Gender diversity and teams' decision-making

Maddalena Ronchi Viola Salvestrini*

August 31, 2025

Abstract

We show that gender composition affects both the nature and quality of team decisions, assembling a novel dataset on collegial rulings from first- and second-instance criminal courts in Italy. Leveraging the quasi-random allocation of both judges and cases to panels, we find that all-female teams convict more often and issue the most appeal-proof rulings (i.e., unappealed or upheld on appeal). In contrast, all-male teams convict less frequently and have the lowest rate of appeal-proof convictions, while mixed-gender teams resemble all-male teams in conviction rates but nearly match the conviction quality of all-female teams. That higher conviction rates coincide with fewer reversals in all-female teams rules out differences in judicial leniency across teams as the main driver. That mixed-gender teams convict less often than all-female teams, despite similar conviction quality, points instead to coordination costs. We formalize these insights in a model of diagnostic skills in which individuals differ in both preferences and their ability to distinguish guilt from innocence, stemming from effort or raw talent. Model estimates, together with reduced-form evidence on conviction and overturn patterns, judicial written opinions, and deliberation times, show that frictions arise in gender-mixed teams and that variations in effort, rather than preferences alone, drive our results.

JEL codes: J16, J82, M5, K4

*Ronchi: Northwestern University, IZA, and CESifo. Salvestrini: Bocconi University and IZA. We thank François Gerard and Anna Raute for their support and guidance. For very helpful comments and suggestions we thank Manuel Bagues, Jan Bakker, Jordi Blanes i Vidal, Wouter Dessein, Adriano De Falco, Elisa Facchetti, Alex Frenkel, Alfonso Gambardella, Andrea Ichino, Mikkel Mertz, Elia Moracci, Massimo Morelli, Barbara Petrongolo, Nicola Persico, Fabiano Schivardi, Martina Zanella, as well as seminar participants at Bocconi University, Einaudi Instutute for Economics and Finance (EIEF), European University Institute, Queen Mary University, Oxford University, LMU Munich, Regensburg University, and conference participants at the CEPR Paris Symposium; 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 financial support from Einaudi Institute for Economic and Finance and Institute for Human Studies.

1 Introduction

The sharp rise in women's labor force participation, together with the growing occupational convergence between men and women, has made gender diversity within teams increasingly common. From a theoretical perspective, diversity may enhance performance through complementarities in skills, knowledge, and perspectives but can also raise coordination and communication costs (Lazear, 1999; Hong and Page, 2001; Prat, 2002; Lazear and Shaw, 2007; Che and Kartik, 2009). Since men and women differ in preferences, attitudes, and social values (e.g. Chattopadhyay and Duflo, 2004; Croson and Gneezy, 2009; Falk et al., 2018; Anwar et al., 2019), a team's gender composition may shape both the nature and the quality of its decisions in ambiguous ways. This makes the impact of gender diversity on decision-making an empirical question – one that is especially important in modern economies, where teams are the primary unit of decision-making (Deloitte, 2016)[1] – yet causal evidence remains scarce.

In this paper, we exploit a unique institutional setting – the Italian criminal justice system – to study the causal effect of the gender composition of a team on decision-making and decision quality. Several features make this context particularly well-suited for answering this question. First, serious criminal offenses are tried by a collegial panel of three judges, with both judges and cases randomly assigned to panels within courts.[2] This exogenous variation in team gender-mix within-court allows us to causally estimate whether gender composition affects judicial ruling. Second, while defining and measuring the intrinsic quality of team decisions is challenging in most real-life settings, the hierarchical structure of the judiciary allows us to use the appellate process and patterns of sentence reversal as a natural benchmark for assessing the performance of judicial teams (Alesina and La Ferrara, 2014). Third, in most developed countries, including Italy, the judiciary is relatively gender-balanced, providing sufficient statistical power to study all-male, all-female, and mixed-gender teams separately.[3]

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

[2] Specifically, constitutional provisions require the Italian judicial system to allocate both judges and cases to panels through quasi-random procedures that rule out discretion or manipulation. This design ensures that judges are not assigned to teams based on gender, and that the cases given to collegial panels do not depend on a team's gender composition. We describe the assignment rules in greater detail in Section 2.3 and provide formal tests in Section 4.2.

[3] For example, the most recent data from the Federal Judicial Center show that in the U.S. almost 40% of District Court judges are female. Comparable figures from the United Nations Economic Commission for Europe (UNECE) indicate that women represent about 50% of judges in Canada, Germany, Italy, and Norway, and 65% in France.

Leveraging these unique features, we answer three main questions. First, we examine the causal relationship between the gender composition of judicial teams and the *type* of decisions they take – measured by conviction rates – as well as their *accuracy* – measured by "appeal-proof" decisions, i.e., rulings that are either unappealed or upheld on appeal.[4] Second, we determine whether observed differences across teams reflect differences in *preferences* (leniency) or *skills* (talent and/or effort); to do so, we develop and structurally estimate a theoretical model of judicial team decisions and test its predictions using rich data on cases and judges' written opinions. Finally, combining the result of our reduced-form analysis with those of the estimation, we shed light on the relation between gender diversity and frictions in teams' decision-making.

To answer these questions, we construct a novel database containing the universe of collegial ruling sentences from a sample of first-instance criminal courts in Italy between 2004 and 2021, merged with the corresponding second-instance sentences. To our knowledge, no comparable digitized dataset exists at the court level, making its construction crucial for enabling our analysis. Specifically, we obtain data from four courts in Central and Southern Italy and the corresponding Courts of Appeal.[5] The sentences were made available by the presidents of each court in paper format, which we then digitized and processed using a combination of OCR and text-parsing algorithms. In total, the dataset includes more than 65,000 judicial decisions.

Using a specification with court fixed effects and crime-by-year fixed effects – thus comparing teams within the same court ruling on similar crimes – we first show that judicial teams reach different sentencing outcomes depending on their gender composition. All-women teams are 7.1 percentage points (16.7%) more likely to convict than all-men teams, and 5.7–9.3 percentage points (13.3–21.8%) more likely to convict than mixed-gender teams, depending on whether the latter include one or two women. The estimates for mixed-gender teams with one versus two women are not statistically different from each other, and conviction rates for all-men and mixed-gender teams are also statistically indistinguishable. Thus, for the same type of crime, all-women teams

[4] As in Alesina and La Ferrara (2014), using patterns of sentence reversal as a measure of quality of judicial decisions rests on the assumption that higher courts can only improve upon the accuracy of first sentencing. We further discuss this assumption and provide empirical evidence in its favor in later Sections.

[5] 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 vary in both size – Naples is the second-largest court in Italy, Florence is medium-sized, and the other two are smaller – and location, with Florence in the Center-North and the others in the South. This diversity reassures us that our results are not specific to a particular type of court or area.

convict more often than any other team type, whereas all-men and mixed-gender teams exhibit similar conviction rates. We do not find that all-women teams impose harsher sentences: conditional on conviction, they impose prison terms that are, on average, 12 months shorter than those imposed by all-men teams and 6 months shorter than those from mixed-gender teams.

Second, we exploit the hierarchical structure of criminal courts to estimate whether a team's gender composition influences the accuracy of its decisions. We begin by examining the *perceived accuracy* of a ruling, proxied by its likelihood of being appealed. Our analysis shows that decisions issued by all-male teams are significantly more likely to be appealed than those of any other team type, while mixed-gender and all-female teams exhibit similar appeal rates. Importantly, this pattern persists when we restrict the sample to conviction decisions, implying that it is not driven by differences in conviction rates across teams, and when we control for sentence length, indicating that the higher appeal rate of all-male teams is not merely a response to the longer sentences they impose. We then measure *effective accuracy* by examining whether a ruling is confirmed or reversed by the Court of Appeal. All-men teams are 17.7 percentage points (24.8%) less likely than all-female teams to issue decisions that withstand appeal. The presence of at least one woman substantially improves outcomes: mixed-gender teams' coefficient is only about one-third that of all-male teams, thus their effective accuracy is just. 7–9% lower than that of all-female teams, which achieve the highest accuracy. In robustness checks, we show that our results on the nature and quality of decisions reflect the impact of gender rather than other judicial characteristics correlated with gender (e.g. age, experience, or geographical mobility), and are not driven by particular crime types or by the gender of the defendant.

Taken together, our findings show that all-female teams both convict more often and issue the most appeal-proof convictions, suggesting they are more effective at distinguishing guilt from innocence. By contrast, all-male teams convict less frequently and have the lowest rate of appeal-proof convictions. Mixed-gender teams resemble all-male teams in conviction rates but achieve conviction quality nearly as high as all-female teams. Intuitively, the fact that higher conviction rates coincide with fewer reversals rules in all-female teams out differences in judicial leniency across teams with different gender composition as the main driver of our results. If gender composition affected only the propensity to convict, but not the evaluation of the case, harsher teams would convict more defendants regardless of guilt and, as a result, should expe-

rience more reversals in appeal. Instead, all-female teams convict more *and* are more accurate. This pattern points to differences in teams' ability to better distinguish guilt from innocence rather than differences in leniency, as the most plausible explanation.[6] Finally, the fact that mixed-gender teams deliver convictions of similar quality to all-female teams but at lower rates suggests the presence of frictions: in gender-diverse teams, it is harder to translate a signal of guilt into a unanimous conviction than in homogeneous female teams.

We rationalize these intuitions with a stylized model of diagnostic skill, adapted from Chan et al. (2022) to a setting in which decisions are made by teams. Judicial panels 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).[7] More precise signals increase conviction rates – because they lower the expected cost of a false positive – and reduce errors, while imprecise signals reduce conviction rates and generate more errors. Signal precision depends on the team's diagnostic skill, which we model as 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. However, while low effort also reduces the share of appeal-proof convictions, frictions in consensus-building primarily reduce conviction rates without lowering decision quality.

We structurally estimate the model and find that female judges experience a greater disutility from making mistakes than male judges. Within our model, this higher cost of errors leads them to exert more effort, which, consistent with our reduced-form results, translates into higher conviction rates and more appeal-proof convictions in all-women teams compared to all-men teams. We refer to this aversion to errors as the diligence parameter, capturing any factor that makes judicial mistakes feel more costly (such as personal preferences, professional norms, or external scrutiny and

[6]Thus, a standard research design that uses the random assignment of cases to judicial panels as an instrument for conviction would fail in this setting. The monotonicity assumption underlying such an approach requires that all judges rank cases in the same way but apply different conviction thresholds. This, however, implies that all panels have the same ability to distinguish guilt from innocence - an implication that our reduced-form results reject.

[7]In criminal law, the *Blackstone's ratio* principle holds 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.

sanctions).[8] We also estimate that mixed-gender teams face greater frictions than homogeneous teams, possibly due to the challenges of reaching an agreement amidst diverse viewpoints. This helps explain why their conviction rates are lower than those of all-female teams, even though our model predicts that women exert more effort when assigned to mixed-gender panels.[9]

While our reduced-form and structural results indicate that preferences are not the main driver of our findings, women's greater ability to distinguish guilt from innocence could, in principle, reflect either innate talent or effort. In our theoretical framework, we model this ability as effort. Although disentangling the two is inherently difficult, we show additional results consistent with an effort-based channel, as well as present further evidence of frictions in mixed-gender teams. First, we digitized the official rankings from national entry examinations for the judiciary and find no performance differences across gender, indicating that, at entry, women are not positively selected.[10] Second, we find that all-male teams decide faster than both mixed-gender and all-female teams, which take equally long. This yields two insights. First, women's higher decision quality is unlikely to be explained by innate talent alone: if that were the case, they could achieve better outcomes in the same time as men. Instead, they achieve better outcomes *and* spend more time per case, consistent with greater effort. Second, the fact that mixed and all-female teams deliberate equally long but reach different conviction rates is consistent with the presence of frictions in mixed teams, where part of the deliberation is absorbed by difficulties in reaching consensus. Finally, we examine the written reasoning that judicial teams must provide for every verdict and construct several measures of writing quality and conciseness. Across all measures, all-male teams produce judgments that are less clear and concise than those of other teams. By contrast, all-female and mixed-gender teams tend to issue shorter opinions but spend more time writing them, suggesting greater effort in ensuring clarity and effective communication.

This paper contributes to the literature on the gender composition of teams and decision-making bodies, which has largely examined three dimensions. The first strand studies whether the share of women in a group affects women's influence on group de-

[8] Finding that women experience higher disutility from professional mistakes aligns with evidence that they are punished more harshly for the same errors and face weaker support from colleagues despite similar performance (Gagliarducci and Paserman, 2012; Egan et al., 2022; Sarsons, 2024).

[9] This is because optimal effort is decreasing in the other team member's aversion to mistakes.

[10] In Italy, judges are not elected; entry into the profession is gained through a nationwide public competition. We discuss this in more detail when reviewing the institutional setting in Section 2.

cisions (Born et al., 2022; Karpowitz et al., 2024; Radbruch and Schiprowski, 2024; Rebolledo et al., 2024). Unlike this literature, which examines how gender composition shapes information aggregation or the relative influence of group members, we focus directly on its impact on team performance. This is possible because our setting provides a natural benchmark for the accuracy and quality of team decisions. Leveraging this feature, we can assess whether differences in outcomes across gender-mixed teams are driven solely by preferences or also by differences in skills. Moreover, this literature generally finds that women's influence grows as their numbers increase from a low baseline until they become the majority. By contrast, we find no differences between mixed-gender teams where women are the majority or the minority. This suggests a different channel: the presence of at least one member of a different gender alters team dynamics, introducing coordination costs relative to all-female teams, but greater effort and more accurate decisions relative to all-male teams.

A second strand examines whether gender diversity among decision-making bodies (e.g. hiring committees, companies boards) benefits women in the populations they evaluate or oversee (Bagues and Esteve-Volart, 2010; De Paola and Scoppa, 2015; Bagues et al., 2017; Bertrand et al., 2019; Deschamps, 2024). Unlike these papers, our focus is not on homophily between evaluators and evaluated – and, in our setting, we find no relationship between the gender composition of judicial panels and that of defendants. Moreover, these studies typically examine whether a higher female presence in a committee increases the number of women hired or promoted, but not whether it improves the quality of decisions.

A final, limited group of studies explores the relationship between team gender mix and performance. These studies come largely from lab or lab-in-the-field experiments, often using students as subjects (Apesteguia et al., 2012; Hoogendoorn et al., 2013; Hardt et al., 2024).[11] We complement this literature in two ways. First, we examine a real-world, high-stakes setting with teams of high-skill professionals, moving beyond the lab-in-the-field evidence that dominates existing work. Second, we contribute to this literature by providing empirical evidence on how gender differences in preferences and effort affects team decision-making and performance, therefore shedding light on

[11]Work on team performance has also considered other dimensions of diversity such as ethnicity, nationality, seniority, and prior collaborations (e.g. Hjort, 2014; Lyons, 2017; Chan, 2021; Chen, 2021; Marx et al., 2021). Advances have also been made on the theoretical side, with recent studies modeling whether and when diversity improves performance (recent examples include Che and Kartik, 2009; Kets and Sandroni, 2021; Hughes et al., 2023).

the costs and benefits of diversity in teams.

Additionally, 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 (e.g., 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 teams 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.

Finally, our results have important implications for the literature using the so-called "judges design," which exploits judges' decision propensities as an instrument to estimate the impact of judicial rulings on other outcomes of interest, including in criminal justice, bankruptcy, approval of patents and disability insurance (e.g. Maestas et al., 2013; Galasso and Schankerman, 2015; Dobbie et al., 2018; Bernstein et al., 2019). This design relies on a monotonicity assumption, which – as argued by Chan et al. (2022) and Mungan (2024a) – holds only if heterogeneity in judges' propensities is driven exclusively by differences in preferences (e.g., leniency) rather than in skills (e.g., talent or effort). Our findings further weaken the rationale for assuming average monotonicity when random assignment to judges is used for identification.

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 proposes 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 a unified, hierarchical body responsible for civil, criminal, administrative, tax, and audit cases. It includes ordinary courts, which hear most civil and criminal matters, specialized courts for administrative and tax disputes, and a separate military judiciary. Rooted in the *civil law* tradition, it relies on compre-

hensive legal codes and prioritizes the application of written law over precedent. The judiciary is constitutionally independent from the legislative and executive branches of government.[12]

Our analysis focuses on the criminal proceedings adjudicated by criminal courts (*Tribunali ordinari*). It therefore excludes all civil cases, as well as criminal cases handled by lower courts (*Giudici di pace*) – which deal with minor offenses – and the Courts of Assize (*Corte di Assise*) – which hear the most serious crimes such as terrorism and aggravated murder. Within criminal tribunals, the mode of adjudication depends on the severity of the charge: less serious cases are heard by a *single* judge, while more serious offenses are tried by a collegial team of *three judges*.[13] Our analysis focuses on decisions made by these three-judges teams.

2.1 The Appeal process

Judgments issued by the court of first instance can be appealed to the Court of Appeal (*Corte di Appello*), which acts as the second-instance court. These courts are composed of teams of three judges who are entirely independent from those involved in the original proceedings.[14] In the decade 2011-2021, nearly 38% of all criminal judgments delivered by the first-instance courts were appealed (Della Torre, 2022). Appeals may be initiated by either the defendant or the public prosecutor, however they are disproportionally advanced by the defendants: between 2012-2016, public prosecutors are responsible for only 5% of all appeals (Della Torre, 2022).[15] The Court of Appeal may confirm, modify, or overturn the original ruling, based on a full review of the

[12]The Italian and U.S. criminal justice systems diverge on several central dimensions, including legal foundations, trial structure, appellate review, jury involvement, and judicial reasoning. A comparative overview of the two systems is provided in Appendix B.

[13]The offenses under the jurisdiction of collegial teams are regulated by law (Art. 33 bis c.p.p.) and include crimes with maximum allowed 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; arms trafficking; criminal associations; usury; sexual violence; prostitution; racial, ethnic, and religious discrimination; and crimes against public safety and order.

[14]Courts of Appeal and first-instance tribunals are separate judicial bodies: each Court of Appeal is located in a district capital, while tribunals sit in individual cities, and their judges do not preside over the same cases. Appellate judges are randomly assigned to collegial teams, and cases are likewise allocated randomly across teams under rules set out in each Court of Appeal's organizational plan. The same mechanism applies in trial courts, as discussed in Section 2.3.

[15]When only the defendant appeals, the Court of Appeal cannot worsen the outcome of the initial judgment (art.579 c.p.p.). 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 cost of appealing.

case, reassessing both the evaluation and interpretation of the *evidence* (*giudizio di legittimità*) – such as the reconstruction of the events, or the credibility of witnesses – and the application and interpretation of the *law* (*giudizio di merito*). However, this review is subject to two important limitations. First, the appellate court can only rule on the specific parts of the judgment that have been challenged: issues that were not explicitly appealed are considered final and are not subject to review. Second, new evidence is generally not admissible, and witnesses are not retried. This implies that the appellate judges decide the case based on the same information available to the trial court.[16] This makes the Court of Appeal the final venue where errors in fact-finding – such as misjudging a defendant's guilt or innocence – can be addressed. Beyond this stage, cases may be brought before the Supreme Court of Cassation, but its review is limited to questions of law, not factual determination.[17]

2.2 Judges' appointment and career

The Italian judiciary is a high-skill, high-stakes, and standardized profession. Judges perform the same core functions, follow a common pay scale, and must meet identical formal qualifications.[18] There are no internal hierarchies: judges differ only in the nature of their assigned duties (e.g., trial vs. appellate functions)

Judges are not elected. Entry into the judiciary is gained through a nationwide public competition.[19] Candidates must hold a law degree and a two-year specialization

[16]The introduction of new evidence is permitted only in exceptional cases, when the appellate team deems it absolutely necessary for the decision. In practice, such requests are rare and granted infrequently.

[17]At the top of the judicial hierarchy is the Supreme Court of Cassation (Corte Suprema di Cassazione), Italy's highest appellate court. Unlike the courts of appeal, it rules only on matters of law, ensuring the correct and uniform application of the law nationwide. Based in Rome with national jurisdiction, the Court may confirm or annul lower rulings. If annulled, the case may be remanded for a new trial or, in rare cases, the sentence may be nullified altogether. A sentence confirmed by the Supreme Court of Cassation is final and definitive, and cannot be appealed further on the same grounds. In what follows, we focus on the first round of appeal, the only stage where factual errors can still be corrected through a full reassessment of both evidence and law.

[18]In addition to career magistrates, courts may appoint honorary judges (Giudici Onorari di Tribunale, or GOTs) on a fixed-term basis. Candidates must hold a law degree and at least five years of legal experience (e.g., as lawyers, notaries, or academics). GOTs serve renewable four-year terms, up to a maximum of 16 years. They typically preside over single-judge proceedings, but in cases of staff shortages or heavy backlogs may join judicial panels. In such cases, only one GOT may sit on a panel, and they are excluded from adjudicating the most serious criminal offenses (e.g., organized crime or sexual violence).

[19]The exam, administered by the Ministry of Justice, includes three written papers - on civil, criminal, and administrative law - followed by an oral examination covering all substantive and procedural areas of law. Candidates are assessed by a panel composed of senior magistrates, professors, and lawyers.

diploma for legal professions.[20] Successful entrants are initially appointed as judicial trainees for an 18-month program. At the end of the traineeship, the High Council of the Judiciary (*Consiglio Superiore della Magistratura*, CSM) publishes a list of vacancies in trial courts.[21] The CSM matches candidates to vacancies using a transparent formula that combines (i) the candidate's ranking, (ii) their expressed preferences, and (iii) organisational needs.[22]

Judges receive a fixed salary that increases with seniority, according to a structured pay scale. However, career progression is conditional on passing mandatory professional evaluations every four years, up to the 28th year of service. These evaluations – conducted by the CSM – assess three core dimensions: (i) productivity (e.g., number of cases handled); (ii) timeliness and diligence (e.g., clearance rates); and (iii) quality of legal reasoning (based on a sample of the judge written judgments).[23]

Beyond salary, evaluations – together with seniority – also determine promotion to higher courts (e.g., the Court of Appeal or the Supreme Court) and to administrative roles (e.g., president of a tribunal), as well as eligibility for transfers to more desirable locations. To safeguard independence and limit internal competition, pay progression is kept separate from hierarchical function.[24] Finally, judges cannot be dismissed or demoted except through formal disciplinary proceedings conducted by the CSM.

Similar to other developed countries, Italy has witnessed a transformation in the gender composition of its workforce, including in the judiciary. Appendix Figure A1 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

[20]This latter requirement is waived if candidates hold a Ph.D. in law or have completed an internship at the State Attorney's Office or a judicial office.

[21]The CSM is a constitutional body, chaired by the President of the Republic, that safeguards judicial independence and oversees the appointment, careers, and discipline of all ordinary judges and prosecutors, thereby insulating the judicial career path from political influence.

[22]A judge must normally remain in the first seat for at least three years before requesting a transfer to another court. Court vacancies are published twice a year, and selection is based on seniority, previous evaluation scores, family circumstances, and the staffing priorities of the destination court. No judge may serve in the court district where they completed significant prior legal practice or where close family members sit or practice, safeguarding both independence and impartiality.

[23]A positive evaluation permits advancement to the next salary class. A non-positive evaluation delays advancement by one year, while a negative evaluation delays it by two. A second negative review results in removal from the judiciary. Although a missed evaluation may temporarily pause salary progression, it does not reduce a judge's accumulated economic seniority: once the judge passes, the missed years are retroactively recognized in their pay grade.

[24]Judges in these directive or semi-directive roles do receive a functional allowance (*indennità di funzione*) as compensation for their added administrative responsibilities. This allowance increases total earnings during the term of service, but does not affect their placement on the formal salary scale.

rise in the share of female judges. In recent years, women constitute approximately 60 percent of new entrants in the judiciary. The fact that the judiciary has been a relatively gende-balanced profession since the 1980s allows us to observe the effects of both all-female and all-male teams within the same profession.

Note that although the gender composition of the judiciary has changed over time, we do not observe differential selection by gender over this period, as proxied by scores on the national entry examination. Figure 1 shows that the distribution of grades is nearly identical for men and women, who on average achieve the same score in every cohort (measured by year of birth or year of examination). We return to this point in Section 6.3.

2.3 Random assignment of judges and cases to teams

Italy's judicial system allocates judges and cases to panels through quasi-random procedures that exclude discretion or manipulation, creating a natural experiment to study the causal effect of team gender composition on team decision-making. We describe the assignment rules below and formally test the statistical exogeneity of both team and case allocation in Section 4.2.

Assignment of judges to judicial teams. Each court adopts an organisation plan – reviewed and approved every four years by the CSM – which sets out the criteria for assigning judges to three-member teams. These criteria must comply with the constitutional requirements of objectivity, predetermination, and impartiality (Art. 25 of the Italian Constitution). Courts implement this mandate in different ways. Some use explicitly random methods, such as lotteries or random number generators; others rely on fixed mechanical rules, like alphabetical or seniority-based rotations, which – once absences and workload balancing are factored in – function as effectively random systems. The organizational plan also determines how long each team remains together. In most courts, teams are fixed for the entire four-year term of the plan, unless a formal variation is required due to illness, transfer, or other justified reasons such as conflicts of interest.[25] Other courts adopt rotating team systems, in which team composition is reshuffled at regular intervals – such as monthly – according to pre-established schedules. Regardless of the model, any reshuffling applies only to new

[25]Substitution rules in these cases must also be predetermined, objective, and clearly specified in the organization plan.

cases: once a team begins hearing evidence in a trial, its composition is locked in and must remain unchanged until the final ruling is delivered (Art. 525 c.p.p.).[26]

Assignment of cases to teams. Case assignment follows a two-step process. First, territorial jurisdiction directs each new criminal filing to the court that covers the geographic area where the alleged offence occurred. Second – within that court – the case is assigned to one of its three-judge team through a procedure that must be strictly random, predetermined, and impartial (Art. 25 of the Italian Constitution). The goal is to eliminate any relationship between the identity of the judges and the characteristics of the case, such as its complexity, the parties involved, or the lawyers representing them. Each court is free to adopt its own randomisation method, as long as the criteria are publicly specified in its organisation plan and abide to the principles of impartiality and neutrality.[27] Judges cannot decline an assigned case except under limited statutory grounds for recusal, such as illness or extended leave (longer than one week), close personal ties to the parties, prior involvement in the same matter, or any other circumstance that might compromise impartiality. In the rare cases in which this occurs, cases are diverted in a random way to other judges, according to objective and pre-approved procedures set out in the court's organization plan.

3 Data

To answer our research questions, we need data on the gender composition of judicial teams, the crimes they rule on, the outcomes of their rulings, and a metric for sentence quality. To the best of our knowledge, no such dataset exists, as Italy lacks a digitized and accessible court-level database. We therefore constructed a novel dataset by manually collecting the universe of collegial ruling sentences from a sample of first-instance criminal courts and merging them 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).

[26] If a judge must be replaced mid-trial, the trial must start over and the evidentiary phase repeated.

[27] The primary tool used in most courts is GIADA (*Gestione Integrata delle Attività e dei Dati Amministrativi*), a centralised case-management system that distributes cases automatically according to predetermined rules. GIADA evaluates the complexity of each case and allocates them on a rotating basis across teams. Once a team receives a case of a given complexity level, it is temporarily excluded from receiving another case of the same type until all other teams have received one, ensuring a fair and balanced distribution of workload.

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 variation reassures us that our results are not specific to one type of court or geographical area.

The sentences were provided in paper format by the presidents of each court.[28] We digitized the sentences issued in collegial rulings and processed them using a combination of OCR and text-parsing algorithms to extract the following information: the name and location of the court, the start and end dates of the trial, the identities of the judicial team members, the crime under prosecution, and the final verdict. Since Italian first names unambiguously indicate an individual's gender, we infer the gender of judges based on their first names as collected during the scraping process.[29] By digitizing official rankings and scores from the national selection exams required for entry into the judiciary we obtain a measure of judges' experience and of their ability at entry.[30] Appendix Figure A2 presents an example sentence, with the relevant data fields highlighted in red. Finally, we retrieved additional personal information – such as judges' age, municipality, and region of birth – through online sources.[31]

3.1 Descriptive statistics

Table 1 provides descriptive statistics for our sample, which covers the years 2004 to 2021.[32] In total, we digitized more than 10,000 cases. Since a single case may involve multiple defendants and multiple charges, this corresponds to over 65,000 decisions. Each individual decision constitutes the unit of observation in our analysis.

[28] With the exception of the Court of Naples, which instead provided a digital extraction of the relevant variables.

[29] We further validate this using the title "Dr.", which precedes judges' first names and varies by gender: "*Dott.*" for men and "*Dott.ssa*" for women.

[30] We digitized all judicial selection exams from 1965 to 2007; exams from 2008 onward are publicly available on the Ministry of Justice website. We were unable to retrieve the rankings for three national exams (those announced with the following ministerial decrees: DM 3 agosto 1967; DM 3 dicembre 1988 e 19 gennaio 1989; DM 24 aprile 1989). Additionally, we were unable to match 84 out of 371 judges (22.6%), 64 of whom (76.2%) are GOTs and thus did not enter the judiciary through the public competitive exam.

[31] We were able to retrieve this information for 96.2% of our judges using information on official bulletins, articles, and bios.

[32] Data from the Court of Florence begin in 2004, while data from Naples Nord start in 2013, when the court was established. Data from Torre Annunziata and Naples are available from 2010.

Panel A of Table 1 provides an overview of the composition of the first-instance judiciary in our courts. There are 371 unique judges operating in our sample's courts in the years considered, of whom 56 percent are women.This gender distribution is fairly consistent across courts, with the share of women ranges from 50 to 59 percent (not reported in the Table). The average age of judges is 49 years, and nearly 20 percent were born in a region different from the one in which they work. As expected, given the dynamics of the judiciary's gender composition, female judges tend to be younger (47 vs 51 years, not reported in the Table), less experienced (19 vs 22 years of service, not reported in the Table) and less likely to have migrated from another region (16 percent vs 24 percent, not reported in the Table).

Among the 1,134 unique teams we cover in our dataset, the majority are mixed-gender (Panel B): in particular, 33.5 percent of the teams consist of two men and one woman, while 41.7 percent include one man and two women. All-female teams are more common than all-male teams (18.4 percent compared to 6.3 percent).[33] The relatively balanced gender composition of the Italian judiciary enables us to study the effects of both all-female and all-male judicial panels within the same profession, unlike much of the existing literature, which focuses on male-dominated fields where women are so few that all-female teams are virtually absent.

Panel C reports characteristics of the 10,676 cases included in our sample. On average, each case has slightly more than two defendants and five distinct charges. The vast majority of defendants are male (89 percent), and 18 percent are foreigners.

Finally, Panel D of Table 1 summarizes the 65,051 decisions issued by judicial teams in our dataset. A guilty verdict is issued in 43 percent of cases. Conditional on conviction, the average prison sentence is five years. Trials are lengthy, lasting on average more than two years. Just under 20 percent of rulings in our sample are appealed, nearly all of which are convictions (97 percent, not reported in the Table). This implies that 42 percent of convictions are appealed.

The types of crimes observed in our sample are summarized in Appendix Table A1. The four most frequent categories – accounting for nearly 80 percent of all offenses – are crimes against property, public order, public trust, and the person. The first category includes offenses against property involving violence or threats (such

[33] In Section 4.2, we show that the observed distribution of teams aligns with a simulated distribution in which teams are randomly formed from a population that has the same proportion of female judges as in our sample.

as theft, robbery, and extortion), and involving deception (such as fraud, fraudulent bankruptcy, usury, and money laundering). The second group includes crimes such as mafia-type criminal association, drug and arms trafficking, devastation, and looting. Crimes against public trust include counterfeiting money or documents, forging public or private acts, impersonation, and falsifying identity or official seals. The final category includes personal injuries, manslaughter, threats, and crimes targeting women and children.

4 Empirical strategy

The main challenge in identifying the causal effect of team gender composition on decision-making is twofold. First, team composition and task assignment are often endogenous to both organizational and individual characteristics. Second, defining and measuring the intrinsic quality of team decisions is challenging in most real-life settings. To overcome these challenges, the ideal experiment would randomize individuals into teams, assign tasks randomly, and use tasks with clear performance benchmarks - allowing a clean comparison of the nature and quality of decisions across groups with different gender mixes.

The Italian judiciary - characterized by a hierarchical structure, and by the random allocation of judges and cases to panels within courts - provides exactly such a setting. Intuitively, we assess whether the gender composition of judicial teams causally affects both the *type* of decisions, measured by conviction rates, and their *accuracy*, measured by "appeal-proof" rulings (i.e., decisions either unappealed or upheld on appeal) by comparing outcomes across teams operating in the same court and ruling on the same crime but differing in gender composition.

The validity of this strategy rests on two assumptions. First, although judges may choose which court to join based on its overall characteristics (including the gender composition of its workforce), they have no discretion over their specific assignments or colleagues. Thus, within a court, variation in both team composition and case assignment can be treated as random (see Section 2.3). Second, measuring decision accuracy through appeal outcomes relies on the assumption that appellate courts can only improve upon first-instance rulings by identifying and correcting judicial mistakes.[34] We

[34]This assumption has been assumed several times in the literature.For instance, Alesina and La Ferrara (2014) use reversals by higher courts as a proxy for the quality and impartiality of lower-

provide evidence in favor of the validity of our identifying assumptions in Section 4.2.

4.1 Regression model

We exploit the random allocation of judges to teams and of cases to teams within courts to test whether gender composition influences judicial outcomes. Our analysis focuses on two dimensions: (i) the likelihood of conviction and (ii) the resilience of rulings to appeal, which we treat as a proxy for the quality and accuracy of decisions. Specifically, we estimate the following specification:

$$Y_{isct} = \beta_0 + \beta_1 F0_{isct} + \beta_2 F1_{isct} + \beta_3 F2_{isct} + \delta X_{ict} + \eta Z_{isct} + \gamma_c + \gamma_{s,t} + \epsilon_{isct} \qquad (1)$$

where Y_{isct} is the outcome of interest for team i in court c ruling on crime s in year t: either (i) an indicator equal to 1 if the team delivers a guilty verdict (conviction rate), or (ii) an indicator equal to 1 if the verdict is not appealed or is upheld on appeal (appeal-proof decision rate). The main variables of interest are $F0_{isct}$, $F1_{isct}$, and $F2_{isct}$, where Fj_{isct} is an indicator equal to 1 if team i has j women and $(3-j)$ men.[35] Because of random assignment, β_1, β_2, and β_3 describe the causal effect of having a judicial team with no woman, one woman, and two women on our outcomes of interest compared to a team with three women.

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 team characteristics across courts as well as for courts' characteristics that are constant across teams within a court. We also include crime type-by-year fixed effects to ensure that we are comparing judicial teams 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 team characteristics, X_{ict}, and a vector of case characteristics, Z_{isct} to increase precision, and we cluster the standard errors at the team level to account for potential correlation in decision outcomes within teams.[36]

court decisions, while Coviello et al. (2019) to study the productivity effect of labor specialization.

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

[36] The vector X_{ict} includes each judge's age, experience, region of birth, and exam grade. 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 non-Italian.

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 teams 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 teams. After controlling for court and crime type-by-year fixed effects, 93.6 percent of the original variation remains for all-men teams, 95.8 percent for teams with two men and one woman, 95.5 percent for teams with one man and two women, and 96.4 percent for all-women teams. This reassures us that our empirical strategy leaves sufficient residual variation in our key variables. Moreover, the fact that nearly all variation survives the inclusion of fixed effects supports the credibility of our identification strategy, indicating that team gender composition is not mechanically determined by court assignment rules or crime types but remains largely idiosyncratic and plausibly random.

Random assignment of judges and cases to teams within court: 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 teams within courts results from random fluctuations rather than from systematic selection. We provide evidence in favor of this assumption in three ways.

First, we formally test for selection on observables by examining whether the variation in the gender composition of the teams is correlated with variation in case characteristics. Table 2 reports the results of separate regressions of trial characteristics on indicators for all-male teams, two-men–one-woman teams (2M1F), one-man–two-women teams (1M2F), and all-female teams, controlling for court and year fixed effects. All variables have been standardized to have mean zero and standard deviation one. Only one of the coefficients is statistically significant at the 5 percent significance level, which is what would be expected by chance. This suggests that cases are well-balanced across teams with different gender composition.

Second, we show that judges assigned to mixed-gender teams and same-gender teams do not differ in terms of observable characteristics. Table 3 displays the results of separate regressions of judges' observable characteristics on the mixed-gender team

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 and they are all small in magnitude.

Finally, we show that the observed distribution of judicial teams by gender composition is consistent with what we would obtain by randomly forming teams within courts. To test this, we repeatedly draw three judges without replacement from each court's judge pool – matching the actual number of observed teams – and record the number of women in each simulated team. Table 4 compares the observed and simulated distributions and reports the results of a chi-square test. Since we fail to reject the null hypothesis that the two distributions are equal, we conclude that the observed team composition does not differ significantly from what would arise under random formation.

Correlated Effects: Because gender composition varies randomly across teams within courts, the parameters of interest β_1, β_2, and β_3 in (1) can be given a causal interpretation. A remaining concern, however, is that these coefficients might reflect not only gender composition but also other observable characteristics correlated with it. For instance, if female and male judges systematically differ in age, then the age profile of a team will vary mechanically with its gender composition, and the estimated effects could partly capture the influence of age rather than gender alone (correlated effects as discussed in Manski, 1993).

We assess the role of correlated team characteristics in two ways. First, to isolate the effect of gender composition from other attributes, we augment the control vector X_{isc} in equation 1 with a comprehensive set of team-level characteristics: judges' age, experience, region of origin. This is important because, as discussed in Section 3.1, women judges tend to be younger, less experienced, and less geographically mobile (i.e., less likely to work in their region of origin) than men. Although men and women do not differ in performance on the national entry exam, we include exam scores to capture ability at entry.[37] In Section 5.4, we show that including these controls does not alter our results. Second, we re-estimate equation 1, replacing gender composition with alternative dimensions of diversity in which men and women differ—namely age, judicial experience, and region of origin. As we show in Section 5.4, none of these alternative forms of diversity explains our findings: they are not systematically asso-

[37] In our setting there is no need to control for educational background, as the path to becoming a judge is standardized and identical across genders.

ciated with differences in team outcomes and therefore cannot account for the effects we observe for gender composition.

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 team. Column (1) of Table 5 reports the results from Equation (1). All-women teams are more likely to issue a conviction decision than any other type of team: in particular, they are 7.1 percentage points (16.7 percent) more likely to convict the defendant than all-men teams, and 5.7-9.3 percentage points (13.3-21.8 percent) more likely to convict than mixed-gender teams. By contrast, conviction rates among all-men and mixed-gender teams do not differ significantly from one another, nor do they vary within mixed-gender teams depending on whether they include one or two women.[38]

As we are interested in examining not only differences in conviction rates but also the severity of sentencing, we next investigate whether teams make different sentencing decisions depending on their gender composition. We estimate the effect of team gender composition on the the length of prison sentences. Column (2) of Table 5 reports results from an unconditional regression of team gender composition on jail time, and shows that all-women teams sentence defendants to 3 months more, on average, than mixed-gender teams, with no difference relative to the all-men teams. Because this unconditional measure reflects both the extensive and the intensive margins, column (3) focuses on conviction decisions only. Conditional on conviction, all-women teams impose prison sentences that are, on average, 12 months shorter than those handed down by all-men teams and 6 months shorter than those from mixed-gender teams.[39] By contrast, and similarly to the previous set of results, we find no statistically significant differences in sentencing severity between all-men and mixed-gender teams, nor within mixed-gender teams depending on whether they include one or two women.

Overall, our findings show that all-women teams have higher conviction rates than

[38]Teams with two men and one woman and teams with one man and two women display only marginally statistically significant conviction rates, and the effect is small (8 percent).

[39]In particular, teams with one woman issue sentences about 8 months longer than all-women teams, while teams with two women issue sentences about 4 months longer (not significant).

both mixed-gender and all-men teams, but do not impose longer prison sentences once a conviction occurs. Put differently, women appear harsher at the extensive margin—whether to convict—but not at the intensive margin—the severity of the sentence.

5.2 Quality of decisions

To understand whether differences in the type of decisions are associated with differences in team performance, we examine the *accuracy* of decisions taken by different teams. We use the appeal process—where first-instance decisions are reviewed by appellate courts—as a proxy for accuracy and structure the analysis in two steps.

First, we study whether team gender composition affects the *perceived* accuracy of rulings, measured by the probability that a decision is appealed. The underlying assumption is that unappealed decisions are generally perceived as sound. As discussed in Section 2, while all types of decisions can, in principle, be appealed by either defendants or prosecutors, appeals are overwhelmingly concentrated among convictions.[40] Second, we investigate the effect of team gender composition on the *effective* accuracy of judicial decisions, by whether an appellate court confirms or overturns them.

Combining both steps, we define a ruling as *appeal-proof* if it satisfies one of the following conditions:

(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) if the defendant was convicted, the prison sentence is also confirmed in length.[41]

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

[40] In our data, 94% of appeals stem from convicted cases. Being convicted increases the probability of appeal by 38 percentage points ($\approx$ 200%), and each additional year of prison time increases the probability of appeal by 1.2 percentage points (Appendix Table A3).

[41] Condition (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, 95 percent of decisions (and 88 percent of convictions) would be classified as appeal proof decisions. By contrast, our more nuanced definition reduces this figure to 87 percent (71 percent of convictions), allowing for a more detailed evaluation of decision-making accuracy. Nevertheless, our results are robust to a) removing condition (b2); b) relaxing condition (b2) to allow for a tolerance in the difference between the jail time ruled by the first instance court and that of the appeal court.

innocence.[42]

The fact that appeals in our data almost exclusively follow from conviction decisions introduces a form of one-sided selection: we can observe whether a conviction was accurate or not (based on confirmation or overturning on appeal), but we cannot distinguish between accurate and inaccurate acquittals, since acquittals are rarely appealed. While this limits our ability to fully assess judicial accuracy, the limitation is substantially mitigated by the random allocation of cases to judicial teams. Since teams are, on average, exposed to similar distributions of guilty and innocent defendants, systematic differences in conviction rates and the accuracy of convictions across teams must reflect differences in judicial behavior, rather than differences in case composition. As a result, if one team convicts more often and its convictions are more likely to be upheld on appeal than another team's, we can infer not only that it makes fewer false convictions, but also that it likely makes fewer false acquittals. In other words, even though we cannot directly observe the accuracy of acquittals, we can still draw meaningful conclusions about overall team performance based on the observable component of accuracy of decisions.

Perceived accuracy. Table 6 examines how the gender composition of judicial teams affects the likelihood that their decisions are appealed. We find that decisions made by all-men teams are 8.2 percentage points (43.2%) more likely to be appealed than decisions taken by all-women teams (column (1)). The presence of at least one woman in the team significantly reduces the probability of appeal, with no meaningful difference between teams with one, two, or three women. Since appeal likelihood may vary due to differences in conviction behavior across teams, column (2) restricts the sample to conviction decisions only. The estimated effect is virtually unaffected: convictions by all-men teams are still 45.2% (or 19.1 percentage points) more likely to be appealed. Again, decisions by mixed-gender teams do not differ significantly from those of all-women teams. Finally, since all-men teams are more likely to impose longer prison sentences upon conviction (see Table 5), one might be concerned that their higher appeal rates are simply a response to harsher punishments. To address this, column (3) adds a control for the length of the prison sentence. The results remain once again virtually unchanged, suggesting that the higher likelihood of appeal for all-men teams is not explained by sentence severity, but instead reflects a lower perceived accuracy

[42]In Appendix Table A4, we show that, in line with this assumption, appeal judges are more experienced and have higher baseline ability.

or fairness of their decisions.

Effective accuracy. We now examine how the gender composition of judicial teams affects the likelihood that their decisions are appeal-proof. Table 7 shows that the effects of gender composition on effective accuracy mirror those on perceived accuracy. Rulings made by all-male teams are not only more likely to be perceived as inaccurate, but are also 8.7 percentage points (10%) more likely to be appealed and overturned compared to other teams (column (1)). In contrast, rulings by all-women and mixed-gender teams are equally likely to withstand appeal. The effects are even more pronounced when we focus on conviction decisions only. All-men teams are 17.7 percentage points (24.8%) less likely to issue convictions that hold up on appeal. The presence of at least one woman on the team significantly improves outcomes: teams with one woman have appeal-proof conviction rates that are 11 percentage points (15.4%) higher than those of all-male teams, while there is no statistically significant difference between teams with one versus two women. In contrast, all-women teams make conviction decisions that are the most likely to withstand appeal, indicating the highest level of accuracy among all team compositions.

5.3 Summary and implications of main findings

Taken together, our results show that all-women teams have both the highest conviction rates and the highest rates of convictions that withstand appeal, suggesting they are more effective at distinguishing between guilty and innocent defendants. In contrast, all-men teams not only convict less frequently but also have the lowest rate of appeal-proof convictions. Mixed-gender teams behave similarly to all-male teams in terms of conviction rates (thus have lower conviction rates than all-women teams), but their convictions are nearly as robust to appeal as those of all-women teams (thus are of higher quality than those of men).

These findings have important implications. The fact that teams with higher conviction rates also have fewer convictions overturned on appeal suggests that the results cannot be explained solely by differences in judicial leniency. If teams with different gender compositions were equally skilled at assessing guilt and differed only in their propensity to convict, harsher teams would convict more defendants regardless of guilt and, as a result, should see a larger share of their rulings overturned on appeal. Instead, we observe the opposite: all-women teams convict more often but also produce

more appeal-proof convictions. This pattern points to differences in teams' ability to better distinguish guilt from innocence rather than differences in leniency, as the most plausible explanation. We return to this interpretation in greater detail in Section 6, where we provide a theoretical framework to rationalize these findings.

5.4 Robustness checks and heterogeneity

In this section, we assess the robustness of our results and perform additional heterogeneity analysis.

Selection on unobservables: The balancing tests in Tables 2 and 3 suggest that the variation in the gender composition of the teams 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 teams 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 paragraph. Appendix Table A5 reports the results on conviction rates (Panel A) and on appeal-proof conviction rates (Panel B). 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: As discussed in Section 4.2, a key concern is that our estimates may reflect not only the effect of gender composition but also that of other characteristics correlated with gender. For example, if certain team compositions systematically differ in other observable traits linked to both gender and case outcomes, the coefficients may partially capture the influence of those confounding factors.

We rule out that our results are driven by these other dimensions of diversity in three ways. First, all our specifications include controls for judges' age, experience, region of origin, and ability, ensuring that the estimated effects of gender composition are not driven by these other characteristics.[43] Including these controls has little

[43]On average, women judges are younger, less experienced, and less geographically mobile than

impact on our estimates, suggesting that these variables do not substantially contribute to the observed effects (see Appendix Table A5).

Second, we examine whether other forms of team composition—specifically age, regional background, experience, and ability—affect conviction rates and appeal-proof conviction rates. Results are reported in Table A6. Panel A analyzes age diversity by classifying teams as all-young, all-old, or mixed-age, based on the share of judges above the court's median age. Panel B focuses on regional background, dividing teams into all-south, all-north, or mixed-region groups according to judges' region of birth. Panel C evaluates experience by grouping teams as all-inexperienced, all-experienced, or mixed, based on the proportion of judges above the median years of service. Finally, in Panel D we focus on ability composition, dividing teams in all-low-ability, all-high-ability, and mixed-ability teams based on the proportion of team members above the median entry exam score. If the effects observed in our main analysis were driven by these characteristics rather than gender composition, we would expect similar estimated coefficients to those reported in Tables 5 and 7: for instance, since women tend to be younger, we would expect the coefficient of all-young teams to be similar to that of all-women teams. However, we find no significant relationship between any of these alternative dimensions and either conviction rates or appeal-proof conviction rates. This supports the conclusion that our main results are indeed driven by gender composition, not by other judge characteristics correlated with gender. More broadly, it suggests that other forms of diversity—such as age, experience, region, or ability—do not play a significant role in this context.

Finally, to further reassure that our results are not driven by judges' unobserved characteristics, we estimate a more restrictive specification with judge fixed effects. That is, we estimate the following model separately for men and women:

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

where Y_{isct} is the outcome of interest for judge j in team 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 A7 reports the results, which are largely consistent with our main findings. Because this specification

their male counterparts. Educational background is not a relevant factor in our context, as the path to becoming a judge is highly standardized and identical for both genders.

includes judge fixed effects, it holds constant all time-invariant characteristics of individual judges and exploits movements of judges across teams with different gender compositions. The consistency of our results under this specification indicates that the same judge behaves differently depending on the team's gender composition. In other words, the effects cannot be explained solely by individual judge characteristics but instead reflect how judicial behavior is shaped by the gender composition of the team.

Different classifications of crimes: Our crime classification follows the broad categories defined by the Penal Code, which groups 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 A8 shows that our results remain robust under this alternative and more detailed classification.

Gender-specific attitudes toward different crimes: One potential explanation for the observed differences in team behavior is that men and women may hold different attitudes, beliefs, or sensitivities regarding certain types of crimes. To explore this possibility, we examine whether specific categories of offenses are driving the effects documented in the previous sections. If this were the case, we might expect the presence of women on judicial teams to have a stronger impact on gender-sensitive crimes – such as violence against women or children (classified under "crimes against the person") – compared to more gender-neutral offenses, such as theft (under "crimes against property"). We test this in Appendix Table A9, where we sequentially exclude one crime category at a time and re-estimate the effect of gender composition on conviction rates (Panel A) and appeal-proof conviction rates (Panel B). Across all specifications, the estimated coefficients remain stable in both sign and magnitude and are not statistically different from our baseline results. This indicates that no single category of crime is driving the effects, and that the impact of gender composition is robust and consistent across a wide range of case types.

Gender and nationality of the defendant: Another potential concern is that our results may be partly driven by gender-based or nationality-based dynamics between judges and defendants. For instance, men and women judges may respond

differently when interacting with defendants of the same or opposite gender, or with those from different national backgrounds. To investigate this possibility, we estimate the effect of judicial gender composition on conviction rates (Panel A of Appendix Table A10) and appeal-proof conviction rates (Panel B of Appendix Table A10), separately for male and female defendants (columns (1) and (2)) and for Italian and foreign defendants (columns (3) and (4)). Across all subsamples, the patterns observed in previous sections remain largely consistent, suggesting that the effects of team gender composition are not driven by specific interactions with defendants of a particular gender or nationality.

6 Interpretation of results

In this section, we shed light on plausible mechanisms by providing and estimating structural models of team behavior that could rationalize our empirical findings. We show that our results are at odds with the predictions of standard decision-making models with no differences in skills or effort (see Mungan, 2024a,b; Chan et al., 2022, for a review of the issue). Such models would *not* predict that teams that convict more often are also teams that make fewer mistakes upon conviction. By contrast, we propose that a framework incorporating differences in skills or effort, alongside frictions arising from gender diversity within a team, can generate the patterns we observe. We then conclude by providing empirical evidence pointing to an effort-based channel as a key driver of our results.

6.1 Theoretical model

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 evaluation phase is followed by a decision phase, 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*, which reflect their taste for conviction, or their leniency. 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.

States and prior – Consider a team of two judges $i \in \{1, 2\}$ who make a binary decision under uncertainty.[44] Their task is to decide whether to convict a defendant, who is either guilty ($S = G$, with prior probability p) or innocent ($S = I$, with probability $(1-p)$). The judges do not observe S but receive a joint, noisy signal $X \in \{G, I\}$, whose precision depends on the effort exerted by the judges into analyzing and interpreting the information of the case.

Signal technology and cost of effort – Judges $i \in \{1, 2\}$ simultaneously exert effort $e_i > 0$ at a convex cost $c(e_i)$, with $c'(e) > 0$ and $c''(e) > 0$. Together, they observe a *joint* binary signal $X \in \{G, I\}$ that is correct with probability $\theta(e_i, e_j) \in \left(\frac{1}{2}, 1\right)$, where $\theta_{e_i} > 0$, $\theta_{e_i e_i} < 0$, and $\theta_{e_i e_j} < 0$.[45] That is, given the state the signal satisfies:

$$\Pr(X = G \mid S = G) = \theta, \quad \Pr(X = G \mid S = I) = 1 - \theta$$

Disutility from mistakes. When taking a conviction decisions, judicial teams can make mistakes. Let $d \in \{g, i\}$ denote the team's decision (g = convict, i = acquit). Despite the decision being a team one, each judge can differ in the disutility they experience from mistakes. A wrong conviction generates a disutility for judge i of $\phi_i \beta$ with $\beta > 1$; a wrong acquittal generates a disutility of ϕ_i; correct decisions yield no disutility:

$$u_i = \begin{cases} -\phi_i \beta - c(e_i), & \text{if } S = I,\ d = g \\ -\phi_i - c(e_i), & \text{if } S = G,\ d = i \\ -c(e_i), & \text{otherwise} \end{cases}$$

The key preference parameter β captures the disutility of a wrong conviction relative to a wrong acquittal. It can be interpreted as a judge's leniency. Given that convicting an innocent person is considered to be more costly than acquitting a guilty

[44] Note that, for simplicity, we model teams as consisting of two judges rather than three. This simplification is justified because (i) the decision rule is unanimity, and (ii) teams with two men and teams with two women are never statistically different from each other.

[45] Subscripts denote partial derivatives (e.g. $\theta_{e_i} = \partial\theta/\partial e_i$).

person (*Blackstone Principle*), we assume $\beta > 1$.[46] The parameter ϕ_i instead captures judge i's aversion to making *any* errors. It can be interpreted as the judge's diligence parameter, capturing anything that makes judicial mistakes feel more costly (such as personal preferences, professional norms, external scrutiny/sanctions). Following Chan et al. (2022), we normalize the utility of correctly classifying defendants to zero.[47]

We keep the parameter β common across judges due to both identification and interpretation reasons. With our moments and a single prior p, letting both β and ϕ vary across judges is weakly identified: optimal effort (and thus precision) depends on the product $\phi_i[p + \beta(1-p)]$, while the convict–acquit threshold depends only on β; heterogeneity in β is therefore largely observationally equivalent to heterogeneity in ϕ, so we let ϕ_i absorb individual differences and interpret β as the common normative asymmetry. In the following section, we will show how the model estimation changes when allowing β to vary across judges while keeping ϕ constant, as well as show you how the model with both variation in β_i and ϕ_i is poorly identified.

Stage 3: Ex-post optimal voting rule – The team convicts if the utility from convicting is larger than the utility from acquitting; that is, if $P(S = G \mid X) \geq \frac{\beta}{1+\beta}$. Note that since $\beta > 1$, the right-hand side is larger than $1/2$: the team requires more evidence than a random coin toss to convict.

Because the signal is binary, the team's decision is driven by which signal realization it obtains. If $X = G$,

$$P(S = G \mid X = G) = \frac{p\theta(e_i, e_j)}{p\theta(e_i, e_j) + (1-p)(1-\theta(e_i, e_j))}$$

Thus, the team will convict after a G signal if

$$\theta(e_i, e_j) \geq \frac{\beta(1-p)}{p + \beta(1-p)} \equiv \underline{\theta} \tag{2}$$

Stage 2: Disagreement about signal accuracy – Working in teams can introduce fric-

[46] 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.

[47] This is without loss of generality, as β is sufficient to determine the agent's optimal decision given the posterior $P(S = G|X)$.

tions, such as imperfect communication between judges or disagreement. We model these frictions as follows. Each judge believes the signal precision is $\theta_i = \theta(e_i, e_j) + \epsilon_i$, with $E(\epsilon_i) = 0$, ϵ_i small so that $\theta_i \in (1/2, 1)$. In other words, $\theta(e_i, e_j)$ represents how reliable the evidence *actually* is, given how hard the two judges have worked to understand it; ϵ_i represents judge i's *noise* in how the evidence is interpreted; and therefore θ_i is judge i's *subjective* estimate of that precision.

Judges bargain over a common belief θ^*. With probability q the judges *agree* and select one of their beliefs at random. Thus $\theta^* = \theta_1$ with probability $1/2$, and $\theta^* = \theta_2$ with probability $1/2$. With probability $1 - q$, instead, they *fail to agree* and keep their own θ_i. Then, the team convicts only when both individual thresholds are met: $d = g \iff \theta_1 \geq \underline{\theta} \ \wedge \ \theta_2 \geq \underline{\theta}$.

Stage 1: Effort choice – Judge i chooses effort to maximize his expected utility taking his teammate's effort as given. We assume each team anticipates frictions imperfectly; that is, they choose effort ignoring q:

$$\max_{e_i} \ \mathbb{E}[u_i] = -\phi_i \big[\beta(1-p) + p\big]\big(1 - \theta(e_i, e_j)\big) - c(e_i)$$

The first-order condition is

$$\theta_{e_i}(e_i, e_j) = \frac{c'(e_i)}{\phi_i \left[p + \beta(1-p)\right]} \tag{3}$$

with $\theta_{e_i e_i} < 0$ and $c'' > 0$ ensuring a maximum.

Conviction rates (C) – Let $P(\theta; \beta, p) \equiv P(\theta \geq \underline{\theta})$. With probability q, the judges bargain successfully and randomly adopt θ_1 or θ_2. Conviction then occurs with probability $qP(\theta; \beta, p)$. With probability $1 - q$ no agreement is reached; conviction requires $\theta_1, \theta_2 \geq \underline{\theta}$, giving probability $(1-q)P(\theta; \beta, p)^2$. Thus, $\Pi(\theta, q; \beta, p) = qP(\theta; \beta, p) + (1-q)P(\theta; \beta, p)^2$. Multiplying by the chance that the signal takes the value $X = G$ yields the unconditional conviction rate, namely:

$$C = \left[p\theta + (1-p)(1-\theta)\right] \Pi\big(\theta, q; \beta, p\big) \tag{4}$$

Appeal-proof conviction rates (APC) – Appeal-proof conviction rate is given by the

share of convictions that are truly guilty. That is,

$$\text{APC} = \Pr(S = G \mid d = g) = \frac{\Pr[S = G,\, d = g]}{\Pr[d = g]} = \frac{p\,\theta}{p\theta + (1-p)(1-\theta)} \tag{5}$$

APC depends only on the underlying precision θ and the prior p, and it is independent of both q and ϵ_i.

Comparative statics – Optimal effort is increasing in own aversion to mistakes: $\partial e_i^*/\partial \phi_i > 0$; hence higher ϕ_i increases effort, precision, and (given $\partial C/\partial \theta > 0$) conviction and accuracy. In constrast, optimal effort is decreasing in the other team member's aversion to mistakes: $\partial e_i^*/\partial \phi_j < 0$: working alongside a more diligent colleague reduces the effort optimally exerted by agent i, and viceversa. Finally, optimal effort is increasing in the prior p. If $\beta > 1$, raising p increases the combined error weight $p + \beta(1-p)$ and therefore effort.

Qualitative predictions – The model generates the following predictions:

(i) Teams whose judges have higher ϕ_i have higher C and APC. From equation (3), a rise in ϕ_i raises e_i, hence θ. From equation (4), $\frac{\partial C}{\partial \theta} = (2p-1)\Pi(\theta, q) + [p\theta + (1-p)(1-\theta)]\Pi_\theta, \qquad \Pi_\theta = q + 2(1-q)P(\theta) > 0$. The second term is always positive; the first term is positive whenever $p > 1/2$.[48] Hence C rises in θ under mild conditions. From equation (5), $\frac{\partial APC}{\partial \theta} = \frac{p(1-p)}{[p\theta+(1-p)(1-\theta)]^2} > 0$. Therefore APC unambiguously rises with θ.

(ii) Teams with lower q have lower C. From equation (4), $\frac{\partial C}{\partial q} = \alpha(\theta)\big[P(\theta) - P(\theta)^2\big] > 0 \quad (0 < P(\theta) < 1)$. Thus, teams experiencing *larger* frictions will have a lower conviction rate, with no effect on the accuracy of their convictions.[49]

6.2 Estimation

For a given vector of parameters, the model is solved by backwards induction. Some of these parameters are fixed in our estimations. We use $\theta(e_i, e_j) = \frac{\exp(\kappa\,(e_i+e_j))}{1+\exp(\kappa\,(e_i+e_j))}$, where we set $\kappa = 0.5$. As cost function, we use $c(e_i) = e_i^2/2$. We assume that agents can be of two types – men and women – and we allow ϕ to differ between types by defining ϕ_m for men and ϕ_f for women. We also allow q to differ between homogeneous

[48] Note that assuming $p > 1/2$ is empirically plausible, once defendants have already been indicted.

[49] Note that if teams with higher q also share a lower ϕ, the resulting decrease in effort pushes θ down, working against the q-channel. Hence the net effect on C is ambiguous. APC is unaffected because its expression does not contain q.

teams (q_0) and mixed-gender teams (q_{mix}). The resulting vector of parameters $\Theta = (\beta, \phi_m, \phi_f, p, q_0, q_{mix})$ is estimated using indirect inference as the vector that minimizes the distance: $Q = (m(\Theta) - \hat{m})'W(m(\Theta) - \hat{m})$. The vector $\hat{m}$ includes the empirical moments used as "targets" to be matched by the structural model:

$$\hat{m} = \begin{bmatrix} C_{AW} \\ C_{AW} - C_{AM} \\ C_{AW} - C_{MG} \\ APC_{AW} \\ APC_{AW} - APC_{AM} \\ APC_{AW} - APC_{MG} \end{bmatrix}$$

Where C_t and APC_t denote the conviction rates and the appeal proof conviction rates of team t (all-women (AW), mixed-gender (MG), all-men (AM)).[50] The vector $m(\Theta)$ includes these same moments as predicted by the model for a given vector of parameters Θ. The weighing matrix W is a diagonal matrix containing the inverse of the variance of the empirical moments. The estimation algorithm selects $\hat{\Theta}$ by randomly drawing a set of N=100 vectors of initial values for the free parameters, by finding the best estimator $\hat{\Theta}_{min}$ for each vector (i.e., a local minimum), and by selecting the best estimator among those 100 local minima.

Table 8 reports the estimated parameters for three specifications. Column (1) is our baseline model with a common preference parameter β and group-specific diligence parameter ϕ. Column (2) extends the model to allow both β and ϕ to differ by gender. Column (3) instead holds diligence fixed across genders (same ϕ) and lets β vary. Table 9 shows the corresponding simulated moments.[51]

Same β, different ϕ – The estimates imply $\phi_f > \phi_m$ (with ϕ_f about 7.1 times ϕ_m) and larger frictions in mixed-gender teams ($q_{mix} < q_0$). Since APC depends only on precision θ (and prior p), not on q, higher ϕ raises effort increasing both C and APC, while lower q depresses C only. This specification closely matches the cross-team

[50] We group mixed-gender teams with one woman and two women together, since the estimated coefficients in our main results are similar in magnitude and not statistically different from each other.

[51] The targeted moments are obtained from a regression that pools the two mixed-gender team types:

$$Y_{isct} = \tilde{\beta}_0 + \tilde{\beta}_1 F0_{isct} + \tilde{\beta}_2 Mixed_{isct} + \delta X_{ict} + \eta Z_{isct} + \gamma_c + \gamma_{s,t} + \epsilon_{isct}$$

where Y_{isct} is either an indicator for conviction or for appeal-proof conviction. We pool mixed teams to mirror the model structure and because teams with one or two women are never statistically different. Results are in Appendix Table A11. Although point estimates differ slightly from Tables 5 and 7, the results and their interpretation are qualitatively unchanged.

differences, while slightly overshooting the *levels*.

Different β, different ϕ – In this specification, women both (a) place greater weight on avoiding wrongful convictions ($\beta_f > \beta_m$) and (b) are more diligent ($\phi_f > \phi_m$). To ensure the model is not underidentified, we set $\phi_m = 1$, so that ϕ_f is estimated to be 7.5 times ϕ_m. This model delivers a better fit of the underlying data as measured by the goodness of fit reported at the bottom of Table 8, but both the estimated moments and the parameters are quite similar to the baseline version of the model. As anticipated in Section 6.1, allowing variation in both β and ϕ delivers a very similar fit to varying ϕ alone, reflecting weak separate identification of the two parameters.

Different β, same ϕ – When the model tries to explain the data using preference differences alone (holding ϕ fixed across genders), it fails to reproduce the key cross-team gaps and over-predicts the levels, especially APC. Constraining ϕ forces the model to inflate β_f to raise effort which pushes θ too high. This indicates that preference heterogeneity by itself cannot generate the effort levels to closely match the observed patterns in the data.

The model helps us disentangle the role of preferences versus diligence in driving our main results, and the model estimation points to an important gender difference in diligence. In the next section, we present a series of empirical tests designed to further support this conclusion and to shed light on whether the observed difference in diligence reflects disparities in raw talent or in effort.

6.3 Preferences, skills and effort

Two important considerations emerge within our theoretical framework. First, the gender composition of a team can induce systematic variation in decision-making either through a *diagnostic skill channel*, affecting a team's assessment of a case, or through a *preference channel*, affecting a team's propensity to convict given the assessment of a case. Second, the model implies that under quasi-random assignment of cases to teams, the joint distribution of conviction rates and appeal-proof conviction rates across teams with different gender compositions 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 all-women teams exhibit both higher conviction rates *and* higher appeal-proof conviction rates (see Tables 5 and 7) is inconsis-

tent with the hypothesis that all teams operate along the same production possibilities frontier. If higher conviction rates simply reflected harsher preferences, we would expect more mistakes and therefore more appeals and reversals. Instead, the evidence that teams convicting more also make fewer wrong convictions points to differences in their ability to distinguish between guilty and innocent defendants – in other words, to differences in skill and/or effort across teams characterized by a different gender composition.

While the evidence strongly suggests that preferences are not the main driver of our results, the findings remain consistent with variation in either *innate ability* or *effort*. Disentangling the two is inherently difficult, but in what follows we present additional evidence indicating that differences in effort, rather than raw talent alone, are most likely to explain the patterns we document.

Entry examination results. First, we examine whether our findings could be driven by men and women being differently selected into the judiciary. Given that the profession was historically male-dominated, one might expect women to be more positively selected on average, raising the possibility that pre-existing differences in baseline ability explain our results. To investigate this, we collected and digitized the official rankings from all national entry examinations for the judiciary held between 1965 and 2019.[52]

If women were more positively selected than men, we would expect them to perform better on the entry exam. Figure 1, however, shows no such pattern. Panel (a) compares the distribution of grades for men and women, which are nearly identical. Panels (b) and (c) plot average grades by exam year and by birth year, respectively, and reveal a general upward trend over time. Crucially, this increase is common to both men and women, suggesting no meaningful gender gap in exam performance.

The absence of performance differences indicates that our results are unlikely to be explained by selection patterns at entry. Appendix Table A12 further confirms this: adding exam performance as a control in our main regressions leaves the estimates virtually unchanged. Taken together, this evidence suggests that ability at entry is not a key driver of the observed effects.[53]

[52]We successfully digitized 68 examinations rounds and were able to match all judges in our dataset, except for 69 *Giudici Onorari di Tribunale* (G.O.T.), who serve as auxiliary judges and are appointed without taking the national entry exam. For more details, see Section 2.

[53]This evidence does not rule out the possibility that women are more positively selected along other traits not captured by the exam. It does, however, show that they are not differentially selected

Duration of the decision-making process. We next investigate whether the gender composition of judicial teams affects the time required to reach a decision. This is an important margin to consider, as faster decision-making is typically desirable in the judiciary, where case backlogs are substantial and costly (Coviello et al., 2019). At the same time, longer decision times may reflect greater deliberation and, potentially, higher accuracy. However, they can also arise from frictions within the team, such as prolonged discussions or difficulties in reaching consensus.

We use trial duration as a proxy for the length of the decision-making process, defined as the number of days between the assignment of a case to a team and the final ruling.[54] Because cases are randomly assigned to teams, any systematic differences in trial duration – once court fixed effects are included – can be attributed to variation in team composition. Table 10 reports the results. All-male teams complete cases substantially faster than any other composition: their trials last, on average, eight months (30 percent) less than those of all-women teams and nearly six months (22 percent) less than those of mixed-gender teams. The difference between all-women and mixed-gender teams, by contrast, is not statistically significant.

Taken together, these results suggest two mechanisms. First, the fact that all-women teams spend more time on cases than all-men teams points to greater effort invested in deliberation. If the higher quality of women's decisions were driven primarily by innate talent, they should be able to achieve better outcomes within the same deliberation time as men. Instead, the evidence shows that women both achieve better outcomes and devote more time to each case – a pattern more consistent with higher effort than with higher raw talent. Second, the finding that all-women and mixed-gender teams take equally long to reach decisions – despite differing conviction and appeal outcomes – points to the presence of frictions in mixed teams: they devote as much time as all-women teams but reach fewer convictions, suggesting that part of their deliberation is absorbed by difficulties in reaching consensus. At the same time, the fact that they have higher appeal-proof rates than all-men teams indicate that when consensus is reached, the additional scrutiny improves the quality of their rulings. In other words, the slower process and lower conviction rate with respect to all-women teams appear to reflect the cost of frictions in deliberation, while the higher

on the dimensions the judiciary itself considers most important at entry.

[54]The analysis excludes the Court of Naples, for which assignment dates were not available. Results are robust to using a coarser measure available for all courts, defined as the difference between the year of assignment and the year of ruling.

appeal-proof rate with respect to all-men teams reflects the benefit of more thorough scrutiny before agreement is reached.

Case complexity: To further probe the role of effort, we examine whether the effects of gender composition vary with the complexity of the cases adjudicated. The intuition is straightforward. Simpler cases – where the information and evidence is arguably clearer and the decision problem more straightforward – should require less effort to extract a precise signal, and hence differences across teams should be minimal. In complex cases – where evidence is noisier and responsibilities harder to disentangle – differences in effort are more likely to matter.

Table 11 presents estimates separately for *more complex cases*, defined as those involving multiple crimes and defendants (Panel A), and *less complex cases*, involving a single crime or defendant (Panel B).[55] The results support the effort hypothesis. In less complex cases, all teams behave similarly regardless of their gender composition. By contrast, in more complex cases, the patterns documented in Section 5 reemerge: all-women teams display both higher conviction rates and higher appeal-proof conviction rates than any other team type, indicating greater accuracy in distinguishing between guilty and innocent defendants. All-male teams, by comparison, convict less frequently and have the lowest appeal-proof rates. Mixed-gender teams resemble all-male teams in terms of conviction rates, but their convictions are significantly more robust to appeal, and are closer to the higher level of quality of all-women teams. Consistent with the role of effort, column (3) of Table 11 shows that deliberation times differ by team composition only in complex cases.

These patterns highlight that gender composition matters precisely when adjudication requires greater effort. In straightforward cases, where less deliberation is needed, differences across teams largely vanish. But in complex cases, where additional effort in processing evidence and weighing competing narratives becomes critical, all-women teams appear to invest more effort, leading to more accurate and defensible decisions, with the opposite happening for all-men teams. Mixed teams – although affected by frictions – still achieve higher-quality convictions than all-male teams.

Written opinions: To further support the idea that differences in exerted effort drive our findings, we examine the written reasoning that judicial teams are required

[55]We adopt this definition of case complexity because it is one of the criteria used by courts algorithms to allocate cases across teams and judges to balance their workload.

to provide for every verdict.[56]. While they do not create binding precedent, these texts are subject to appellate scrutiny and must clearly set out the logic and evidentiary evaluation underlying the decision.

We focus on three features of these documents. First, we focus on their *length* (in words). Long opinions often lack clarity and conciseness, making it difficult for parties, lawyers, and appellate courts to follow legal reasoning (Butt, 2013).[57] Second, we build an index of drafting quality drawing on several linguistic measures of judicial texts described in Giommoni et al. (2024) and combine them using principal component analysis.[58] Finally, our third measure of drafting quality refers the frequency of citation of Supreme Court maxims. These are short summaries of a legal principle distilled from Cassation judgments meant to be concise reference points rather than binding precedents. While citation of maxims is not inherently discouraged, excessive reliance on them can weaken judicial reasoning and lead to the arguments appearing mechanically constructed, rather than demonstrating a thoughtful and case-specific analysis (Ministero della Giustizia, 2016).[59]

Results are reported in Table 12.[60] All-male teams produce judgments that are more than twice as long as those of other teams and cite Court of Cassation maxims nearly three times as often as mixed-gender or all-women teams. Their written reasons are also over one standard deviation less clear than those of the other team types. By contrast, mixed-gender and all-women teams are statistically indistinguishable on these measures, suggesting similar levels of effort devoted to produce clear and concise judicial reasoning.

[56]These "Reasons for Judgment" resemble U.S. judicial opinions but differ in key ways. See Appendix B for a detailed comparison.

[57]Shorter reports have also been used as a proxy for greater effort in clarity and communication by radiologists in Chan et al. (2022).

[58]The clarity index is standardized to have mean 0 and standard deviation 1. The measures used to build the index include: average sentence length; average word length; average length of root words; frequency of gerunds, demonstrative adjectives, modal verbs, and contingency clauses. Higher values of these features indicate less clear, and lower-quality drafting. As argued in Giommoni et al. (2024), the reasons are intuitive: long sentences are harder to understand and more prone to ambiguous subjective interpretations; gerunds and contingency clauses make sentences pompous or convoluted while plainer, more concise wording improves comprehension.

[59]As detailed in Appendix B, the Italian system is not based on common law. Therefore, judges do not heavily rely on court precedents in formal adjudications. Instead, their reasoning is primarily grounded in legal codes, with occasional reference to Supreme Court maxims.

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

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 teams, in which ruling takes place collegially; ii) empirically assess the accuracy 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, men-only, and mixed-gender teams. Specifically, all-female teams both convict more often and issue the most appeal-proof convictions, suggesting they are more effective at distinguishing guilt from innocence. By contrast, all-male teams convict less frequently and have the lowest rate of appeal-proof convictions. Mixed-gender teams resemble all-male teams in conviction rates but achieve conviction quality nearly as high as all-female teams.

Guided by a stylized model of diagnostic skill (Chan et al., 2022), we interpret the finding that higher conviction rates coincide with fewer reversals in all-female teams as ruling out that judicial leniency is the only driver of our results. Instead, this pattern points to the presence of differences in teams' ability to distinguish guilt from innocence. Such ability could, in principle, reflect either innate talent or effort. Although disentangling the two is inherently difficult, we provide a series of tests consistent with an effort-based channel. Our model estimates, together with additional reduced-form tests, also rationalize why mixed-gender teams deliver convictions of similar quality to all-female teams but at lower rates: in gender-diverse teams, it is harder to translate a signal of guilt into a unanimous conviction than in homogeneous female teams.

References

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

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.

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.

Bernstein, S., Colonnelli, E., and Iverson, B. (2019). Asset allocation in bankruptcy. The Journal of Finance, 74(1):5–53.

Bertrand, M., Black, S. E., Jensen, S., and Lleras-Muney, A. (2019). Breaking the glass ceiling? the effect of board quotas on female labour market outcomes in norway. The Review of Economic Studies, 86(1):191–239.

Born, A., Ranehill, E., and Sandberg, A. (2022). Gender and willingness to lead: Does the gender composition of teams matter? Review of Economics and Statistics, pages 1–46.

Butt, P. (2013). Modern legal drafting: A guide to using clearer language. Cambridge University Press.

Chan, D. C. (2021). Influence and information in team decisions: evidence from medical residency. American Economic Journal: Economic Policy, 13(1):106–137.

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.

Che, Y.-K. and Kartik, N. (2009). Opinions as incentives. Journal of Political Economy, 117(5):815–860.

Chen, Y. (2021). Team-specific human capital and team performance: evidence from doctors. American economic review, 111(12):3923–3962.

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.

De Paola, M. and Scoppa, V. (2015). Gender discrimination and evaluators' gender: evidence from italian academia. Economica, 82(325):162–188.

Della Torre, J. (2022). La crisi dell'appello penale nel prisma della statistica giudiziaria. Archivio penale: rivista quadrimestrale di diritto, procedura e legislazione penale speciale, europea e comparata: LXXIV, 2, 2022, pages 471–530.

Deloitte, B. (2016). Global human capital trends 2016, the new organization: different by design.

Deschamps, P. (2024). Gender quotas in hiring committees: A boon or a bane for women? Management Science, 70(11):7486–7505.

Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. American Economic Review, 108(2):201–240.

Egan, M., Matvos, G., and Seru, A. (2022). When harry fired sally: The double standard in punishing misconduct. Journal of Political Economy, 130(5):1184–1248.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. The quarterly journal of economics, 133(4):1645–1692.

Gagliarducci, S. and Paserman, M. D. (2012). Gender interactions within hierarchies: evidence from the political arena. The Review of Economic Studies, 79(3):1021–1052.

Galasso, A. and Schankerman, M. (2015). Patents and cumulative innovation: Causal evidence from the courts. The Quarterly Journal of Economics, 130(1):317–369.

Giommoni, T., Guiso, L., Michelacci, C., and Morelli, M. (2024). The economic cost of ambiguous laws. Technical report.

Hardt, D., Mayer, L., and Rincke, J. (2024). Who does the talking here? the impact of gender composition on team interactions. Management Science. Forthcoming.

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

Hong, L. and Page, S. E. (2001). Problem solving by heterogeneous agents. Journal of economic theory, 97(1):123–163.

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.

Hughes, N., Khan, Z. U., and Mengel, F. (2023). Diversity in committees. Available at SSRN 4495035.

Karpowitz, C. F., O'Connell, S. D., Preece, J., and Stoddard, O. (2024). Strength in numbers? gender composition, leadership, and women's influence in teams. Journal of Political Economy, 132(9):3077–3114.

Kets, W. and Sandroni, A. (2021). A theory of strategic uncertainty and cultural diversity. The Review of Economic Studies, 88(1):287–333.

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.

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.

Lyons, E. (2017). Team production in international labor markets: Experimental evidence from the field. American Economic Journal: Applied Economics, 9(3):70–104.

Maestas, N., Mullen, K. J., and Strand, A. (2013). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. American economic review, 103(5):1797–1829.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. The review of economic studies, 60(3):531–542.

Marx, B., Pons, V., and Suri, T. (2021). Diversity and team performance in a kenyan organization. Journal of Public Economics, 197:104332.

Ministero della Giustizia (2016). Relazione: Gruppo di lavoro sulla sinteticità degli atti processuali. Technical report, Ministero della Giustizia, Roma. Accesso: 22 agosto 2025.

Mungan, M. C. (2024a). Skill, preferences, and monotonicity violations with endogenous judgement quality. Texas A&M University School of Law Legal Studies Research Paper.

Mungan, M. C. (2024b). Symmetry, presumptions, and judges design. Working Paper. Available at SSRN: https://ssrn.com/abstract=4322684.

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.

Radbruch, J. and Schiprowski, A. (2024). Committee deliberation and gender differences in influences. Working paper.

Rebolledo, E. B., Bruhn, J., Choon, T. H., and Weber, E. A. (2024). Gender composition and group behavior: Evidence from us city councils. Technical report, National Bureau of Economic Research.

Sarsons, H. (2024). Interpreting signals in the labor market: evidence from medical referrals. R&R Review of Economic Studies.

Tables and Figures

Table 1: Descriptive statistics

	Mean	St.dev	Min	Max
A. Judges				
Female	0.561	0.497	0	1
Age	48.588	9.200	28	74
Experience	20.429	10.091	0	48
Entry exam grade	106.376	10.079	86	148
From central Italy	0.148	0.356	0	1
From northern Italy	0.040	0.196	0	1
From southern Italy	0.812	0.391	0	1
From a different region	0.196	0.397	0	1
N. judges	371			
B. Panels				
3M team	0.063	0.244	0	1
2M1F team	0.335	0.472	0	1
1M2F team	0.417	0.493	0	1
3F team	0.184	0.388	0	1
N. panels	1134			
C. Case characteristics				
Number of defendants	2.257	4.657	1	220
Number of crimes	5.515	19.402	1	609
Defendant is foreign	0.184	0.380	0	1
Defendant is female	0.108	0.270	0	1
N. Cases	10676			
D. Decisions characteristics				
Guilty	0.426	0.494	0	1
Jail time (cond.)	5.215	4.797	0	30
Length of trial (years)	2.316	1.704	0	22
Appeal	0.190	0.392	0	1
Appeal (cond.)	0.422	0.494	0	1
N. Decisions	65051			

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 teams' characteristics. *All-men team* denotes a team consisting of all men; *2M1F team* denotes a team made by two men and one woman; *1M2F team* denotes a team made by one man and two women; *All-women team* denotes a team consisting of all women. Panel C describes case characteristics. Panel D describes decisions. *Guilty* is an indicator for whether the team 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		2M1F		1M2F		All-women	
	Estimate	Std.Err.	Estimate	Std.Err.	Estimate	Std.Err.	Estimate	Std.Err.
Dependent variables:								
N. of crimes	-0.021	(0.032)	0.049	(0.062)	-0.045	(0.076)	0.008	(0.058)
N. of defendants	0.008	(0.041)	0.108	(0.075)	-0.091	(0.059)	-0.023	(0.033)
Defendant is female	0.012	(0.011)	-0.006	(0.015)	0.012	(0.015)	-0.015	(0.015)
Defendant is foreign	-0.004	(0.023)	-0.047**	(0.019)	0.027	(0.020)	0.026	(0.018)

Notes. This table reports the coefficients of separate regressions of each characteristic on the all-men, 2M1F, 1M2F, all-women team indicators. All regressions include court and year fixed effects. All variables have been standardized. Standard errors clustered at the team 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	-0.025*	(0.010)	-0.058	(0.056)
Experience	-0.021	(0.012)	-0.084	(0.103)
Experience	0.027	(0.031)	0.027	(0.032)
From different region	-0.007	(0.013)	0.044	(0.030)
From northern Italy	-0.001	(0.005)	0.027	(0.016)
From central Italy	-0.024	(0.018)	-0.054	(0.046)
From southern Italy	0.023	(0.015)	0.040	(0.038)

Notes. This table reports the coefficients of separate regressions of each characteristic on the mixed-gender team 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: Observed and simulated distribution of teams

	All-men	2M1F	1M2F	All-women	Total
Observed distribution	72	380	473	209	1,134
Simulated distribution	89	348	488	209	1,134
Total	161	728	961	418	2,268
Pearson chi2(3) = 3.4358, P-val = 0.329					

Notes. This table reports the observed and the simulated distributions of teams, as well as the results of a Chi-square test on the equality of the two distributions.

Table 5: Effect of gender composition of the team on conviction rates and jailtime (years)

	(1) Guilty	(2) Jail time	(3) Jail time (cond.)
All-men team	-0.071**	-0.089	0.992**
	(0.031)	(0.290)	(0.483)
2M1F team	-0.093***	-0.368*	0.734**
	(0.025)	(0.207)	(0.349)
1M2F team	-0.057**	-0.236	0.345
	(0.024)	(0.190)	(0.312)
Observations	65051	65051	27686
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.426	2.220	5.215

Notes. In column (1), the dependent variable is an indicator for whether the team convicts the defendant. In columns (2) and (3), the dependent variable is the duration of imprisonment (in years) the defendant is sentenced to. In column (2) the analysis is done on the whole sample; in column (3), the analysis is restricted to conviction decisions only. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 6: Effect of gender composition of the team on probability of appeal

	(1) Unconditional	(2) Conditional	(3) Controlling for jailtime
All-men team	0.082**	0.191***	0.179***
	(0.037)	(0.056)	(0.056)
2M1F team	-0.009	0.044	0.035
	(0.017)	(0.030)	(0.030)
1M2F team	-0.008	0.009	0.005
	(0.016)	(0.028)	(0.027)
Observations	65051	27686	27686
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.190	0.422	0.422

Notes. In all columns, the dependent variable is an indicator for whether the team's decision is appealed against. In column (1), the analysis is done on the whole sample; in columns (2) and (3), the analysis is restricted to conviction decisions only. Column (3) also controls for the length of jailtime. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 7: Effect of gender composition of the team on appeal-proof decisions

	(1) Unconditional	(2) Conditional
All-men team	-0.087**	-0.177***
	(0.034)	(0.051)
2M1F team	-0.009	-0.067**
	(0.014)	(0.026)
1M2F team	-0.014	-0.050**
	(0.013)	(0.025)
Observations	65051	27686
Court and crime x year FEs	X	X
Controls	X	X
Mean dep.var.	0.867	0.714

Notes. In all columns, the dependent variable is an indicator for whether the team takes an *appeal-proof decision*, defined as a decision that is either not appealed against or it is appealed and confirmed. In columns (1), the analysis is done on the whole sample; in columns (2) , the analysis is restricted to conviction decisions only. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 8: Estimated parameters for the structural models

	Same β different ϕ (1)	Different βs different ϕ (2)	Different β same ϕ (3)
β_m		1.258 (0.003)	1.211 (0.006)
β_f		3.109 (0.011)	6.940 (0.051)
β	1.476 (0.011)		
ϕ_m	1.392 (0.005)	1.000 (fixed)	
ϕ_f	10.041 (0.033)	7.469 (0.022)	
ϕ			7.975 (0.053)
p	0.672 (0.003)	0.675 (0.004)	0.740 (0.002)
q_0	0.748 (0.004)	0.714 (0.002)	0.616 (0.002)
q_{mix}	0.5202 (0.003)	0.573 (0.002)	0.604 (0.001)
Moments used	6	6	6
Estimated parameters	6	6	6
Goodness of fit	0.820	0.086	1.143

Notes. The table displays parameter estimates for the three versions of the structural model described in Section 6.2. The parameters are estimated by minimizing the distance between the target empirical moments and the same moments as predicted by the models for a given vector of free parameters.

Table 9: Model fit for the structural models

	Empirical moment [95% CI]	Simulated moments		
		Same β different ϕ	Different β different ϕ	Different β same ϕ
	(1)	(2)	(3)	(4)
C_{AW}	0.484 [0.446,0.522]	0.586	0.597	0.638
$C_{AW} - C_{AM}$	0.066 [0.005,0.127]	0.067	0.089	0.038
$C_{AW} - C_{MG}$	0.072 [0.028,0.116]	0.028	0.077	0.080
APC_{AW}	0.767 [0.732,0.802]	0.862	0.874	0.930
$APC_{AW} - APC_{AM}$	0.175 [0.074,0.276]	0.151	0.147	0.055
$APC_{AW} - APC_{MG}$	0.057 [0.012,0.101]	0.053	0.059	0.021

Notes. The table reports the fit of the three models in Table 8 with respect to the target empirical moments capturing conviction rates and appeal-proof conviction rates in our data.

Table 10: Effect of gender composition of the team on decision-making time

	Length of trial (days)
	(1)
All-men team	-252.781***
	(87.487)
2M1F team	-46.731
	(61.851)
1M2F team	-106.494*
	(56.782)
Observations	17106
Court and crime x year FEs	X
Controls	X
Mean dep.var.	845.6

Notes. In all columns, the dependent variable is the length of the trial (in days). The analysis excludes the Court of Naples for which this data is not available. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 11: Effect of gender composition by case complexity

	(1) Guilty	(2) Appeal-proof	(3) Length of trial
A: More complex cases			
All-men team	-0.073*	-0.188***	-405.945***
	(0.039)	(0.064)	(99.677)
2M1F team	-0.110***	-0.099***	-103.917
	(0.030)	(0.033)	(69.004)
1M2F team	-0.064**	-0.063*	-118.986**
	(0.030)	(0.033)	(55.695)
Observations	49584	18111	11161
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.365	0.681	963.7
B: Less complex cases			
All-men team	-0.033	-0.081**	21.617
	(0.027)	(0.033)	(121.208)
2M1F team	-0.025	0.011	76.581
	(0.021)	(0.022)	(99.559)
1M2F team	0.009	0.001	-46.760
	(0.022)	(0.022)	(103.339)
Observations	15454	9557	5934
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.619	0.775	623.9

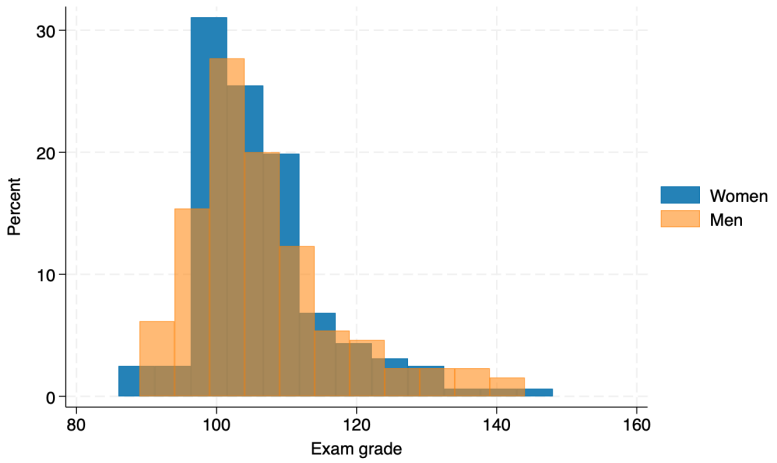
Notes. The dependent variable in columns (1) is an indicator for whether the team convicts the defendant; in columns (2) is an indicator for whether the team takes an *appeal-proof decision*, defined as a decision that is either not appealed against or it is appealed and confirmed; in column (3) is the length of the trial (in days). The analysis in column (3) excludes the Court of Naples for which this data is not available. Panel A reports the results for more complex cases (defined as those with more than one defendant/crime), while Panel B reports the results for less complex cases (defined as those with one defendant and one crime). *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team consists of one man and two women. All regressions include court and crime type-by-year fixed effects. Controls include gender and nationality of the defendant, and judges' age, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table 12: Effect of gender composition on written opinions

	(1) Length (words)	(2) Clarity index	(3) N. maxims
All-men team	31597.291***	-1.266***	8.514***
	(7933.499)	(0.298)	(2.397)
2M1F team	3758.976	-0.061	1.221
	(3899.982)	(0.158)	(0.837)
1M2F team	3844.780	-0.000	0.847
	(4188.345)	(0.135)	(0.730)
Observations	18824	18793	18824
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	18932.349	0.000	5.028

Notes. The dependent variable in columns (1) length (in words) of the the written opinion; in columns (2) is a text clarity index built using PCA from the following measures: average sentence length, average word length, average length of root words, frequency of gerunds, demonstrative adjectives, modal verbs, and contingency clauses; in column (3), is the number of times a maxim from the Supreme Court is cited within the document. The analysis excludes the Court of Naples, for which data on written opinions is not available. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team consists of one man and two women. All regressions include court and crime type-by-year fixed effects. Controls include gender and nationality of the defendant, and judges' age, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

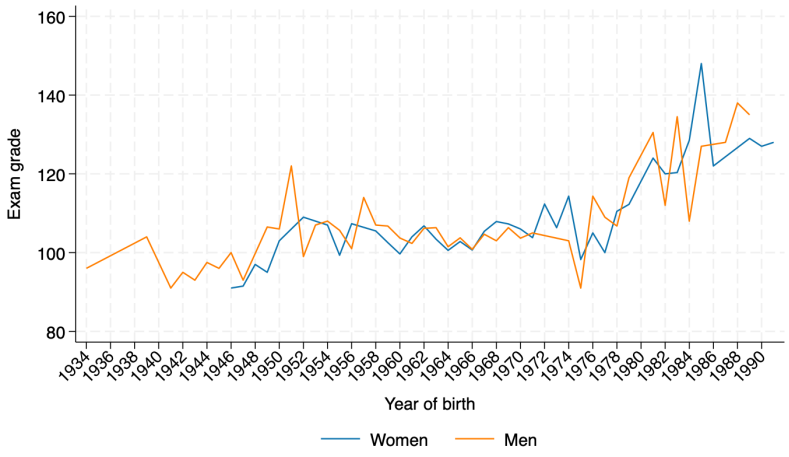
Figure 1: Grades differences between men and women



(a) Distribution of grades by gender



(b) Grades over exam years by gender

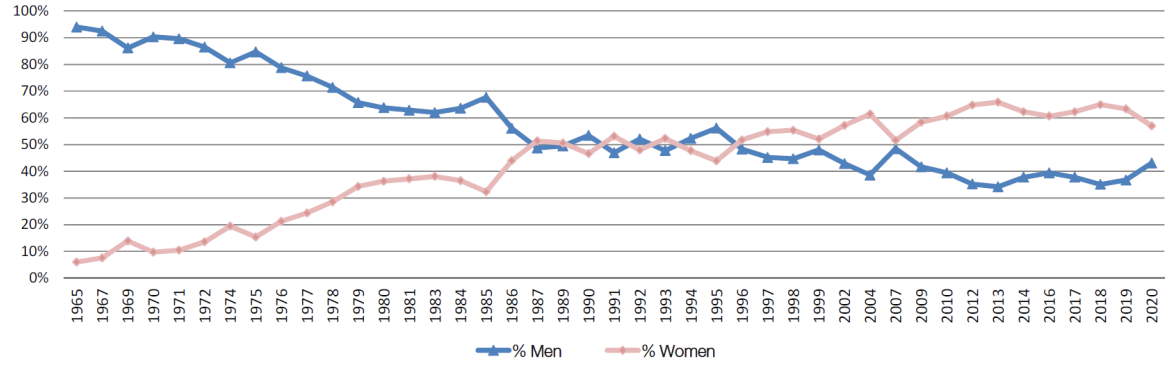


(c) Grades over year of birth by gender

Notes. The figure shows the distribution of the judiciary national exam grades by gender in panel (a); the distribution of grades over exam year for men and women in panel (b); the distribution of grades over cohorts for men and women in panel (c).

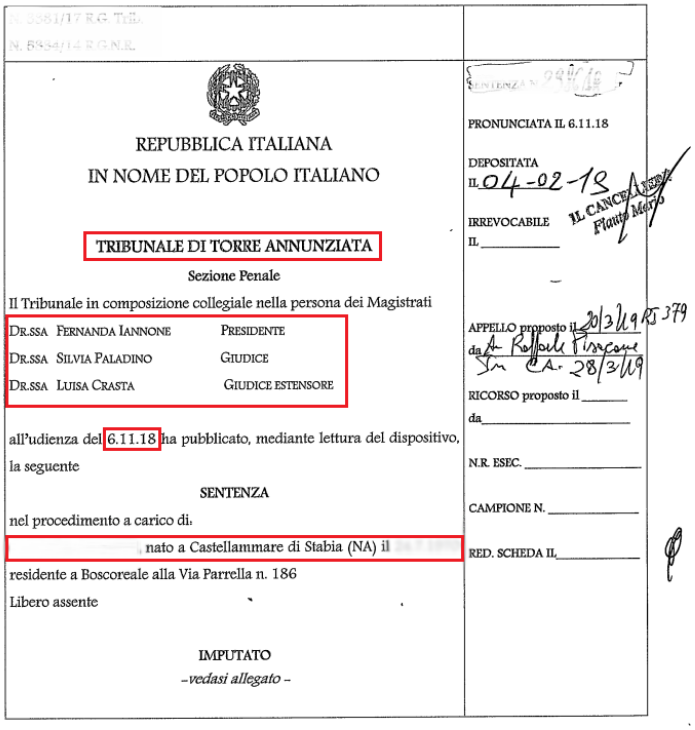
A Other Tables and Figures

Figure A1: Gender composition of the Italian judiciary over time



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

Figure A2: 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 A. Raffaele Fiorenza

RICORSO proposto il

da

N.R. ESEC.

CAMPIONE N.

RED. SCHEDA IL

(a) Court, date, team, 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	25163	38.68
against public order and the state	11559	17.77
against public trust	7651	11.76
against the person	6880	10.58
against the public administration	6051	9.30
against public safety and the environment	4447	6.84
against the economy	1311	2.02
against the administration of justice	1099	1.69
against the family	606	0.93
against public morality and sentiment	284	0.44
Total	65051	100.00

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

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

	Mean	St.Dev.
A. All-men team		
Raw variable	0.059	0.236
Residuals after removing court and crime x year FE	0.000	0.221
B. 2M1F team		
Raw variable	0.354	0.478
Residuals after removing court and crime x year FE	0.000	0.458
C. 1M2F team		
Raw variable	0.402	0.490
Residuals after removing court and crime x year FE	-0.000	0.468
D. All-women team		
Raw variable	0.184	0.388
Residuals after removing court and crime x year FE	0.000	0.374

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 teams in Panel A; teams consisting of two men and one woman in Panel B; teams consisting of one man and two women in Panel C; and all-women teams in Panel D.

Table A3: Effect of conviction and jail time on appeal probability

	Probability of appeal		
	(1) Unconditional	(2) Unconditional	(3) Conditional
Guilty	0.395***		
	(0.018)		
Jail time		0.037***	0.012***
		(0.002)	(0.003)
Observations	65051	65051	27686
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.190	0.190	0.422

Notes. In all columns, the dependent variable is an indicator for whether the team's decision is appealed against. *Guilty* is an indicator for whether the judicial team 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A4: Effect of conviction and jail time on appeal probability

	First instance	Appeal	Difference
Age	49.826	53.853	-4.027***
	(4.832)	(3.167)	
Experience	21.685	25.475	-3.790***
	(5.093)	(3.555)	
Female	0.568	0.628	-0.061***
	(0.279)	(0.277)	
Year of entry	1991.993	1992.018	-0.025
	(6.931)	(5.841)	
Exam score	105.186	104.597	0.589***
	(5.904)	(3.841)	
Exam rank	131.143	136.272	-5.129***
	(58.495)	(52.809)	
N. cases	10676	2531	13207

Notes. The table reports descriptive statistics of first-instance judges (column (1)) and appeal judges (column (2)), as well as differences (column (3)). * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A5: Inclusion\exclusion of controls

	(1)	(2)	(3)	(4)
A. Dep. var: Guilty				
All-men team	-0.013	-0.014	-0.060**	-0.071**
	(0.038)	(0.036)	(0.031)	(0.031)
2M1F team	-0.102***	-0.098***	-0.108***	-0.093***
	(0.027)	(0.026)	(0.027)	(0.025)
1M2F team	-0.060**	-0.059**	-0.049*	-0.057**
	(0.030)	(0.027)	(0.025)	(0.024)
Observations	65051	65051	65051	65051
Mean dep.var.	0.426	0.426	0.426	0.426
B. Dep. var: Appeal-proof (cond.)				
All-men team	-0.263***	-0.197***	-0.166***	-0.177***
	(0.068)	(0.056)	(0.052)	(0.051)
2M1F team	-0.095***	-0.060**	-0.059**	-0.067**
	(0.031)	(0.026)	(0.025)	(0.026)
1M2F team	-0.068**	-0.046*	-0.037	-0.050**
	(0.028)	(0.024)	(0.024)	(0.025)
Observations	27691	27686	27686	27686
Mean dep.var.	0.714	0.714	0.714	0.714
Court FEs	X	X	X	X
Crime x year FEs		X	X	X
Judge-level controls			X	X
Case-level controls				X

Notes. The dependent variable in Panel A is an indicator for whether the team convicts the defendant; in Panel C is an indicator for whether the decision taken is appeal-proof (conditional on conviction). *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator for whether the team consists of one man and two women. 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, experience, ability and area of birth. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A6: Effect of age, geographic, experience and ability composition of the team

	(1) Guilty	(2) Appeal-proof (cond.)
A. Age diversity		
All-old team	-0.027	0.007
	(0.042)	(0.053)
Mixed-age team	-0.024	-0.013
	(0.027)	(0.037)
Observations	60888	26079
Mean dep.var.	0.428	0.710
B. Regional diversity		
All-north team	0.059	0.026
	(0.048)	(0.049)
Mixed-zone team	0.069**	-0.024
	(0.034)	(0.039)
Observations	59379	25625
Mean dep.var.	0.432	0.707
C. Experience diversity		
High-experience team	-0.033	0.020
	(0.069)	(0.064)
Mixed-experience team	0.007	0.009
	(0.053)	(0.044)
Observations	46738	20475
Mean dep.var.	0.438	0.707
C. Ability diversity		
Low-ability team	-0.034	0.025
	(0.039)	(0.067)
Mixed-ability team	-0.022	0.063
	(0.037)	(0.064)
Observations	46738	20475
Mean dep.var.	0.438	0.707

Notes. The dependent variable in column (1) in an indicator for whether the team convicts the defendant; and in column (2) is an indicator for whether the decision taken is appeal-proof (conditional on conviction). Panel A focuses on age diversity; Panel B focuses on geographical diversity; Panel C focuses on experience diversity; and Panel D focuses on ability diversity. *All-old team* is an indicator of whether all judges in the team are above the median age. *Mixed-age team* is an indicator of whether at least one judge in the team is below and one is above the median age. *All-north team* is an indicator of whether all judges were born in central or northern Italy. *Mixed-zone team* is an indicator of whether at least one judge was born in southern Italy and one in central or northern Italy. *High-experienced team* is an indicator of whether all judges in the team are above the median experience. *Mixed-experience team* is an indicator of whether at least one judge in the team is below and one is above the median experience. *Low-ability team* is an indicator of whether all judges in the team are below the median exam grade. *Mixed-ability team* is an indicator of whether at least one judge in the team is below and one is above the median exam grade. The omitted category in Panels A, B, C and D is a all-young team, an all-south team, a low-experience team and a high-ability team, 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' gender, age, region of birth, experience and ability (excluding the category the classification in each panel is referring to). Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A7: Model with judge fixed effects

	Guilty		Appeal-proof (cond.)	
	(1)	(2)	(3)	(4)
	Women	Men	Women	Men
Same-gender team	0.054***	-0.012	0.053***	-0.084**
	(0.021)	(0.020)	(0.018)	(0.036)
Observations	111352	83795	48316	34743
Court and crime x year FEs	X	X	X	X
Judge FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	0.434	0.415	0.740	0.677

Notes. The dependent variable in columns (1) and (2) is an indicator for whether the team convicts the defendant; in columns (3) and (4) is an indicator for whether the conviction is appeal-proof. Columns (1) and (3) focus on women only, columns (2) and (4) focus on men only. *Same-gender team* is an indicator for whether the team is gender-homogeneous. 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 A8: Different classification of crimes

	(1)	(2)
	Guilty	Appeal-proof (cond.)
All-men team	-0.061**	-0.173***
	(0.029)	(0.049)
2M1F team	-0.080***	-0.063**
	(0.022)	(0.025)
1M2F team	-0.047**	-0.046*
	(0.022)	(0.024)
Observations	65051	27688
Court and crime x year FEs	X	X
Controls	X	X
Mean dep.var.	0.426	0.714

Notes. The dependent variable in column (1) is an indicator for whether the team convicts the defendant; in column (2) is an indicator for whether a conviction is appeal-proof. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team consists of one man and two women. 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A9: 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-men team	-0.063**	-0.071**	-0.072**	-0.069**	-0.074**	-0.071**	-0.071**	-0.073**	-0.074**	-0.083**
	(0.031)	(0.031)	(0.031)	(0.031)	(0.030)	(0.031)	(0.031)	(0.031)	(0.034)	(0.039)
2M1F team	-0.086***	-0.094***	-0.100***	-0.097***	-0.077***	-0.095***	-0.093***	-0.094***	-0.099***	-0.087***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.027)	(0.029)
1M2F team	-0.046**	-0.058**	-0.069***	-0.055**	-0.042*	-0.058**	-0.056**	-0.058**	-0.056**	-0.075***
	(0.024)	(0.024)	(0.024)	(0.025)	(0.023)	(0.024)	(0.024)	(0.024)	(0.026)	(0.028)
Observations	59000	63952	53492	60604	57400	63740	64767	64445	58171	39888
Mean dep.var.	0.436	0.427	0.427	0.426	0.441	0.431	0.427	0.424	0.407	0.403
B. Appeal-proof (cond.)										
All-men team	-0.181***	-0.179***	-0.165***	-0.186***	-0.126**	-0.183***	-0.178***	-0.176***	-0.199***	-0.205***
	(0.051)	(0.052)	(0.048)	(0.052)	(0.052)	(0.052)	(0.051)	(0.052)	(0.057)	(0.053)
2M1F team	-0.062**	-0.068***	-0.070**	-0.069***	-0.053**	-0.069***	-0.068**	-0.068***	-0.086***	-0.058**
	(0.027)	(0.026)	(0.027)	(0.027)	(0.025)	(0.026)	(0.026)	(0.026)	(0.028)	(0.029)
1M2F team	-0.042*	-0.052**	-0.058**	-0.050**	-0.035	-0.051**	-0.050**	-0.050**	-0.059**	-0.047*
	(0.025)	(0.025)	(0.025)	(0.025)	(0.023)	(0.025)	(0.025)	(0.025)	(0.028)	(0.028)
Observations	25743	27331	22821	25840	25292	27460	27630	27332	23648	16077
Mean dep.var.	0.713	0.714	0.714	0.714	0.721	0.713	0.714	0.713	0.710	0.708

Notes. The dependent variables in Panels A and B are an indicator for whether the team convicts the defendant and an indicator for whether a conviction is appeal-proof, respectively. In columns (1)-(10) we exclude one specific type of crimes at the time. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A10: Gender and ethnicity of defendants

	Gender of defendant		Nationality of defendant	
	(1)	(2)	(3)	(4)
	Male	Female	Italian	Foreign
A. Guilty				
All-men team	-0.077**	-0.070	-0.060*	-0.115**
	(0.030)	(0.056)	(0.033)	(0.053)
2M1F team	-0.091***	-0.129***	-0.092***	-0.091**
	(0.024)	(0.040)	(0.026)	(0.038)
1M2F team	-0.051**	-0.129***	-0.064**	-0.030
	(0.024)	(0.036)	(0.025)	(0.033)
Observations	57447	7224	56005	8580
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	0.429	0.391	0.409	0.538
B. Appeal-proof (cond.)				
All-men team	-0.163***	-0.306***	-0.191***	-0.104*
	(0.051)	(0.079)	(0.057)	(0.062)
2M1F team	-0.068**	-0.028	-0.068**	-0.047
	(0.027)	(0.047)	(0.030)	(0.041)
1M2F team	-0.053**	-0.024	-0.055*	-0.018
	(0.025)	(0.047)	(0.029)	(0.042)
Observations	24627	2808	22882	4611
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var.	0.712	0.710	0.698	0.784

Notes. The dependent variable in Panel A is an indicator for whether the team convicts the defendant; in Panel B is an indicator for whether a conviction is appeal-proof. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team 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 (only in columns (3)-(4) and (1)-(2), respectively), and judges' age, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A11: Conviction and appeal-proof conviction rates for estimation

	(1) Guilty	(2) Appeal-proof
All-men team	-0.066**	-0.175***
	(0.031)	(0.051)
Mixed gender team	-0.072***	-0.057**
	(0.022)	(0.023)
Observations	65051	27686
Court and crime x year FEs	X	X
Controls	X	X
Mean dep.var.	0.426	0.714

Notes. The dependent variable in column (1) is an indicator for whether the team convicts the defendant; in column (2) is an indicator for whether a conviction is appeal-proof. *All-men team* is an indicator for whether the team consists of men only. *Mixed-gender team* is an indicator for whether the team is mixed-gender. 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, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A12: Role of judges' ability measure

	Guilty		Appeal-proof	
	(1)	(2)	(3)	(4)
All-men team	-0.050	-0.071**	-0.199***	-0.177***
	(0.032)	(0.031)	(0.053)	(0.051)
2M1F team	-0.087***	-0.093***	-0.070**	-0.067**
	(0.025)	(0.025)	(0.028)	(0.026)
1M2F team	-0.051**	-0.057**	-0.058**	-0.050**
	(0.024)	(0.024)	(0.026)	(0.025)
Observations	65051	65051	27686	27686
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Judge ability		X		X
Mean dep.var.	0.426	0.426	0.714	0.714

Notes. The dependent variable in columns (1)-(2) is an indicator for whether the team convicts the defendant; in columns (3)-(4) is an indicator for whether a conviction is appeal-proof. *All-men team* is an indicator for whether the team consists of men only. *2M1F team* is an indicator for whether the team consists of two men and one woman. *1M2F team* is an indicator of whether the team consists of one man and two women. All regressions include court and crime type-by-year fixed effects. Controls in columns (1) and (3) include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age, region of birth, and experience. In columns (2) and (4), we also control for judges' ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

B Comparison with the U.S. criminal justice system

The Italian and U.S. criminal justice systems diverge on several central dimensions, including legal foundations, trial structure, appellate review, jury involvement, and judicial reasoning. Italy belongs to the civil law tradition, with judges applying codified statutes, while the U.S. follows a common-law tradition shaped by binding judicial precedents. Italian criminal trials operate under a mixed model that grants judges limited inquisitorial powers, in contrast to the fully adversarial U.S. framework where judges act primarily as referees. In appellate proceedings, Italian courts may reassess both facts and legal issues, whereas U.S. appeals are largely confined to questions of law. Jury participation is the norm in U.S. trials but limited in Italy, where most cases are judge-only. Finally, Italian judges must provide detailed written justifications for every verdict, unlike in the U.S., where jury verdicts require no reasoning and even judicial opinions often precede, rather than follow, the verdict. Below we review these differences in greater detail.

Legal Foundation: The Italian criminal justice system belongs to the *civil law* family and is grounded in comprehensive legal codes such as the *Codice Penale* (Criminal Code) and *Codice di Procedura Penale* (Code of Criminal Procedure). Judges are expected to apply these written laws in a systematic and structured way, and their decisions do not have binding force as precedent. By contrast, the United States follows a *common-law* tradition in which judicial rulings from higher courts serve as binding precedents (*case law*) and play a central role in shaping future interpretations of the law.

Trial Structure: Italian criminal trials follow a *mixed model* (adversarial with limited inquisitorial powers), in which the judge (or team of judges) plays an active role during the trial.[61] While prosecutors and defense lawyers are responsible for presenting and challenging evidence, judges can question witnesses at the end of the examination, highlight factual issues that need further exploration, and request new expert opinions or additional evidence when necessary. This limited inquisitorial power

[61] Since the 1988 reform of the Code of Criminal Procedure, Italy has adopted an accusatorial (or adversarial) framework while retaining limited inquisitorial safeguards. In an adversarial model the parties control the proof and the judge acts mainly as a referee, whereas an inquisitorial model gives the court a direct duty to investigate the facts and seek the material truth.

contrasts with the fully *adversarial* approach in U.S. courts, where the judge rarely goes beyond ruling on objections and ensuring procedural fairness. This approach reflects the civil law emphasis on the importance of judicial truth-seeking, regardless of how effectively the parties argue their case. In line with this, in Italy plea bargaining (*patteggiamento*) is generally available for offenses with a prison sentence of no more than five years, and is not considered an admission of guilt: the judge, after examining the case, may reject the agreement and acquit the defendant.

Appelate Review: Because the trial judge maintains some fact-finding responsibility, Italian Courts of Appeal may re-examine both the facts and the application of the law. In contrast, U.S. appellate courts review legal and procedural errors only.[62]

Jury Involvement: Moreover, most Italian trials are judge-only and do not involve a jury. A jury appears only in the *Corte d'Assise*, which tries the gravest crimes and sits with a mixed panel of two professional and six lay judges (*giudici togati e giudici popolari*).

Justification of Verdict: Italian judges are required to issue a reasoned judgment that details the logical path and evidentiary evaluation underlying the verdict. For every case – whether resulting in a guilty or not-guilty verdict – judges must provide a written explanation analyzing all relevant evidence, including witness testimonies, expert opinions, and documentary records. These written motivations are themselves subject to appellate scrutiny. Importantly, the verdict is delivered first, after which judges have a legal deadline – typically 30 to 90 days, though longer for complex cases – to file the written justification. This procedure applies both to individual judges and to collegial courts. By contrast, in the United States, reasoning often precedes the verdict, and in jury trials verdicts are rendered without any written justification, leaving appellate courts to assess only the sufficiency of the evidence rather than the logic of the decision.

[62]U.S. appellate courts can rule on factual findings only under narrow "clearly erroneous" or "no reasonable jury" standards.