoni, J. and Vesterlund, L. (2001). Which is the fair sex? gender differences in altruism. The Quarterly Journal of Economics, 116(1):293–312.

Anwar, S., Bayer, P., and Hjalmarsson, R. (2012). The impact of jury race in criminal trials. The Quarterly Journal of Economics, 127(2):1017–1055.

Anwar, S., Bayer, P., and Hjalmarsson, R. (2019). A jury of her peers: The impact of the first female jurors on criminal convictions. The Economic Journal, 129(618):603–650.

Apesteguia, J., Azmat, G., and Iriberri, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. Management Science, 58(1):78–93.

Arnold, D., Dobbie, W., and Yang, C. S. (2018). Racial bias in bail decisions. The Quarterly Journal of Economics, 133(4):1885–1932.

Ash, E., Asher, S., Bhowmick, A., Bhupatiraju, S., Chen, D. L., Devi, T., Goessmann, C., Novosad, P., and Siddiqi, B. (2023). In-group bias in the indian judiciary: Evidence from 5 million criminal cases. Technical report, Center for Global Development.

Ashraf, N., Bandiera, O., Minni, V., and Quintas-Martínez, V. (2022). Gender roles and the misallocation of labour across countries.

Bagues, M., Sylos-Labini, M., and Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? American Economic Review, 107(4):1207–38.

Bagues, M. F. and Esteve-Volart, B. (2010). Can gender parity break the glass ceiling? evidence from a repeated randomized experiment. The Review of Economic Studies, 77(4):1301–1328.

Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., and Topalova, P. (2009). Powerful women: does exposure reduce bias? The Quarterly Journal of Economics, 124(4):1497–1540.

Bertrand, M. (2020). Gender in the twenty-first century. In AEA Papers and proceedings, volume 110, pages 1–24.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. Journal of Economic Literature, 55(3):789–865.

Born, A., Ranehill, E., and Sandberg, A. (2020). Gender and willingness to lead: Does the gender composition of teams matter? Review of Economics and Statistics, pages 1–46.

Carrell, S. E., Hoekstra, M., and West, J. E. (2019). The impact of college diversity on behavior toward minorities. American Economic Journal: Economic Policy, 11(4):159–82.

Chan, D. C., Gentzkow, M., and Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. The Quarterly Journal of Economics, 137(2):729–783.

Chattopadhyay, R. and Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in india. Econometrica, 72(5):1409–1443.

Coviello, D., Ichino, A., and Persico, N. (2019). Measuring the gains from labor specialization. The Journal of Law and Economics, 62(3):403–426.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. Journal of Economic Literature, 47(2):448–74.

Delfino, A. (2024). Breaking gender barriers: Experimental evidence on men in pink-collar jobs. American Economic Review, 114(6):1816–1853.

Dessein, W. and Santos, T. (2006). Adaptive organizations. Journal of political Economy, 114(5):956–995.

Dufwenberg, M. and Muren, A. (2006). Gender composition in teams. Journal of Economic Behavior & Organization, 61(1):50–54.

Goldin, C. (2014). A grand gender convergence: Its last chapter. American Economic Review, 104(4):1091–1119.

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. Journal of Public Economics, 143:73–88.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. Journal of Political Economy, 111(3):465–497.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2012). Diversity and productivity in production teams. In Advances in the Economic Analysis of participatory and Labor-managed Firms. Emerald Group Publishing Limited.

Hjort, J. (2014). Ethnic divisions and production in firms. The Quarterly Journal of Economics, 129(4):1899–1946.

Hoogendoorn, S., Oosterbeek, H., and Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. Management Science, 59(7):1514–1528.

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The allocation of talent and us economic growth. Econometrica, 87(5):1439–1474.

Knepper, M. (2018). When the shadow is the substance: Judge gender and the outcomes of workplace sex discrimination cases. Journal of Labor Economics, 36(3):623–664.

Lavy, V. and Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. American Economic Journal: Applied Economics, 3(2):1–33.

Lazear, E. P. (1999). Globalisation and the market for team-mates. The Economic Journal, 109(454):15–40.

Lazear, E. P. and Shaw, K. L. (2007). Personnel economics: The economist's view of human resources. Journal of Economic Perspectives, 21(4):91–114.

O*NET (2022). Work context: Work with work group or team. Retrieved from https://www.onetonline.org/find/descriptor/result/4.C.1.b.1.e.

Prat, A. (2002). Should a team be homogeneous? European Economic Review, 46(7):1187–1207.

Schaede, U. and Mankki, V. (2022). Quota vs quality? long-term gains from an unusual gender quota.

Tables and Figures

Table 1: Descriptive statistics

	Mean	St.dev	Min	Max
A. Judges				
Female	0.559	0.497	0	1
Age	48.703	9.092	28	74
From central Italy	0.146	0.353	0	1
From northern Italy	0.040	0.196	0	1
From southern Italy	0.814	0.389	0	1
From a diff region	0.197	0.397	0	1
N. judges	370			
B. Panels				
3M panel	0.065	0.247	0	1
2M1F panel	0.339	0.474	0	1
1M2F panel	0.413	0.493	0	1
3F panel	0.183	0.387	0	1
N. panels	1093			
C. Case characteristics				
Number of defendants	2.293	4.821	1	220
Number of crimes	5.819	20.271	1	609
Defendant is foreign	0.163	0.362	0	1
Defendant is female	0.107	0.270	0	1
N. Cases	9729			
D. Decisions characteristics				
Guilty	0.470	0.499	0	1
Jail time	5.301	4.865	0	30
Length of trial	2.950	2.652	0	35
Appeal	0.190	0.393	0	1
N. Decisions	61302			

Notes. This table reports descriptive statistics for our sample. Panel A reports judges characteristics. *Female* is an indicator for whether the judge is a woman. *From a diff region* is an indicator for whether the judge works in a region that is different from the one they were born in. Panel B reports panels' characteristics. *3M panel* denotes a panel made by all men; *2M1F panel* denotes a panel made by two men and one woman; *1M2F panel* denotes a panel made by one man and two women; *3F panel* denotes a panel made by all women. Panel C describes cases characteristics. Panel D describes decisions. *Guilty* is an indicator for whether the panel convicts the defendant. *Jail time* is the years of jail time the defendant is sentenced for, conditional on conviction. *Appeal* is an indicator for whether the decision is appealed against.

Table 2: Balancing tests - case characteristics

	All-men		Mixed-gender		All-women	
	Estimate	Std.Err.	Estimate	Std.Err.	Estimate	Std.Err.
Dependent variables:						
N. of crimes	-0.018	(0.023)	0.008	(0.039)	0.002	(0.040)
N. of defendants	0.005	(0.033)	0.014	(0.027)	-0.018	(0.026)
Defendant is female	0.008	(0.010)	0.015	(0.014)	-0.021	(0.013)
Defendant is foreign	-0.001	(0.021)	-0.006	(0.019)	0.007	(0.019)

Notes. This table reports the coefficients of separate regressions of each characteristic on the mixed-gender panel indicator. All regressions include court and crime type times year fixed effects. All variables have been standardized. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 3: Balancing tests - judge characteristics

	Women		Men	
	Estimate	Std.Err.	Estimate	Std.Err.
Dependent variables:				
Age	-0.033*	(0.012)	-0.073	(0.046)
From different region	-0.009	(0.015)	0.034	(0.022)
From northern Italy	-0.009*	(0.003)	0.025	(0.013)
From central Italy	-0.018	(0.015)	-0.044	(0.039)
From southern Italy	0.021	(0.015)	0.031	(0.032)

Notes. This table reports the coefficients of separate regressions of each characteristic on the mixed-gender panel indicator, separately for men and women. All regressions include court and crime type times year fixed effects. All variables have been standardized. Standard errors clustered at the court level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 4: Effect of gender composition of the panel on probability of conviction

	Guilty					
	(1)	(2)	(3)	(4)	(5)	(6)
Mixed gender panel	-0.113***	-0.091***	-0.084***			
	(0.025)	(0.022)	(0.024)			
All-women panel				0.118***	0.101***	0.102***
				(0.026)	(0.024)	(0.025)
All-men panel				0.095**	0.059	0.025
				(0.044)	(0.039)	(0.039)
Observations	61302	61302	61302	61302	61302	61302
Court and crime x year FEs		X	X		X	X
Controls			X			X
Mean dep.var.	0.470	0.470	0.470	0.470	0.470	0.470

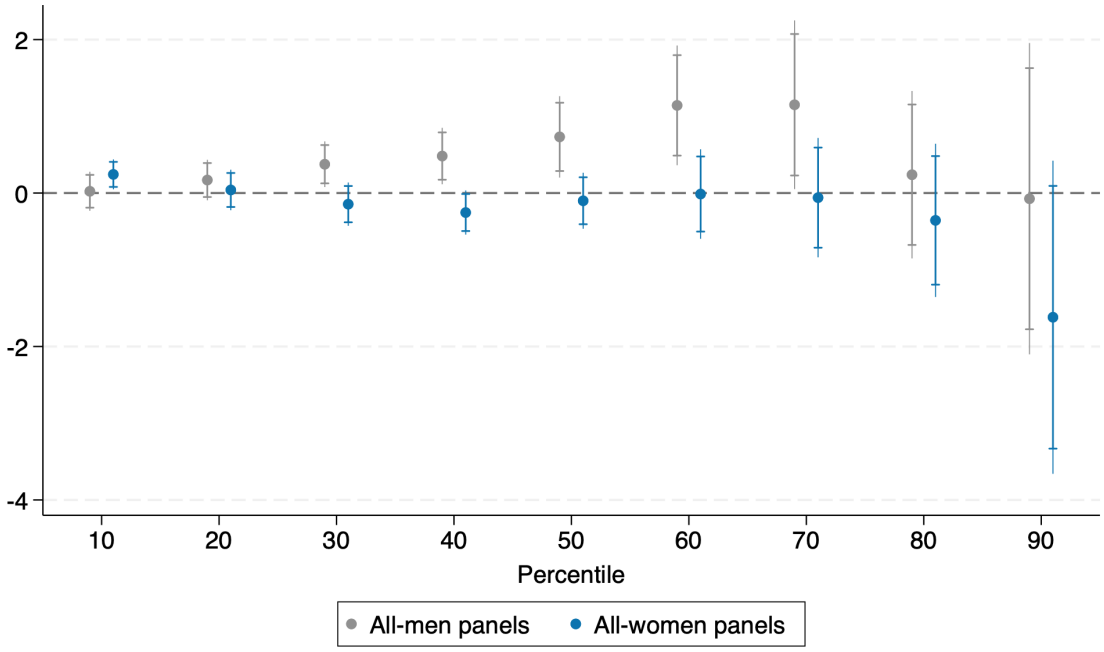
Notes. In all columns, the dependent variable is an indicator for whether the panel convicts the defendant. *Mixed-gender panel* is an indicator for whether the gender composition of the panel is mixed. *All-women panel* is an indicator for whether the panel consists of women only. *All-men panel* is an indicator for whether the panel consists of men only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 5: Effect of gender composition of the panel on jail time

	Jail sentence (years)			
	(1) Unconditional	(2) Unconditional	(3) Conditional	(4) Conditional
Mixed gender panel	-0.453**		0.268	
	(0.176)		(0.275)	
All-women panel		0.456**		-0.514
		(0.195)		(0.329)
All-men panel		0.443		0.511
		(0.321)		(0.362)
Observations	59541	59541	27017	27017
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	2.448	2.448	5.395	5.395

Notes. In all columns, the dependent variable is the duration of imprisonment (in years) the defendant is sentenced to. In columns (1) and (2), the analysis is done on the whole sample; in columns (3) and (4), the analysis is restricted to conviction decisions only. *Mixed gender panel* is an indicator for whether the gender composition of the panel is mixed. *All-women panel* is an indicator for whether the panel consists of women only. *All-men panel* is an indicator for whether the panel consists of men only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Figure 1: Effects of gender composition of the panel on percentiles of jail time



Notes. This figure plots the regression coefficients and the 90% and 95% confidence intervals from equation (2) by percentiles of the dependent variable. The dependent variable is the duration of imprisonment (in years) the defendant is sentenced to. *All-women panel* is an indicator for whether the panel consists of women only. *All-men panel* is an indicator for whether the panel consists of men only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors are clustered at the panel level.

Table 6: Effect of gender composition of the panel on decision-making time

	Length of trial (years)			
	(1) Unconditional	(2) Unconditional	(3) Conditional	(4) Conditional
Mixed gender panel	0.153		0.085	
	(0.171)		(0.143)	
All-women panel		-0.062		-0.003
		(0.197)		(0.166)
All-men panel		-0.453*		-0.355**
		(0.234)		(0.170)
Observations	61302	61302	28778	28778
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	2.906	2.906	2.464	2.464

Notes. In all columns, the dependent variable is the length of the trail (in years). In columns (1) and (2), the analysis is done on the whole sample; in columns (3) and (4), the analysis is restricted to conviction decisions only. *Mixed gender panel* is an indicator for whether the gender composition of the panel is mixed. *All-women panel* is an indicator for whether the panel consists of women only. *All-men panel* is an indicator for whether the panel consists of men only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 7: Effect of conviction and jail time on appeal probability

	Probability of appeal	
	(1)	(2)
Guilty	0.379***	
	(0.017)	
Jail time		0.036***
		(0.003)
Observations	61302	59541
Court and crime x year FEs	X	X
Controls	X	X
Mean dep.var.	0.190	0.195

Notes. In all columns, the dependent variable is an indicator for whether the panel's decision is appealed against. *Guilty* is an indicator for whether the judicial panel pronounced a guilty verdict. *Jail time* measures the imprisonment duration the defendant is sentenced for. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 8: Effect of gender composition of the panel on probability of appeal

	Probability of appeal			
	(1) Unconditional	(2) Unconditional	(3) Conditional	(4) Conditional
Mixed gender panel	-0.039**		-0.033	
	(0.018)		(0.027)	
All-women panel		0.012		-0.026
		(0.017)		(0.027)
All-men panel		0.130***		0.227***
		(0.044)		(0.051)
Observations	61302	61302	28778	28778
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	0.190	0.190	0.394	0.394

Notes. In all columns, the dependent variable is an indicator for whether the panel's decision is appealed against. In columns (1) and (2), the analysis is done on the whole sample; in columns (3) and (4), the analysis is restricted to conviction decisions only. *Mixed gender panel* is an indicator for whether the gender composition of the panel is mixed. *All-men panel* is an indicator for whether the panel consists of men only. *All-women panel* is an indicator for whether the panel consists of women only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 9: Effect of gender composition of the panel on decision quality

	Decision quality			
	(1) Unconditional	(2) Unconditional	(3) Conditional	(4) Conditional
Mixed gender panel	0.029**		0.035*	
	(0.014)		(0.021)	
All-men panel		-0.107***		-0.197***
		(0.035)		(0.044)
All-women panel		-0.005		0.014
		(0.012)		(0.019)
Observations	61075	61075	28742	28742
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	0.883	0.883	0.755	0.755

Notes. In all columns, the dependent variable is an indicator for whether the panel takes a *correct decision*, defined as a decision that is either not appealed against or it is appealed and confirmed with a difference in jail time in absolute value below the average. In columns (1) and (2), the analysis is done on the whole sample; in columns (3) and (4), the analysis is restricted to conviction decisions only. *Mixed gender panel* is an indicator for whether the gender composition of the panel is mixed. *All-men panel* is an indicator for whether the panel consists of men only. *All-women panel* is an indicator for whether the panel consists of women only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 10: Effect of gender composition by case complexity

	Guilty		Length of trial		Quality (cond.)	
	(1)	(2)	(3)	(4)	(5)	(6)
A: More complex cases						
Mixed gender panel	-0.098***		0.296		0.046*	
	(0.029)		(0.239)		(0.028)	
All-women panel		0.117***		-0.182		0.014
		(0.031)		(0.274)		(0.022)
All-men panel		0.036		-0.674**		-0.240***
		(0.048)		(0.326)		(0.063)
Observations	47327	47327	47327	47327	19940	19940
Court and crime x year FEs	X	X	X	X	X	X
Controls	X	X	X	X	X	X
Mean dep.var.	0.422	0.422	3.374	3.374	0.781	0.781
B: Less complex cases						
Mixed gender panel	-0.019		0.049		0.010	
	(0.017)		(0.119)		(0.019)	
All-women panel		0.028		-0.071		0.005
		(0.019)		(0.141)		(0.021)
All-men panel		-0.012		0.026		-0.063
		(0.029)		(0.111)		(0.042)
Observations	13962	13962	13962	13962	8781	8781
Court and crime x year FEs	X	X	X	X	X	X
Controls	X	X	X	X	X	X
Mean dep.var.	0.630	0.630	1.514	1.514	0.696	0.696

Notes. The dependent variable in columns (1) and (2) is an indicator for whether the panel convicts the defendant; in columns (3) and (4) is the length of jail time in years; and in columns (5) and (6) is an indicator for whether the decision taken is correct (conditional on conviction). Panel A reports the results for complex cases, while panel B reports the results for simple cases. *Mixed gender panel* is an indicator for whether the gender composition of the panel is mixed. *All-men panel* is an indicator for whether the panel consists of men only. *All-women panel* is an indicator for whether the panel consists of women only. All regressions include court and crime type-by-year fixed effects. Controls include gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 11: Effect of gender composition on written opinions

	Time to write (days)		Length (words)		N. maxims	
	(1)	(2)	(3)	(4)	(5)	(6)
Mixed gender panel	54.019***		-9209.492***		-2.132*	
	(19.397)		(2893.104)		(1.186)	
All-men panel		-71.821***		22440.227***		6.994***
		(27.261)		(5080.692)		(2.471)
All-women panel		-43.804**		1500.645		-0.723
		(20.326)		(2548.218)		(0.801)
Observations	14562	14562	15167	15167	15196	15196
Court and crime x year FEs	X	X	X	X	X	X
Controls	X	X	X	X	X	X
Mean dep.var.	119.933	119.933	20402.466	20402.466	6.181	6.181

Notes. The dependent variable in columns (1) and (2) is the number of days taken to write the opinions; in columns (3) and (4) is the number of words in the written opinion; and in columns (5) and (6) is the number of references to a Court of Cassation maxim. *Mixed gender panel* is an indicator for whether the gender composition of the panel is mixed. *All-men panel* is an indicator for whether the panel consists of men only. *All-women panel* is an indicator for whether the panel consists of women only. The analysis excludes the Court of Naples, for which data on written opinions is not available. All regressions include court and crime type-by-year fixed effects. Controls include gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

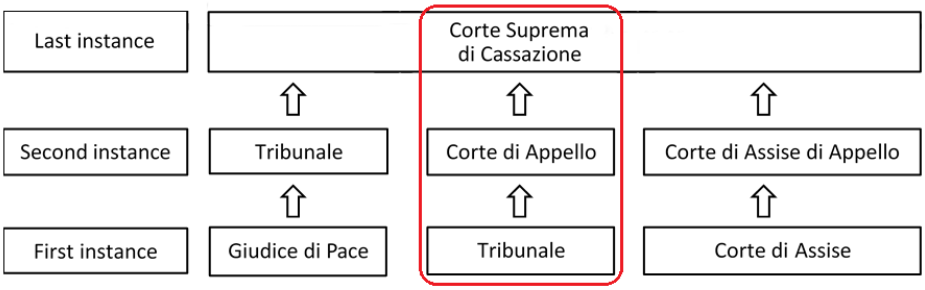
Table 12: Effect of mixed-gender experience by gender composition of the panel

	(1) Guilty	(2) Length of trial	(3) Quality (cond)
All-women panel	0.101**	-0.105	0.013
	(0.048)	(0.343)	(0.034)
High mixed-gender exp.	-0.007	-0.017	0.009
	(0.040)	(0.284)	(0.030)
All-women panel × High mixed-gender exp.	-0.029	-0.311	0.049
	(0.065)	(0.416)	(0.049)
All-men panel	0.049	-0.814**	-0.240***
	(0.055)	(0.330)	(0.053)
All-men panel × High mixed-gender exp.	-0.095	1.092**	0.192***
	(0.087)	(0.518)	(0.071)
Observations	61302	61302	28742
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.470	2.950	0.755
Experienced - inexperienced F3	-0.036	-0.328	0.058
Experienced - inexperienced F0	-0.102	1.075**	0.201***

Notes. The dependent variable in columns (1) is an indicator for whether the panel convicts the defendant; in columns (2) is the duration of the trial; and in columns (3) is an indicator for whether the decision taken is correct (conditional on conviction). *High mixed-gender exp.* is an indicator for whether at least two judges have a higher than average share of cases ruled in mixed-gender teams within the previous 15 cases. Controls include gender and nationality of the defendant, number of crimes, number of defendants, and judges' region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

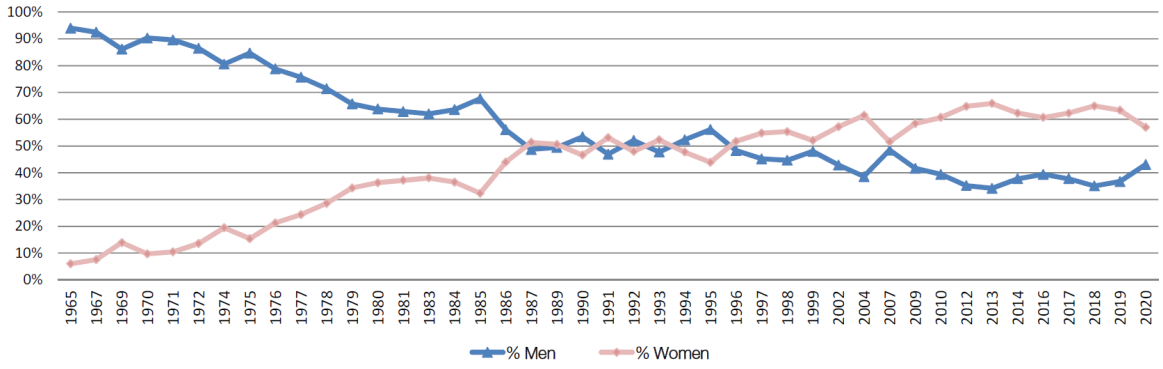
A Other Tables and Figures

Figure A1: Structure of the Italian judiciary



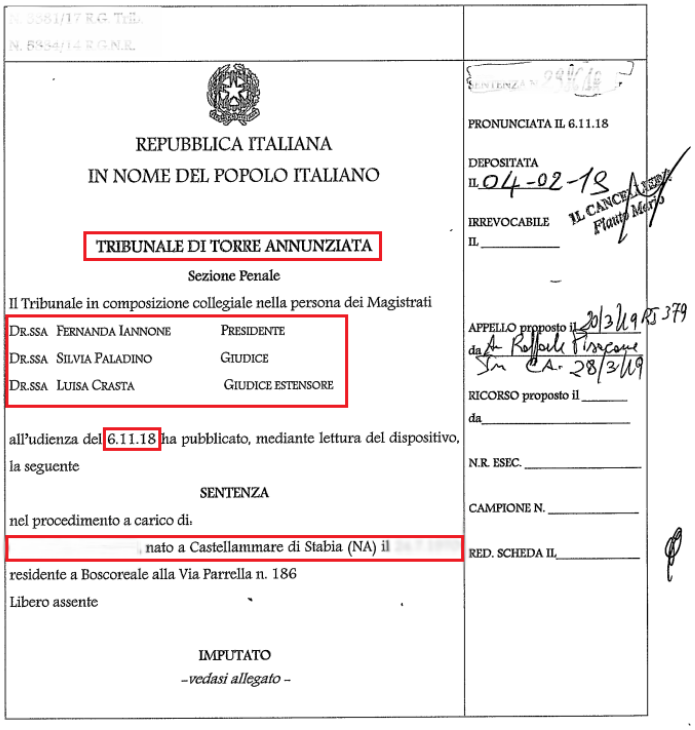
Notes. Structure of the Italian judiciary. The focus of our analysis is highlighted in red.

Figure A2: Gender composition of the Italian judiciary over time



Notes. New magistrates by year and gender. Source: Consiglio Superiore della Magistratura.

Figure A3: Overview of a sentence



REPUBBLICA ITALIANA

IN NOME DEL POPOLO ITALIANO

TRIBUNALE DI TORRE ANNUNZIATA

Sezione Penale

Il Tribunale in composizione collegiale nella persona dei Magistrati

DR.SSA FERNANDA IANNONE PRESIDENTE

DR.SSA SILVIA PALADINO GIUDICE

DR.SSA LUISA CRASTA GIUDICE ESTENSORE

all'udienza del 6.11.18 ha pubblicato, mediante lettura del dispositivo, la seguente

SENTENZA

nel procedimento a carico di:

nato a Castellammare di Stabia (NA) il

residente a Boscoreale alla Via Parrella n. 186

Libero assente

IMPUTATO

-vedasi allegato -

PRONUNCIATA IL 6.11.18

DEPOSITATA IL 04-02-19

IRREVOCABILE IL

APPELLO proposto il 20/3/19 RS 379

da

RICORSO proposto il

da

N.R. ESEC.

CAMPIONE N.

RED. SCHEDA IL

(a) Court, date, panel, defendant

IMPUTATO

per il delitto p. e p. dagli artt. 216 comma 1, nn. 1 e 2,219 comma 2 n. 1, R.D. 267/1942,223, comma 1, R.D. 267/1942, perché, quale titolare della ditta individuale "

dichiarata fallita con sentenza del Tribunale di Torre Annunziata del 27.3.2014, sottraeva, allo scopo di procurare a sé o ad altri un ingiusto profitto o di recare pregiudizio ai creditori, parte dei libri e delle altre scritture contabili della predetta società e comunque le tenevano in modo da non rendere possibile la ricostruzione del patrimonio e del movimento degli affari, nonché distraeva n. 7 veicoli della società meglio descritti di seguito, alienandoli a terzi nel 2012 e 2013:

(b) Crime committed

P.Q.M.

Letti gli artt. 533 e 535 c.p.p. dichiara colpevole dei reati ascrittigli riconosciute le circostanze attenuanti generiche prevalenti sulla contestata aggravante, condanna alla pena di anni due di reclusione, oltre al pagamento delle spese processuali.

Pena sospesa.

Letto l'art. 216 ult.co. l.f., dichiara l'imputato inabilitato per anni dieci all'esercizio un'impresa commerciale e incapace per la stessa durata ad esercitare uffici direttivi pres qualsiasi impresa per la durata di anni dieci.

Leto l'art. 544 c.p.p. fissa in giorni novanta il termine per il deposito della motivazione del sentenza.

Torre Annunziata, 6.11.2018

Il giudice estensore

Il Presidente

TRIBUNALE DI TORRE ANNUNZIATA
CANCELLERIA PENALE

(c) Sentencing outcome

Notes. The figure shows an overview of a sentence, with the parts containing the data of interest highlighted in red. Sentences were made available in paper format and have been scraped using Python algorithms.

Table A1: Types of crime

	Freq.	Percent (%)
against property and cultural heritage	23248	37.92
against public order and the state	11437	18.66
against public trust	7485	12.21
against the person	5983	9.76
against the public administration	5664	9.24
against public safety and the environment	4349	7.09
against the economy	1305	2.13
against the administration of justice	1074	1.75
against the family	507	0.83
against public morality and sentiment	250	0.41
Total	61302	100.00

Notes. This table reports some descriptive statistics on the types of crime ruled against by the panels in our sample. The classification of crimes follows the Italian Penal Code.

Table A2: Raw and residual variation in share of different panels

	Mean	St.Dev.
A. Mixed-gender panel		
Raw variable	0.746	0.435
Residuals after removing court and crime x year FE	-0.000	0.413
B. All-women panel		
Raw variable	0.192	0.394
Residuals after removing court and crime x year FE	-0.000	0.378
C. All-men panel		
Raw variable	0.062	0.241
Residuals after removing court and crime x year FE	-0.000	0.227

Notes.. This table reports the raw and residual (net of court and crime type-by-year fixed effects) variation in the share of all-men panels in Panel A; mixed-gender panels in Panel B; and all-women panels in Panel C.

Table A3: Inclusion\exclusion of controls

	(1)	(2)	(3)	(4)
A. Guilty				
All-women panel	0.115***	0.101***	0.108***	0.102***
	(0.026)	(0.024)	(0.025)	(0.025)
All-men panel	0.059	0.059	0.045	0.025
	(0.038)	(0.039)	(0.036)	(0.039)
Observations	61302	61302	56301	61302
Court FEs	X	X	X	X
Crime x year FEs		X	X	X
Judge-level controls			X	X
Case-level controls				X
Mean dep.var.	0.470	0.470	0.477	0.470
B. Length of trial				
All-women panel	0.053	-0.088	-0.155	-0.131
	(0.250)	(0.219)	(0.220)	(0.219)
All-men panel	-0.458*	-0.470**	-0.533**	-0.485*
	(0.242)	(0.208)	(0.247)	(0.247)
Observations	61302	61302	56301	61302
Court FEs	X	X	X	X
Crime x year FEs		X	X	X
Judge-level controls			X	X
Case-level controls				X
Mean dep.var.	2.950	2.950	2.942	2.950
C. Quality (cond.)				
All-women panel	0.060**	0.029	0.025	0.014
	(0.026)	(0.018)	(0.020)	(0.019)
All-men panel	-0.206***	-0.195***	-0.183***	-0.197***
	(0.047)	(0.044)	(0.042)	(0.044)
Observations	28746	28742	26797	28742
Court FEs	X	X	X	X
Crime x year FEs		X	X	X
Judge-level controls			X	X
Case-level controls				X
Mean dep.var.	0.755	0.755	0.754	0.755

Notes. The dependent variable in Panel A is an indicator for whether the panel convicts the defendant; in Panel B is the duration of the trial; and in Panel C is an indicator for whether the decision taken is correct (conditional on conviction). *All-women panel* is an indicator for whether the panel consists of men only. *All-men panel* is an indicator for whether the panel consists of women only. All regressions include court fixed effects. Columns (2)-(4) include also crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A4: Effect of age and regional composition of the panel

	(1) Guilty	(2) Length of trial	(3) Quality (cond)
A. Age diversity			
All-young panel	-0.033	-0.683**	0.033
	(0.030)	(0.279)	(0.026)
All-old panel	0.002	0.112	-0.025
	(0.044)	(0.239)	(0.030)
Observations	56301	56301	26797
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.477	2.942	0.754
B. Regional diversity			
All-north panel	-0.016	0.116	0.054
	(0.041)	(0.234)	(0.037)
All-south panel	-0.072	0.401	-0.044
	(0.047)	(0.284)	(0.028)
Observations	56301	56301	26797
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.477	2.942	0.754

Notes. The dependent variable in columns (1) and (2) in an indicator for whether the panel convicts the defendant; in columns (3) and (4) is the length of jail time in years; and in columns (5) and (6) is an indicator for whether the decision taken is correct (conditional on conviction). Panel A focuses on age diversity, while Panel B focuses on regional diversity. *All-old panel* is an indicator of whether all judges in the panel are above the median age. *All-young panel* is an indicator of whether all judges in the panel are below the median age. *All-north panel* is an indicator of whether all judges were born in central or northern Italy. *All-south panel* is an indicator of whether all judges were born in southern Italy. The omitted category in panels A and B is a mixed-age panel and a mixed-south-north panel, respectively. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A5: Model with judge fixed effects

	(1) Guilty	(2) Length of trial	(3) Quality (cond)
A. Women			
Same-gender panel	0.078***	0.100	-0.002
	(0.025)	(0.120)	(0.016)
Observations	105252	105252	51023
Court and crime x year FEs	X	X	X
Judge FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.485	3.020	0.779
B. Men			
Same-gender panel	-0.029	0.169	-0.131***
	(0.025)	(0.157)	(0.026)
Observations	78666	78666	35211
Court and crime x year FEs	X	X	X
Judge FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.448	2.857	0.720

Notes. The dependent variable in column (1) is an indicator for whether the panel convicts the defendant; in column (2) is the duration of the trial; and in column (3) is an indicator for whether the decision taken is correct (conditional on conviction). *All-women panel* is an indicator for whether the panel consists of men only. *All-men panel* is an indicator for whether the panel consists of women only. Controls include proxies of case complexity (number of defendants and crimes), gender and nationality of the defendant. Standard errors clustered at the judge level in parentheses. * $p< 0.1$, ** $p< 0.05$, *** $p< 0.01$.

Table A6: Different classification of crimes

	(1) Guilty	(2) Length of trial	(3) Quality (cond)
All-women panel	0.095***	-0.129	0.017
	(0.023)	(0.210)	(0.018)
All-men panel	0.025	-0.439*	-0.194***
	(0.038)	(0.239)	(0.040)
Observations	61302	61302	28742
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.470	2.950	0.755

Notes. The dependent variable in column (1) is an indicator for whether the panel convicts the defendant; in column (2) is the duration of the trial; and in column (3) is an indicator for whether the decision taken is correct (conditional on conviction). *All-women panel* is an indicator for whether the panel consists of men only. *All-men panel* is an indicator for whether the panel consists of women only. All regressions include court and crime type-by-year fixed effects. These regressions are estimated including a more granular classification that separates the most common offenses into distinct categories: mafia-type crimes, robbery, money laundering, extortion, personal injury, fraud, sexual violence, and fraudulent bankruptcy. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A7: Excluding specific types of crime

	Excl. crimes against									
	(1) the P.A.	(2) admin of justice	(3) public order	(4) public safety	(5) public trust	(6) the economy	(7) public morality	(8) the family	(9) the person	(10) the property
A. Guilty										
All-women panel	0.094***	0.103***	0.115***	0.103***	0.074***	0.101***	0.102***	0.103***	0.107***	0.120***
	(0.025)	(0.026)	(0.026)	(0.026)	(0.024)	(0.025)	(0.025)	(0.025)	(0.028)	(0.029)
All-men panel	0.027	0.027	0.037	0.036	-0.003	0.024	0.025	0.024	0.028	0.016
	(0.040)	(0.039)	(0.037)	(0.040)	(0.036)	(0.039)	(0.039)	(0.040)	(0.044)	(0.044)
Observations	55638	60228	49865	56953	53817	59997	61052	60795	55319	38054
Mean dep.var.	0.482	0.472	0.463	0.471	0.484	0.474	0.471	0.469	0.455	0.447
B. Length of trial										
All-women panel	-0.237	-0.135	-0.059	-0.208	-0.117	-0.066	-0.130	-0.130	-0.104	-0.104
	(0.225)	(0.219)	(0.211)	(0.211)	(0.201)	(0.211)	(0.219)	(0.221)	(0.234)	(0.280)
All-men panel	-0.456*	-0.483*	-0.547**	-0.451*	-0.458**	-0.499**	-0.485*	-0.495**	-0.536**	-0.391
	(0.255)	(0.247)	(0.240)	(0.249)	(0.227)	(0.241)	(0.247)	(0.249)	(0.272)	(0.293)
Observations	55638	60228	49865	56953	53817	59997	61052	60795	55319	38054
Mean dep.var.	2.946	2.952	2.819	2.989	2.865	2.921	2.948	2.964	3.057	3.060
C. Quality (cond.)										
All-women panel	0.013	0.014	0.012	0.015	0.017	0.014	0.014	0.014	0.019	0.006
	(0.020)	(0.019)	(0.020)	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.021)	(0.022)
All-men panel	-0.191***	-0.198***	-0.194***	-0.199***	-0.167***	-0.201***	-0.197***	-0.198***	-0.212***	-0.217***
	(0.042)	(0.044)	(0.045)	(0.044)	(0.046)	(0.045)	(0.044)	(0.045)	(0.051)	(0.046)
Observations	26790	28368	23060	26803	25998	28374	28694	28450	25157	16984
Mean dep.var.	0.755	0.756	0.749	0.757	0.745	0.754	0.755	0.756	0.765	0.761

Notes. The dependent variables in panels A, B and C are an indicator for whether the panel convicts the defendant, the length of jail time in years, and an indicator for whether the decision taken is correct (conditional on conviction), respectively. In columns (1)-(10) we exclude one specific type of crimes at the time. *All-women panel* is an indicator for whether the panel consists of women only. *All-men panel* is an indicator for whether the panel consists of men only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A8: Gender and ethnicity of defendants

	Gender of defendant		Nationality of defendant	
	(1)	(2)	(3)	(4)
	Male	Female	Italian	Foreign
A. Guilty				
All-women panel	0.101***	0.136***	0.105***	0.105***
	(0.025)	(0.043)	(0.026)	(0.034)
All-men panel	0.028	0.082	0.054	-0.042
	(0.037)	(0.054)	(0.038)	(0.055)
Observations	49814	6123	48873	6981
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	0.480	0.436	0.463	0.570
B. Length of trial				
All-women panel	-0.175	0.006	-0.127	-0.162
	(0.216)	(0.300)	(0.231)	(0.196)
All-men panel	-0.450	-0.768***	-0.633**	0.142
	(0.274)	(0.238)	(0.286)	(0.243)
Observations	49814	6123	48873	6981
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	2.943	2.982	3.044	2.167
C. Quality (cond.)				
All-women panel	0.002	0.011	0.022	-0.068
	(0.021)	(0.036)	(0.022)	(0.045)
All-men panel	-0.188***	-0.258***	-0.194***	-0.193***
	(0.045)	(0.092)	(0.046)	(0.070)
Observations	23890	2659	22625	3961
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	0.752	0.766	0.753	0.752

Notes. The dependent variable in Panel A is an indicator for whether the panel convicts the defendant; in Panel B is the duration of the trial; and in Panel C is an indicator for whether the decision taken is correct (conditional on conviction). *All-women panel* is an indicator for whether the panel consists of men only. *All-men panel* is an indicator for whether the panel consists of women only. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant (only in columns (3)-(4) and (1)-(2), respectively) and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A9: Unpacking the mixed-gender teams

	(1) Guilty	(2) Length of trial	(3) Quality (cond.)
All-men panel	-0.083**	-0.349	-0.213***
	(0.039)	(0.286)	(0.048)
2M1F panel	-0.129***	0.152	-0.022
	(0.029)	(0.251)	(0.025)
1M2F panel	-0.082***	0.115	-0.008
	(0.028)	(0.229)	(0.020)
Observations	61302	61302	28742
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.470	2.950	0.755

Notes. The dependent variable in columns (1) is an indicator for whether the panel convicts the defendant; in column (2) is the length of jail time in years; and in column (3) is an indicator for whether the decision taken is correct (conditional on conviction). *All-men panel* is an indicator for whether the panel consists of women only. *2M1F* is an indicator for whether the panel consists of two men and one woman. *1M2F* is an indicator for whether the panel consists of one man and two women. All regressions include court and crime type-by-year fixed effects. Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age and region of birth. Standard errors clustered at the panel level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.