Gender diversity and decision-making in teams

Maddalena Ronchi Viola Salvestrini*

March 26, 2026

Abstract

We study how gender composition affects the nature and quality of team decisions using a novel dataset on collegial rulings from Italian criminal courts. Exploiting the quasi-random assignment of judges and cases to panels, we find that all-women teams convict more often and issue the most appeal-proof rulings (unappealed or upheld on appeal). All-men teams convict less frequently and have the lowest appeal-proof rate, while mixed-gender teams resemble all-men teams in conviction rates but nearly match the conviction accuracy of all-women teams. Higher conviction rates and fewer reversals for all-women teams rule out differences in leniency across gender compositions as the main explanation, while mixed teams' lower conviction rate relative to all-women teams despite similar accuracy points to coordination frictions. We formalize these patterns in a model of diagnostic decision-making where judges differ in leniency and diagnostic skills shaped by effort or ability. In line with the reduced-form evidence, model estimates and supplementary analyses of written opinions, deliberation times, and entry exam scores indicate that effort differences are the main channel and that frictions in mixed-gender teams reduce performance relative to all-women teams, without eliminating their accuracy edge over all-men teams.

JEL codes: J16, J82, M5, K4

*Ronchi: Northwestern University, IZA, and CESifo. Salvestrini: Bank of Italy and IZA. We thank François Gerard and Anna Raute for their support and guidance. For very helpful comments and suggestions we thank Ricardo Alonso, Manuel Bagues, Jan Bakker, Jordi Blanes i Vidal, Wouter Dessein, Florian Englmaier, Elisa Facchetti, Alex Frenkel, Alfonso Gambardella, Andrea Ichino, Ambar La Forgia, Mikkel Mertz, Elia Moracci, Massimo Morelli, Barbara Petrongolo, Nicola Persico, Fabiano Schivardi, Martina Zanella, as well as seminar participants at the Conference in honor of Ghazala Azmat, NBER Organizational Economics Fall 2025, Bocconi University, Einaudi Institute for Economics and Finance (EIEF), European University Institute, Queen Mary University, Oxford University, LMU Munich, Regensburg University, 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 Economics 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; Prat, 2002; Lazear and Shaw, 2007). Since men and women differ in preferences, attitudes, social values, and skills (e.g. Chattopadhyay and Duflo, 2004; Croson and Gneezy, 2009; Kim and Starks, 2016; Schaede and Mankki, 2022; Truffa and Wong, 2025), 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; O*NET, 2022) – 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.[1,2] This exogenous variation in team gender-mix within-court allows us to causally estimate whether gender composition affects judicial rulings. 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 accuracy of judicial team decisions (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-men, all-women, and mixed-gender teams separately.[3] This is important, as it provides the variation needed to determine whether gender homogeneity masks potential differences between all-men and all-women teams.

Leveraging these 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-

[1]Constitutional provisions require the Italian judicial system to allocate both judges and cases to panels through quasi-random procedures that preclude discretion or manipulation, ensuring that neither team composition nor case assignment depends on gender. We describe the assignment rules in Section 2.1 and formally test their validity in Section 4.2, showing that: (i) within-court variation in team gender composition is unrelated to case characteristics; (ii) judges in mixed- and same-gender teams do not differ on observables; (iii) the observed distribution of teams by gender composition is consistent with random formation.

[2]Panel decisions are taken collectively, and convictions are treated as requiring unanimity.

[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.

proof" decisions, i.e., rulings that are either unappealed or upheld on appeal.[4] Second, we determine whether observed differences in outcomes across teams with different gender compositions reflect differences in leniency or in the accuracy of their case assessments, and whether any differences in such diagnostic performance are more consistent with fixed ability or with endogenous effort, drawing on rich data on cases and judges' written opinions. Third, guided by these reduced-form findings, we develop and estimate a model of team diagnostic decision-making that formalizes the mechanisms suggested by the data and can rationalize the full set of empirical patterns.

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, our dataset includes more than 65,000 judicial decisions. Finally, we augment our dataset with judges' entry-exam scores, which we digitize from nationwide public competitions for judicial appointments.

Using a specification with court fixed effects and crime-by-year fixed effects – thus comparing teams within the same court ruling on comparable crimes – we first show that judicial teams reach different sentencing outcomes depending on their gender composition. All-women teams are 7.1 percentage points (13.7%) more likely to convict than all-men teams, and 5.7–9.3 percentage points (13.6-24.1%) more likely to convict than mixed-gender teams, depending on whether the latter include one or two women. 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 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. As with conviction rates, sentencing by all-men and mixed-gender teams is not statistically different.

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

[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 appellate courts are not systematically biased with respect to the gender composition of the panel of first-instance judges and are at least as accurate as first-instance courts. 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 should not be specific to a particular type of court or area.

the *perceived accuracy* of a ruling, proxied by its likelihood of being appealed. Our analysis shows that decisions issued by all-men teams are significantly more likely to be appealed than those of any other team type, while mixed-gender and all-women 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-men teams is not merely a response to the longer sentences they impose. We then measure teams' *appeal-proof decision rates* by examining whether a ruling is either not appealed or appealed and confirmed by the Court of Appeal.[6] All-men teams are 17.7 percentage points (22.6%) less likely than all-women teams to issue conviction decisions that are appeal-proof. The presence of at least one woman substantially improves outcomes: mixed-gender teams' appeal-proof conviction rate is just 5-7 percentage points lower than that of all-women 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.

The combination of higher conviction rates and fewer reversals for all-women teams rules out differences in judicial leniency as the main driver of our results. If gender composition affected only the propensity to convict, but not the evaluation of the case, teams that convict more would do so by bringing in weaker marginal cases and should therefore experience more appeals and reversals. Instead, all-women teams convict more *and* are more accurate – a pattern more consistent with differences in teams' ability to distinguish guilt from innocence than with leniency alone.[7] Finally, the patterns for mixed-gender teams point to coordination frictions as an additional mechanism. Mixed-gender teams achieve conviction accuracy comparable to all-women teams yet convict less often, suggesting that converting similar evidence into a unanimous verdict is harder in gender-diverse teams. At the same time, the presence of at least one woman yields an accuracy advantage over all-men teams, indicating that these frictions reduce how often mixed teams convict without undermining the quality of convictions when they do occur.

Women's greater diagnostic ability could in principle reflect innate talent or effort. Although disentangling the two is inherently difficult, we show additional results consistent with an effort-based channel and with frictions in mixed-gender teams. First, digitized rankings from nationwide public exams to enter the judiciary show no gender differences in performance, indicating that women are not positively selected at entry. As we discuss in Section 2, this measure of baseline ability is unique to the Italian institutional setting

[6]In Italy, the Court of Appeal may reassess the evaluation of the evidence (on the existing record) as well as the application of the law. Thus, "appeal-proof" outcomes reflect both substantive correctness and procedural/legal soundness.

[7]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.

and sets our study apart from the U.S. literature, which cannot construct comparable measures given that judges in the U.S. are elected. Second, all-men teams decide cases faster than both mixed-gender and all-women teams, who instead take equally long. This finding yields two insights. If higher decision quality for all-women teams were due solely to innate talent, it could be achieved without longer deliberation; instead, all-women teams obtain better outcomes *and* spend more time per case, consistent with greater effort.[8] Moreover, the fact that mixed-gender and all-women teams deliberate equally long but reach different conviction rates is consistent with 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-men teams produce judgments that are less clear and concise than those of other teams. By contrast, mixed-gender and all-women teams are not statistically different across the measures used, suggesting similar levels of effort devoted to producing clear and concise judicial reasoning.

To formalize this interpretation, we then propose and structurally estimate a model of team decision-making in the spirit of Chan et al. (2022) that introduces: (i) heterogeneity in diagnostic skills, modeled via endogenous effort provision, and (ii) within-team frictions arising from gender diversity. A panel of judges must decide whether to convict a defendant whose true guilt they cannot observe. Mistakes are costly and asymmetric: judges dislike both wrongful convictions and wrongful acquittals, but a wrongful conviction carries a larger cost. We define judges' overall aversion to mistakes as their diligence, and the additional weight placed on wrongful convictions as their leniency. To reduce the risk of error, each judge chooses how much costly effort to devote to analyzing and interpreting the evidence of the case. Their joint effort determines the quality of the information environment; based on this common body of evidence, each judge independently forms a binary assessment of the defendant's guilt. Greater effort increases the accuracy of these assessments, which raises both conviction rates and conviction quality. When both judges reach the same assessment, they follow it. When they disagree, they enter a deliberation process. With some probability they reach agreement and one judge's view prevails; if instead they do not, the panel acquits by default. Greater frictions therefore reduce convictions but, because they do not affect the effort judges exert or the precision of their signals, they do not reduce the quality of convictions when they do occur.

We structurally estimate the model and find sizable gender gaps in diligence – higher for women – and smaller, though still positive, gaps in leniency. As predicted by the model, women's higher diligence levels lead them to exert more effort than men, which, consistent with our reduced-form results, translates into higher conviction rates and more appeal-proof

[8]Evidence consistent with women exerting more effort than men when completing comparable tasks has been found in other real-world settings, including medicine, academia, and corporate boards (Coviello et al., 2025; Chu et al., 2022; Hengel, 2022; Gupta et al., 2019; Adams and Ferreira, 2009). This greater effort to avoid mistakes is consistent with evidence that women are punished more harshly for identical errors and receive weaker support at work despite similar performance (Gagliarducci and Paserman, 2012; Egan et al., 2022; Sarsons, 2024; Dupas et al., 2025).

convictions in all-women teams compared to all-men teams. Because each judge's optimal effort decreases with teammates' diligence and leniency, men in mixed-gender teams free-ride (exerting less effort than in all-men teams) while women over-compensate (working harder than they do in all-women teams), to the extent that mixed-gender teams' total effort – and thus diagnostic precision – exceeds that of all-men teams and nearly matches that of all-women teams. As a result, their convictions are more accurate than all-men teams' and close to all-women teams. Finally, we estimate greater coordination frictions in mixed-gender panels, which explains why mixed teams convict less frequently than all-women teams, despite exhibiting similar levels of accuracy.

We then use the estimated model to quantify the sources of cross-team performance gaps and to evaluate how the growing presence of women in the judiciary is likely to affect aggregate court performance.[9] Shutting down one channel at a time, we find that the diligence gap is the dominant driver of performance differences: equalizing men's diligence to women's level would largely close the quality gap across team types, whereas equalizing leniency alone would not. Simulating changes in the share of women in the judiciary pool, we find that increasing female representation raises aggregate decision accuracy but at the cost of longer trials, revealing a quality–speed tradeoff that will become increasingly salient as this demographic trend continues to reshape the composition of judicial panels.

This paper contributes to two main strands of the literature. First, it adds to research on gender and group decision-making. Within this literature, one set of studies investigates whether gender diversity within decision-making bodies (e.g., hiring committees, corporate boards) benefits women among those being evaluated (Bagues and Esteve-Volart, 2010; Bagues et al., 2017; Bertrand et al., 2019; Deschamps, 2024). Another group of studies examines how women's relative presence in a group shapes their influence on collective decisions, emphasizing how gender composition affects the aggregation of information and preferences across members (e.g., Born et al., 2022; Karpowitz et al., 2024; Radbruch and Schiprowski, 2024; Rebolledo et al., 2024). Our study differs from both groups of prior work in its focus. Relative to the first, we do not study homophily between evaluators and those evaluated, and in our setting we find no association between the gender mix of judicial panels and defendants' gender. Relative to the second, we move beyond internal decision dynamics to examine their ultimate consequence – overall team performance. This is possible because our setting offers a natural benchmark for the accuracy and quality of team decisions - appeal outcomes. Leveraging this feature, we develop a conceptual framework that disentangles the roles of gender differences in preferences (leniency), skills (via endogenous effort), and within-team coordination frictions in explaining performance differences across teams with different gender mixes. Using this framework, we show that both effort-driven diagnostic ability and coordination costs are necessary to explain the observed outcomes, thereby clarifying the behavioral and interactional channels through which gender composition shapes collective decision-making and performance. In doing so, we also advance the largely experimental

[9]Women now constitute the majority of new entrants into the Italian judiciary, a sharp reversal from the profession's historically male-dominated composition (Appendix Figure A1).

literature on the relation between gender diversity and team performance (e.g., Apesteguia et al., 2012; Hoogendoorn et al., 2013; Aparicio Fenoll and Zaccagni, 2022; Hardt et al., 2024; La Forgia and Singh, 2025).[10]

In addition, this paper relates to research on differences in decision-making across judges. Most of this work documents biases linked to judges' (or jurors') and defendants' demographics – 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; Abrams et al., 2022; Ash et al., 2023, 2024) – with little to no focus on the determinants of decision quality. We advance this work by leveraging the fact that we can simultaneously observe conviction and appeal rates, allowing us to relate differences in judicial decision-making not only to judicial preferences or bias, but also to differences in diagnostic ability.[11] This result has wider implications for the "judges design" literature, which uses decision-makers' propensities as instruments in courts and beyond (e.g., Maestas et al., 2013; Galasso and Schankerman, 2015; Dobbie et al., 2018; Bernstein et al., 2019). As emphasized by Chan et al. (2022) and Mungan (2024a), the key monotonicity assumption this design relies on holds only when heterogeneity in decision-making is purely preference-based (e.g., leniency) and not skill-based (e.g., talent or effort). Our evidence of systematic variation in diagnostic ability across team compositions weakens the validity of this assumption and cautions against designs that implicitly assume uniform skills across judges. Moreover, by extending the single-agent insights of Chan et al. (2022) and Mungan (2024a) to a team setting, we show that when diagnostic skill stems from endogenous effort, effort provision itself becomes team-dependent: it responds to teammates' incentives through strategic interactions. Consequently, differences in diagnostic ability across agents are shaped not only by individual traits but also by who sits together on the panel.

2 Institutional setting

The Italian judicial system, which is constitutionally independent from the legislative and executive branches of government, 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. Crucially for our purposes, Italy's civil-law tradition – built around comprehensive legal codes and professional judges rather than judge-made precedent and jury

[10] 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).

[11] Using appeal outcomes as a performance benchmark is typically infeasible in the U.S. for three reasons: (i) the system is dominated by guilty pleas – about 90% of convictions are pleas (Motivans, 2025) – often with appeal waivers, so few convictions arise from trials and even fewer receive appellate review; (ii) in most criminal cases juries render the verdict, so trial outcomes largely reflect juror fact-finding leaving little scope to isolate judges' diagnostic ability; (iii) data allowing case-level linkage between trial and appellate system is usually not available.

fact-finding – provides a more standardized and judge-centered environment in which decisions are written, reasoned, and attributable to judicial actors. This institutional structure makes the Italian setting particularly well suited to our research question, and in several respects more informative than the U.S. common-law context, where the prominence of precedent, the frequent involvement of juries, and different trial and appellate dynamics can complicate the mapping from judicial composition to observable case outcomes.[12]

Our analysis focuses on the criminal proceedings adjudicated by criminal courts (*Tribunali ordinari*).[13] 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*.[14] We focus on the decisions of these three-judge panels.

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 rise in the share of women entering the judiciary, and, in recent years, women constitute approximately 60% of new entrants in the judiciary. The fact that the judiciary has been a relatively gender-balanced profession for several decades allows us to observe the effects of both all-women and all-men teams within the same profession. This provides the necessary variation to assess whether gender homogeneity conceals potential differences between all-men and all-women teams - a comparison typically not feasible in previous real-world studies focused on male-dominated environments (Bagues and Esteve-Volart, 2010; Bagues et al., 2017; Radbruch and Schiprowski, 2024; Rebolledo et al., 2024, e.g.), where all-women teams are too rare to be analyzed separately. 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 both by year of birth and year of examination). We return to this point in Section 6.1.

[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. Appendix B offers a comparative overview and explains why the Italian setting is particularly well-suited to our research question relative to the U.S. context.

[13]It therefore excludes all 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.

[14]The offenses under the jurisdiction of collegial teams are regulated by law (Art. 33 bis c.p.p.) and include crimes carrying maximum sentences of 10–24 years, as well as offenses of significant criminal relevance or social alarm not pertaining to the Court of Assize. Examples include 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.

2.1 Random assignment of judges and cases to teams

Italy's judicial system allocates judges and cases to panels through quasi-random procedures that are designed to preclude 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 empirically assess the exogeneity of both team and case allocation in Section 4.2.

Assignment of judges to judicial teams. Each court adopts an organization plan – reviewed and approved every four years by the High Council of the Judiciary (Consiglio Superiore della Magistratura, CSM) – which sets out the criteria for assigning judges to three-member teams.[15] 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.[16] Other courts adopt rotating team systems, in which team composition is reshuffled at regular intervals according to pre-established schedules. Regardless of the model, any reshuffling applies only to new 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.).[17] Thus, a judge may serve on multiple panels during the same period.

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 offense 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 randomization method, as long as the criteria are publicly specified in its organization plan and abide by the principles of impartiality and neutrality.[18] Judges cannot

[15]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.

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

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

[18]The primary tool used in most courts is GIADA (*Gestione Integrata delle Attività e dei Dati Amministrativi*), a centralized 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 an-

decline an assigned case except under limited statutory grounds for recusal, such as illness or extended leave, 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.

Decision-making process. Within the team, judges assume different roles. The *presiding judge* is assigned according to the organizational tables by the president of the court, based on seniority and experience. The presiding judge directs the trial, maintains order in the courtroom, manages the hearing, and coordinates the deliberation. The *judge rapporteur* is entrusted with drafting the reasoning of the judgment: this role may be performed by the presiding judge or delegated to another panel member at the time of deliberation. Decisions are taken collectively in the deliberation chamber. Although rulings are formally adopted by majority vote, interviews with judges indicate that, in practice, formal voting does not occur, and convictions are treated as requiring unanimity.

2.2 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.[19] With the same mechanism discussed in Section 2.1, 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. Appeals may be initiated by either the defendant or the public prosecutor; when only the defendant appeals, the Court of Appeal can either uphold the initial sentence or reduce it, creating an incentive for defendants to appeal. As a result, appeals are overwhelmingly advanced by the defendants: between 2012-2016, public prosecutors file only 5% of all appeals (Della Torre, 2022).

The Court of Appeal may confirm, modify, or overturn the original ruling, based on a full review of the case, reassessing both the evaluation and interpretation of the *evidence* – such as the reconstruction of the events, or the credibility of witnesses – and the application and interpretation of the *law*. 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.[20] This makes the Court of Appeal the final venue where errors

other case of the same type until all other teams have received one, ensuring a fair and balanced distribution of workload.

[19]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.

[20]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.

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.[21]

2.3 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.[22] Unlike in the United States, judges are not elected; instead, entry into the judiciary is obtained through a nationwide public examination.[23] As we discuss later in the paper, this institutional feature, combined with our data collection and digitization effort, provides the opportunity to use judges' exam scores as a measure of their ability at the time of entry. Candidates must hold a law degree and a two-year specialization diploma for legal professions. 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 (CSM) publishes a list of vacancies in trial courts. The CSM matches candidates to vacancies using a transparent formula that combines (i) the candidate's ranking, (ii) their expressed preferences, and (iii) organizational needs.[24]

Judges receive a fixed salary that increases with seniority, according to a structured pay scale and 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's written judgments).[25]

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.,

[21]The Supreme Court of Cassation rules only on matters of law, ensuring the correct and uniform application of the law nationwide. A sentence confirmed by the Supreme Court of Cassation is final, 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.

[22]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 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.

[23]The exam, administered by the Ministry of Justice, includes three written papers - on civil, criminal, and administrative law - followed by an oral exam covering all substantive and procedural areas of law.

[24]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 (defined later in this section), family circumstances, and the staffing priorities of the destination court.

[25]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. However, these are extremely rare events: according to data analyzed by the CSM Statistical Office, more than 98% of evaluation have a positive outcome.

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.[26] Finally, judges cannot be dismissed or demoted except through formal disciplinary proceedings conducted by the CSM.

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). 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 helps ensure that our estimates are not driven by the idiosyncrasies of a single court or a single geographic area.

The sentences were provided in paper format by the presidents of each court.[27] We digitized the sentences issued in collegial rulings and processed them using 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.[28] Appendix Figure A2 presents an example sentence, with the relevant data fields highlighted in red.

We also obtained and digitized the official rankings and scores from the national selection exams required for entry into the judiciary. These data allow us to construct a measure of judges' work experience (captured by the time elapsed since they passed the entry exam) as well as a proxy for their ability at entry, understood in terms of the qualities the judiciary

[26] 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.

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

[28] 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.

itself considers most important at the time of entry.[29] Finally, for 96.2% of our judges we retrieved additional personal information – such as judges' age, municipality, and region of birth – through online sources.

3.1 Descriptive statistics

Table 1 provides descriptive statistics for our sample, which covers the years 2004 to 2021.[30] 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 decision constitutes the unit of observation. This is the relevant level for our analysis, as verdicts pertain to a specific defendant–charge pair, and parties can appeal a single decision rather than the entire case.

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% are women. This gender distribution is fairly consistent across courts, with the share of women ranging from 50 to 59% (not reported in the Table). The average age of judges is 49 years, and nearly 20% 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% vs 24%, 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% of the teams consist of two men and one woman, while 41.7% include one man and two women. All-women teams are more common than all-men teams (18.4% compared to 6.3%).[31] 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%), and 18% 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% of cases. Conditional on conviction, the average prison sentence is five years. Trials last on average a little over two years. Just under 20% of rulings in our sample are appealed, nearly all of which are convictions (97%, not reported in the Table). This implies that 42% 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% of all offenses – are crimes against

[29] 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). Overall, we were able to match 92.5% of all judges in our sample who undertook a national exam.

[30] 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.

[31] 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.

property, public order, public trust, and the person. The first category includes offenses against property involving violence or threats (such 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, possible unobservables characteristics of the cases or the judges are not systematically different across judicial teams of a given gender-mix. As seen in Section 2.1, while judges may choose which court to join based on its overall characteristics, they have no discretion over their specific assignment to a case and to a panel. We provide further empirical evidence in favor of the validity of this assumption in Section 4.2.

Second, measuring decision accuracy with appeal outcomes relies on the assumption that appellate review is independent of the trial panel's gender mix, and that appellate judges are at least as accurate as first-instance judges; under these conditions, decisions that are reversed on appeal are, on average, less likely to be correct than those that are not.[32] If appellate courts tend to agree with lower courts merely because they share a similar gender composition (e.g., both female-dominated), then an upheld decision could reflect

[32] A similar assumption has been made before in the literature. For instance, Alesina and La Ferrara (2014) use higher-court reversals to proxy the quality and impartiality of lower-court rulings, and Coviello et al. (2019) similarly treat the probability of appeal as an indicator of decision quality in their analysis of judicial specialization.

compositional alignment rather than the accuracy of first-instance decisions. We assess this by testing whether the appeal panel's gender mix differentially affects the probability that convictions issued by all-women, mixed-gender, and all-men trial panels are upheld. Specifically, we interact indicators for each trial-panel type with an indicator for whether the appellate panel's gender composition matches it. Appendix Table A4 shows that a matching gender mix between appellate and trial panels does not change the probability that a conviction withstands appeal, providing no support for this alternative explanation.[33] Relatedly, our interpretation would also be violated if the appeal court judges were of *lower* diagnostic ability. Note that we do not need appellate courts to be strictly more accurate and skilled than trial courts: even if their accuracy were comparable, having a second, independent assessment makes an affirmed decision more likely to be correct than a reversed one. To support this assumption we compare the characteristics of appellate judges with those of first-instance judges. Appendix Table A5 shows that appellate judges have more experience and higher baseline ability – proxied by entry-exam performance – making it unlikely that they are systematically less accurate than trial judges.

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 one if the team delivers a guilty verdict (conviction rate), or (ii) an indicator equal to one 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. Thus, the omitted category is a women-only team, $F3_{isct}$. Because of random assignment, β_1, β_2, and β_3 describe the causal effect of having zero, one, or two women on the judicial team - each relative to a team with three women - on the outcomes.

The coefficients γ_c and $\gamma_{s,t}$ denote court fixed-effects 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

[33] Our interpretation would also be invalid if appellate courts were biased towards lower-court panels with a particular gender composition irrespective of the appellate panel's own mix. Note that, however, to generate our results such bias would need to run *in favor* of women – the opposite of what has been documented in a variety of settings including private firms, academia, the judiciary, orchestras, or the classroom (Ash et al., 2022; Benson et al., 2021; Card et al., 2020; Bagues et al., 2017; Boring, 2017; Bagues and Esteve-Volart, 2010; Goldin and Rouse, 2000).

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.[34] 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.[35]

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% of the original variation remains for all-men teams, 95.8% for teams with two men and one woman, 95.5% for teams with one man and two women, and 96.4% 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 the gender mix of a team 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, 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-men teams, two-men–one-woman teams (2M1F), one-man–two-women teams (1M2F), and all-women 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% significance level, which is what would be expected by chance. This suggests that, within a court, 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

[34] Estimating a more restrictive model that includes court-by-year fixed effects yields virtually identical results (Appendix Table A8).

[35] 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.

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. Only one of the coefficients is statistically significant at the 10% significance level, which is again what would be expected by chance. 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. Specifically, we repeatedly sample triads of judges without replacement from each court's judge pool – preserving the court's actual number of male and female judges and the observed number of panels – 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, since 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 (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 set of team-level characteristics: each judge's age, experience, region of origin.[36] 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.[37] 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 5.3, none of these alternative forms of diversity explains our findings: they are not systematically associated 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

[36] 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.

[37] Although men and women do not differ in their performance on the national entry exam, we also include exam scores to capture ability at entry.

Equation (1). All-women teams are more likely to issue a conviction decision than any other type of team: they are 7.1 percentage points (13.7%) more likely to convict the defendant than all-men teams, and 5.7-9.3 percentage points (13.6-24.1%) 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 show meaningful variation within mixed-gender teams depending on whether they include one or two women.[38]

We complement this evidence by investigating whether teams make different sentencing decisions depending on their gender composition. Specifically, we estimate the effect of team gender composition on 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 differences at the extensive and the intensive margins (conviction rates and sentence length), 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.

5.1.1 Interpreting the results

Overall, we find that all-women teams convict more often than mixed-gender and all-men teams, but – conditional on conviction – impose shorter prison sentences. Taken in isolation, this pattern is consistent with multiple mechanisms. A first interpretation, consistent with the "judges design" literature, is that panels share a common assessment of cases but differ in leniency as a function of their gender composition. In this view, teams agree on how cases rank in terms of the likelihood of guilt but differ in the cutoff for conviction. Applied to our setting, this interpretation implies that all-women teams use a lower evidentiary threshold, so their higher conviction rate is driven by marginal cases with weaker evidence. If so, all-women teams' convictions should be less robust on average, and therefore more likely to be appealed and overturned.[40]

A second interpretation is that judicial panels with different gender composition differ in their diagnostic performance, i.e., their ability to distinguish guilty from innocent defendants holding the distribution of cases fixed. If all-women teams assess guilt more accurately, they can convict more often by reducing wrongful acquittals without increasing – and possibly

[38]The difference in conviction rate between teams with two men and one woman and teams with one man and two women is only marginally statistically significant (p-val 0.08), and it is small relative to the difference with all-women teams.

[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).

[40]Shorter prison sentences could also arise mechanically if marginal convictions involve less severe conduct or fewer aggravating circumstances, or if teams impose lighter punishment when residual doubt remains.

while reducing – wrongful convictions. Under this mechanism, higher conviction rates do not reflect a different cutoff applied to the same ordering of cases; instead, all-women teams may process and weigh the evidentiary record differently, improving the accuracy of their decisions.

In the next subsection, we leverage the appellate structure of the Italian judiciary to distinguish between these mechanisms. We use the appeal outcomes as an empirical proxy for the accuracy of first-instance decisions. If the conviction gap reflects a lower evidentiary threshold, convictions by all-women teams should be less likely to withstand appellate review, especially after conditioning on sentence length. By contrast, if the gap reflects differences in diagnostic performance, the additional convictions generated by all-women teams should be at least as accurate as those of other teams, and therefore no more likely to be reversed on appeal.

5.2 Accuracy of decisions

To study how a panel's gender composition affects decision accuracy, we proceed 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 – with the underlying assumption that unappealed decisions are perceived as more likely to be accurate. 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.[41] Second, we investigate the effect of team gender composition on the *effective* accuracy of judicial decisions exploiting whether they are *ever overturned* on appeal. Specifically, we define a ruling as *appeal-proof* if it satisfies one of the following conditions:

(a) it has not been appealed; or

(b) it has been appealed and both (i) the conviction/acquittal decision is confirmed and (ii) if the defendant was convicted, the prison sentence is confirmed in length.[42]

Two features of our appeal-proof measure are worth highlighting. First, because it is defined for all rulings regardless of whether they are appealed, it does not condition on appeal and is therefore not mechanically affected by differences in the composition of appealed cases across team types. Second, since appeals in our data almost exclusively follow from conviction decisions, our ability to assess accuracy is one-sided: we can observe whether a conviction was accurate, but we cannot directly assess the accuracy of acquittals. While this limits our ability to fully evaluate judicial accuracy, the limitation is substantially mitigated by

[41] In our data, 94% of appeals stem from convicted cases. Being convicted increases the probability of appeal by 39.5 percentage points ($\approx 200\%$), and each additional year of prison time increases the probability of appeal by 1.2 percentage points (Appendix Table A3).

[42] According to this definition, 87% of rulings, and 71% of convictions are considered "appeal-proof". Appendix Table A6 shows that our results are robust to relaxing condition (b-ii) to allow for a degree of variation in jail time between the first-instance and appellate decisions. Specifically, we show that the results are unchanged when we classify a decision as appeal-proof if it is unappealed or, if appealed, confirmed and the absolute percentage difference in prison time is below the sample median (7.6%).

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 conviction accuracy must reflect differences in judicial behavior rather than case composition. As a result, if one team convicts more often and its convictions are not less likely to be appeal-proof, we can infer that it makes not only fewer inaccurate convictions but also fewer inaccurate acquittals – allowing us to draw meaningful conclusions about overall team performance from the observable component of decision accuracy.

Perceived accuracy of decisions. 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 (46.3%) 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. Column (2) restricts the sample to conviction decisions only. This matters because all-men teams convict less frequently, meaning a larger share of their decisions are acquittals – which are rarely appealed. This mechanically suppresses their overall appeal rate, so the gap estimated in column (1) likely understates the true difference in perceived accuracy. Consistent with this, once we restrict to convictions the estimated gap grows to 19.1 percentage points (54.6%). The pattern for mixed-gender teams is also informative: despite convicting at similar rates to all-men teams, their convictions are appealed at rates statistically indistinguishable from those of all-women teams. This wedge between conviction rates and appeal rates is difficult to reconcile with teams differing only in leniency, and instead suggests that diagnostic skills shape the perceived quality of convictions independently of how often teams convict. 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 at the intensive margin. To address this, column (3) adds a control for the length of the prison sentence. The takeaway remains unchanged, suggesting that the higher likelihood of appeal for all-men teams is not driven by sentence severity per se. While these results already are at odds with a pure leniency interpretation, perceived accuracy alone is not sufficient to distinguish leniency from skill, as appeal decisions are made by the parties and may reflect strategic considerations in addition to the underlying accuracy of the ruling. We therefore turn next to appeal-proof decisions, which provide a closer proxy for the effective accuracy of first-instance rulings.

Appeal-proof decisions. 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 this measure of effective accuracy mirror those on perceived accuracy. Specifically, rulings made by all-men teams are 8.7 percentage points (nearly 10%) less likely to be appeal-proof compared to other teams (column (1)). In contrast, rulings by all-women and mixed-gender teams are equally likely to survive the appeal stage. The effects are even more pronounced when we focus on conviction decisions only (column (2)). All-men teams are 17.7 percentage points (24.8%) less likely to issue convictions that are appeal-proof than

all-women teams. 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 (20.9%) higher than those of all-men teams, while there is no statistically significant difference between teams with one versus two women. Teams with at least one woman issue conviction decisions which are almost as appeal-proof as those made by all-women teams (5-6.7pp lower), who display the highest level of accuracy among all team compositions. Finally, given that all-women teams impose shorter sentences when they convict (Table 5), and longer sentences are mechanically more likely to be reduced on appeal,[43] one might be concerned that their lower reversal rates could, in principle, reflect more cautious sentencing rather than greater accuracy: if all-women teams systematically impose lighter punishments, appellate courts have less room to reduce the sentence, mechanically raising the share of convictions that are confirmed. Column (3) addresses this concern directly by controlling for sentence length, thereby isolating the component of appeal-proof rates that reflects the accuracy of the conviction decision itself rather than its severity. The results are virtually unchanged, confirming that the higher appeal-proof rates of all-women teams are not driven by more lenient sentencing, but instead point to differences in the underlying accuracy of conviction decisions.

5.2.1 Interpreting the results:

Combining the results on conviction rates with those on appeal-proof decisions, two important implications emerge. First, the combination of higher conviction rates and higher accuracy for all-women teams is difficult to reconcile with a view in which panels share the same assessment of cases but apply different conviction cutoffs – that is, where judicial panels' gender composition affects leniency but not diagnostic ability. Under this mechanism, panels that convict more would do so by bringing in weaker marginal cases, implying that their convictions should be less likely to be upheld on appeal. We instead find the opposite: all-women teams convict more often and their convictions are more likely to be upheld, even conditional on sentence length. This pattern points to differences in diagnostic skill – i.e. a team's ability to distinguish guilt from innocence – across panels with different gender compositions.[44]

Second, the patterns for mixed-gender teams point to a wedge between their propensity to convict and the accuracy of convictions. Specifically, mixed-gender teams convict less

[43] When only the defendant appeals, the appellate court cannot increase the term.

[44] An alternative interpretation is that all-men teams write decisions with weaker or less persuasive motivations, while being equally accurate in assessing guilt. This channel could increase the propensity to appeal, because parties may be more likely to challenge decisions that are poorly reasoned. However, it is less consistent with a higher probability of reversal. In the Italian system, the criminal appeal is a review on the merits: the Court of Appeal reassesses the factual and evidentiary basis of the decision, within the issues raised by the appellant, rather than limiting review to the quality of the written motivation. If all-men teams differed mainly in drafting quality, we would expect higher appeal rates but also a higher confirmation rate conditional on appeal. Instead, we find that convictions by all-men teams are more likely to be appealed but not more likely to be confirmed, which points to differences in the accuracy of the underlying assessment rather than solely differences in written justification.

often than all-women teams, yet their convictions are almost as likely to be appeal-proof. One interpretation is that gender diversity raises coordination costs at the conviction stage, making it harder to translate similar evidence into a unanimous verdict, but without affecting the accuracy of convictions once such a verdict is reached. At the same time, the fact that conviction rates do not increase monotonically with the number of women on the panel, whereas conviction quality improves sharply once at least one woman is present, suggests that the relationship between gender composition and decision-making is not purely proportional. We return to these mechanisms in greater detail in Section 6, where we develop a theoretical framework to rationalize these patterns.

5.3 Robustness checks and heterogeneity

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

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 traits linked to both gender and case outcomes, the coefficients may partially capture the influence of those confounding factors. We address the concern that our results could reflect other forms of diversity, beyond gender, in three ways. First, since male and female judges differ along other demographic characteristics (see Section 3.1) all our baseline specifications include controls for judges' age, experience, and region of origin, ensuring that the estimated effects of gender composition are not driven by these other characteristics. Second, because gender correlates with the characteristics mentioned above, we investigate whether team diversity along age, regional background, and experience also affects conviction rates and appeal-proof conviction rates. Results are reported in Table A7. 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. Finally, 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. If the effects observed in our main analysis were driven by these correlated 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, or region of birth – do not play a significant role in our context. Finally, to further reassure that our results are not driven by judges' unobserved characteristics, we estimate the following

specification with judge fixed effects 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 A9 reports the results, which are consistent with our main findings. Because this specification includes judge fixed effects, it holds constant all time-invariant characteristics of individual judges and identification comes from within-judge variation as the same judge appears on teams with different gender mixes. The persistence of our results under this specification suggests that case outcomes associated with a given judge vary systematically with team gender composition, consistent with team-driven dynamics rather than fixed individual characteristics alone.

Court-specific time trends: Our baseline specification includes court fixed effects and crime-type-by-year fixed effects, which together absorb time-invariant differences across courts and crime-specific trends common to all courts. A remaining concern is that courts may also face idiosyncratic, time-varying shocks – such as changes in local judicial culture or shifts in caseload composition – that evolve differently across courts. If these shocks are correlated with the gender composition of judicial teams, our baseline estimates could be biased. To address this concern, we re-estimate our baseline model replacing court fixed effects with court-by-year fixed effects, thereby absorbing any court-specific shock in a given year. This more restrictive specification compares teams of different gender compositions within the same court and year, controlling for crime-type-specific trends over time. Appendix Table A8 reports the results and shows that the estimates are virtually unchanged, with coefficients that are stable in sign, magnitude, and statistical significance.

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 A10 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 knowledge 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. For example, 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 hypothesis in Appendix

Table A11, 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).[45] 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: An alternative explanation is that our results are partly driven by homophily between judges and defendants of the same gender or nationality. 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 A12) and appeal-proof conviction rates (Panel B of Appendix Table A12), 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 consistent, suggesting that the effects of team gender composition are not driven by specific interactions with defendants of a particular gender or nationality.

6 Mechanisms: Effort, Skill, and Team Dynamics

Our results point to systematic differences in the accuracy of conviction decisions across teams with different gender compositions. A natural question is whether these differences reflect (i) fixed differences in diagnostic ability (innate talent), or (ii) differences in the amount of effort judges exert in gathering and processing information, deliberating, and scrutinizing the evidentiary record. Distinguishing between these channels is important because they have different implications for policy and for the interpretation of gender composition effects. If the accuracy advantage reflects fixed diagnostic ability, it suggests that who sits on the panel matters primarily through stable differences in judgment and information processing. If instead it reflects effort responses, it implies that the composition of a team can influence the type and quality of decision through endogenous effort responses, pointing to potentially malleable margins – such as workload, deliberation time, or incentives – through which institutions could improve decision quality. We begin by presenting empirical evidence that points to an effort-based channel rather than an innate talent-based one. We then propose and estimate a model of team behavior that can rationalize our findings. Standard judicial decision-making models that allow judges to differ only in their conviction thresholds but not in skill or effort (e.g., Maestas et al., 2013; Galasso and Schankerman, 2015; Dobbie et al., 2018; Bernstein et al., 2019) cannot reproduce the key results of our analysis: that teams who convict more frequently also make fewer errors conditional on

[45] Results are similar when we estimate separate regressions by crime type. We report the leave-one-out specification as it is more precisely estimated.

conviction. By contrast, we show that a framework incorporating two departures from this benchmark – heterogeneity in diagnostic skills, modeled via endogenous effort provision, and within-team frictions arising from gender diversity – can generate the full pattern of results.

6.1 Is diagnostic ability fixed or effort-driven?

In this subsection, we ask whether the gender differences in diagnostic performance documented above are better explained by differences in effort – how intensively judges gather and process information and deliberate – rather than by fixed differences in innate ability alone. Evidence from several real-world settings, including medicine, academia, and corporate boards, suggests that women tend to work more diligently and exert greater effort (Coviello et al., 2025; Chu et al., 2022; Hengel, 2022; Gupta et al., 2019; Adams and Ferreira, 2009). The mechanisms behind these patterns are likely multifaceted and may include differences in intrinsic motivation at work, differences in risk aversion, and asymmetries in workplace environments—most notably that women often face greater scrutiny and harsher penalties for comparable mistakes, while receiving less support and credit for similar performance (Borghans et al., 2009; Gagliarducci and Paserman, 2012; Benson et al., 2021; Egan et al., 2022; Burbano et al., 2024; Sarsons, 2024; Dupas et al., 2025). Yet, fixed differences in ability could also generate higher accuracy, and ability may interact with effort by lowering its cost or increasing its returns. Although our data do not allow us to cleanly separate differences in effort from fixed differences in ability, their richness helps us assess which of the two mechanisms is more consistent with the empirical patterns we document. Overall, the patterns we document below are more consistent with differences in effort than with fixed differences in innate ability alone.

Entry examination results. First, we examine whether our findings could be driven by gender differences in selection into the judiciary. Given that the profession was historically male-dominated, women may have faced higher entry barriers and thus been more positively selected on average, raising the possibility that pre-existing differences in baseline ability account for our results. To investigate this, we collected and digitized the official rankings from all national entry examinations for the judiciary held between 1965 and 2022. If women were indeed more positively selected than men, we would expect them to perform better on the entry exam. Figure 1 shows that any such difference is negligible in practice. Panel (a) displays the grade distributions for men and women, which are nearly identical in shape, with women scoring on average only 2 points higher than men. Panel (b) plots average grades by exam year and reveals a general upward trend over time that is common to both men and women, with confidence intervals overlapping throughout the entire period, suggesting no meaningful gender gap in exam performance over time. While it remains possible that women are more positively selected on traits not captured by the exam, this evidence shows that they are not differentially selected on the dimensions the judiciary itself considers most important at entry.

Duration of the decision-making process. We next investigate whether the gender

composition of judicial teams affects the time required to reach a decision. Longer decision times may reflect more intensive information processing and deliberation, but they may also arise from within-team frictions, such as more protracted discussions or difficulty 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.[46] Table 8 reports the results. All-men teams complete cases substantially faster than any other composition: their trials last, on average, eight months (30%) less than those of all-women teams and nearly six months (22%) 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 are consistent with two complementary interpretations. First, the fact that all-women teams spend more time on cases than all-men teams is consistent with greater time devoted to case processing and deliberation. If differences in decision quality reflected only fixed differences in innate ability, one might expect all-women teams to achieve better outcomes without requiring systematically more time per case. Instead, we find that all-women teams both achieve better outcomes and devote more time to each case, a pattern that is more naturally aligned with higher effort than with higher innate ability alone. Second, the finding that all-women and mixed-gender teams take equally long to reach decisions – despite showing different conviction rates – is consistent with 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 might be absorbed by difficulties in reaching consensus. At the same time, the fact that they have higher appeal-proof rates than all-men teams indicates that when consensus is reached, the quality of their convictions nearly matches that of all-women teams. We return to these mechanisms below, and formalize them in the model.

Written opinions: To better understand the role of differences in effort in driving our findings, we examine the written reasoning that judicial teams are required to provide for every verdict.[47] 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).[48] 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.[49] Finally, our third measure of drafting quality captures the frequency of citation

[46]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.

[47]These texts differ from U.S. judicial opinions in key ways. See Appendix B for more details.

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

[49]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

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).[50] Results are reported in Table 9.[51] All-men 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 producing clear and concise judicial reasoning.

6.2 A model of team diagnostic decision-making

The empirical evidence presented above is more consistent with effort differences than with fixed differences in innate ability. We now propose and structurally estimate a model of team decision-making that formalizes this interpretation. Our framework builds on existing single-agent models of diagnostic decision-making that incorporate heterogeneity in skills and/or effort (e.g. Mungan, 2024a,b; Chan et al., 2022), and extends them to a team setting with strategic interactions modeled through endogenous effort provision and within-team frictions. To the best of our knowledge, we are the first to generalize this approach to teams. We show that this framework can rationalize the full pattern of results: the coexistence of higher conviction rates and higher conviction quality for all-women teams, and the wedge between conviction rates and conviction quality for mixed-gender teams. We then use the estimated model to decompose the sources of cross-team performance gaps and to evaluate how the growing presence of women in the judiciary is likely to affect aggregate court performance.

6.2.1 Model

We conceptualize judicial decision-making in two stages. First, an *evaluation* stage in which judges assess the evidence and form a belief about the defendant's guilt. The accuracy of this assessment depends on the team's *diagnostic skills* – namely, its ability to distinguish guilty from innocent defendants – which we model as being shaped by judges' *chosen effort*.[52]

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.

[50] 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.

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

[52] Endogenizing effort allows behavior to adjust to who sits on the panel, capturing strategic responses to teammates. Explaining the full pattern of results with innate ability alone would require strong assumptions about how individual abilities aggregate into team outcomes. Under most natural aggregation rules, mixed-

Second, a *decision* stage in which, given their assessment, judges decide whether to convict weighing the costs of wrongful conviction against those of wrongful acquittal. These weights can be interpreted as judges' leniency. In other words, during stage one judges order cases by the likelihood of guilt of the defendant, and in stage two they select a conviction cutoff along that ordering.[53]

States and prior. Consider a team of two judges $i \in \{1, 2\}$ making a binary decision under uncertainty.[54] 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 true state of guilt S of the defendant cannot be observed by the judges.

Signal technology and cost of effort. The panel's information about the defendant comes from judicial effort. Judges $i \in \{1, 2\}$ simultaneously exert effort $e_i > 0$ at a convex cost $c(e_i)$ (with $c'(e) > 0$ and $c''(e) > 0$) to analyze and interpret the evidence of the case. Their joint effort determines the quality of the information environment, which we capture through a common precision $\theta(e_i, e_j) \in \left(\frac{1}{2}, 1\right)$, increasing and concave in each judge's effort and exhibits decreasing returns across judges ($\theta_{e_i e_j} < 0$).[55] Based on this shared evidentiary record, each judge independently forms an individual assessment of the defendant's guilt, modeled as a binary signal $X_i \in \{G, I\}$ which is correct with probability θ:

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

That is, conditional on the true state, each judge's signal is drawn independently.[56] Two judges examining the same body of evidence may therefore reach different assessments of guilt even though both signals are informed by the same underlying precision. The probability of such a disagreement is $2\theta(1 - \theta)$.[57]

Disutility from mistakes. Judicial teams can make mistakes. Let $d \in \{g, i\}$ denote the

gender teams would be expected to fall between all-men and all-women teams on every outcome. Yet we observe that mixed-gender teams resemble all-men teams in conviction rates but all-women teams in appeal-proof rates, a pattern that would require different aggregation rules for different outcomes. An effort-based model captures this asymmetry more parsimoniously, as strategic responses to teammates naturally generate outcome-specific composition effects without additional structure.

[53] If judges share the same diagnostic skill, they rank cases identically and differ only in conviction thresholds; then any case convicted by the higher-threshold (more lenient) judge must also be convicted by the lower-threshold judge, satisfying the judge-IV monotonicity assumption. If diagnostic skills differ across judges, rankings can cross, leading to violations of monotonicity: there can be cases that a more accurate, higher-threshold judge convicts while a less accurate, lower-threshold judge acquits.

[54] Note that, for simplicity, we model teams as consisting of two judges rather than three. This simplification is justified because (i) in practice convictions require unanimity, and (ii) our outcomes of interest are not statistically different between teams with two men and one woman and teams with two women and one man, so conceptually we can pool mixed-gender teams together.

[55] The assumption that effort is a public good – both judges' efforts contribute to a shared θ – reflects the institutional feature that panel members share the case file and access the same evidence.

[56] We assume a symmetric signal for tractability; allowing asymmetric error rates would not change the logic as long as higher effort makes the signal better at separating guilty from innocent cases.

[57] This formulation captures a key feature of judicial panels: although both judges study the same case file (so total effort determines the overall information quality), they may weigh evidence differently, focus on different aspects of the case, or interpret witness credibility differently, leading to genuinely different assessments.

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 $\phi_i\beta_i$; a wrong acquittal generates a disutility of ϕ_i; correct decisions yield no disutility:[58]

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

The parameter β_i measures how much judge i penalizes a wrongful conviction relative to a wrongful acquittal; a higher β_i implies greater *leniency* (a higher conviction threshold). Since the law regards convicting an innocent person as more costly than acquitting a guilty one (*Blackstone Principle*), we assume $\beta > 1$.[59] The parameter ϕ_i instead captures judge i's aversion to *any* errors. We interpret ϕ_i as a judge's *diligence* parameter, which increases when mistakes are perceived as more costly (for instance due to personal ethics, professional norms, external scrutiny, or potential sanctions).

Therefore, judges can differ along two dimensions: (i) how they trade off wrongful convictions and wrongful acquittals (their *leniency*, captured by β_i), and (ii) their aversion to making mistakes of either type (their *diligence*, captured by ϕ_i). Both parameters shape the judge's chosen *effort* e_i, which raises the precision of the team's information and thus its ability to distinguish guilt from innocence. Consequently, cross-team differences in diagnostic skills arise from different effort choices.[60]

Stage 2: Voting and deliberation. After observing their individual signals, each judge forms a posterior belief about the defendant's guilt and votes accordingly. Given a guilty signal and precision θ, judge i's posterior is

$$\Pr(G \mid X = G) = \frac{p\,\theta}{p\,\theta + (1-p)(1-\theta)}$$

Judge i votes to convict if this posterior exceeds their personal conviction threshold $\beta_i/(1+\beta_i)$. Since conviction effectively requires unanimity, the binding constraint is the threshold of the most lenient member of the panel. The panel can convict only if signal precision

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

[59] 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 reflects caution against convicting an innocent, and implies that any reasonable doubt about a defendant's guilt should result in acquittal.

[60] By endogenizing effort and allowing heterogeneity in both ϕ and β, we depart from benchmark models that hold diagnostic skill fixed and vary only β across judges. A model where judges differ only in leniency (à la Mungan (2024a)) as driver of effort is not able to match the joint patterns in conviction rates and appeal-proof rates.

clears this team threshold (otherwise it acquits regardless of the signals):

$$\theta \geq \frac{\beta(1-p)}{p+\beta(1-p)} \equiv \bar{\tau} \tag{2}$$

where $\beta = \max(\beta_i, \beta_j)$. In this case, three outcomes are possible. First, if both signals indicate guilt ($X_i = X_j = G$), the panel convicts. Second, if both signals indicate innocence ($X_i = X_j = I$), the panel acquits. Third, if the signals conflict ($X_i \neq X_j$), the judges disagree and enter a deliberation process. With probability q, the judges reach an agreement: one of the two judges is selected at random and their assessment prevails. With probability $1-q$, they fail to reach agreement, and the panel acquits by default.[61] Under this rule, the probability of conviction given disagreement is $q/2$: the judges must first reach agreement (probability q) and then the judge whose assessment prevails must be the one who received signal G (probability 1/2).

Stage 1: Effort choice. Judges simultaneously choose costly effort e_i that increases the precision $\theta(e_i, e_j)$ of their signals, treating the deliberation outcome as exogenous. Since judges do not anticipate how disagreements will be resolved, each judge's objective depends only on the accuracy of their own signal. The probability that judge i's own signal is incorrect is $1-\theta$ for both types of error: a guilty signal when the defendant is innocent, and a not-guilty signal when the defendant is guilty. Each judge therefore maximizes their expected utility:

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

The first-order condition is

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

At the optimum, the marginal gain in signal precision from an additional unit of effort equals the marginal cost of that effort, scaled by the judge's expected cost of an incorrect signal. Jointly solving the two first-order conditions pins down equilibrium efforts $e_i^* = f(\phi_i, \phi_j, \beta_i, \beta_j, p)$.[62]

We obtain the following comparative statics.[63] Optimal effort e_i^* is increasing in judge i's own diligence ϕ_i and leniency β_i: a higher ϕ_i makes any error more costly, while a higher β_i makes wrongful convictions especially costly, so both prompt judge i to invest more in signal

[61] The default of acquittal when deliberation fails is consistent with the legal principle that the burden of proof lies with the prosecution. The symmetric selection upon agreement captures the idea that, when judges do engage constructively in deliberation, neither has a systematic advantage in persuasion.

[62] This first-order condition characterizes an interior solution in which the equilibrium precision exceeds the conviction threshold ($\theta > \bar{\tau}$). If instead the precision implied by judges' effort falls below $\bar{\tau}$, both judges acquit regardless of their signal realization, and the marginal return to effort is zero, yielding a corner solution at $e_i = 0$. In Section 6.2.2, we verify that at estimated parameters the equilibrium lies in the interior region for all team types.

[63] Proofs easily follow by differentiating equation (3) and applying the Implicit Function Theorem.

precision.[64] By contrast, e_i^* is decreasing in the teammate's diligence ϕ_j and leniency β_j: since efforts are strategic substitutes, a more diligent or more lenient partner – who works harder, and who, if setting the binding threshold, tightens the team's conviction standard – reduces the marginal payoff to i's additional effort, leading i to free-ride. Symmetrically, when paired with a less diligent or less lenient teammate, i compensates by exerting more effort. Finally, optimal effort is decreasing in the prior p when $\beta > 1$: a higher p makes the costlier error (convicting the innocent) rarer, so the return to extra effort falls. Importantly, the first-order condition does not depend on q: judges choose how carefully to study the case based on their own preferences over accuracy and leniency, without anticipating the specific bargaining outcome. This implies that effort differences across team types are driven entirely by the preferences parameters ϕ and β.

Conviction rates (C). Combining the signal technology and the deliberation rule, the unconditional conviction rate is:

$$C = p\left[\theta^2 + q\,\theta(1-\theta)\right] + (1-p)\left[(1-\theta)^2 + q\,\theta(1-\theta)\right]$$

The first bracket collects convictions when the defendant is guilty: either both judges receive signal G (which happens with probability θ^2), or the signals conflict and deliberation leads to conviction (which happens with probability $q\,\theta(1-\theta)$, since disagreement occurs with probability $2\theta(1-\theta)$ and each disagreement leads to conviction with probability $q/2$). The second bracket captures the analogous channel when the defendant is innocent. Since the disagreement term $q\,\theta(1-\theta)$, does not depend on the true state, the expression simplifies to:

$$C = \left[p\theta^2 + (1-p)(1-\theta)^2\right] + q\,\theta(1-\theta) \tag{4}$$

Higher signal precision θ increases conviction rates when the prior probability of guilt is sufficiently high. Intuitively, as θ rises, the increase in unanimous convictions from truly guilty defendants outweighs both the decrease from truly innocent defendants and the reduction in convictions obtained through deliberation (since disagreements become rarer). Formally, $\partial C/\partial\theta = 2(\theta + p - 1) + q(1-2\theta) > 2p - 1 > 0$ when $p > 1/2$. In our context – conditional on a case having passed charging and reached trial before a collegial panel – it is reasonable to expect $p > 1/2$, so that C is increasing in θ.

The effect of β on convictions works through two opposing channels. First, a higher β_i makes wrongful convictions more costly for i, so e_i^* rises; because efforts are strategic substitutes, e_j^* falls. The net effect on θ is positive if i's own response dominates the partner's free-riding, pushing C up. Second, if i is the more lenient judge so that $\beta = \beta_i$, an increase in β_i tightens the team's conviction threshold, which pushes C down. The net effect

[64] Note that variation in innate ability could generate the same comparative statics: a more talented judge either produces more precision per unit of effort or faces lower effort costs, which the model maps into higher effective effort in the same way as ϕ. Thus, our model captures effective effort: the observational effect of working harder (diligence-driven effort) and/or being more talented (innate ability). As discussed in Section 6.1, the empirical evidence presented above is more consistent with the diligence-driven interpretation than with fixed differences in innate ability.

on C is therefore ambiguous and depends on whether the increase in precision outweighs the tightening of the conviction standard.

Greater within-team frictions (lower q) reduce conviction rates. From equation (4), $\partial C/\partial q = \theta(1-\theta) > 0$. Because frictions arise at the deliberation stage – after effort is chosen – and are not anticipated by judges when choosing effort, they leave precision θ unchanged. Instead, they operate at the decision stage by reducing the probability that disagreement cases result in conviction.

Appeal-proof conviction rates (APC). The appeal-proof conviction rate is given by the share of convictions that are truly guilty.[65] That is,

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

The numerator reflects the two paths to a correct conviction: both judges independently receiving guilty signals from a truly guilty defendant (contributing $p\theta^2$), and one judge receiving a guilty signal while the other does not, with the team reaching agreement and the judge favoring conviction being selected (contributing $pq\theta(1-\theta)$).

The effects of θ and p on APC are positive and intuitive. Higher precision θ raises APC by making guilty signals more likely to be true positives rather than false positives, so the share of convictions that are truly guilty rises. A higher prior p raises the chance that a case is actually guilty before observing the signals, so – conditional on conviction – the probability of true guilt increases.

The role of q is less obvious. Since the probability of disagreement $2\theta(1-\theta)$ does not depend on the defendant's true guilt status, cases reaching deliberation have composition $\Pr(S = G \mid \text{disagree}) = p$ – exactly the prior. These cases are therefore of lower average quality than unanimous convictions, among which the share of truly guilty defendants is $\theta^2/[\theta^2 + (1-\theta)^2] > p$. Friction acts as a filter: when agreement is hard to reach, most disagreements end in acquittal by default, so only high-quality unanimous convictions survive into the conviction pool. As q rises, more of these lower-quality disagreement cases result in convictions, dragging APC toward p. Formally, $\partial APC/\partial q = \theta(1-\theta)(p - APC)/C < 0$.

Qualitative predictions. We now summarize how the forces identified above – diligence, leniency, and frictions – map into cross-team comparisons of conviction rates and conviction quality.

(i) Diligence: teams with more diligent members (higher ϕ) exert more effort and thus achieve higher precision θ, leading to higher C and APC than less diligent teams. When comparing two teams, if one team has a judge with higher ϕ and the other with lower ϕ than

[65] For expositional simplicity in the model, we treat every decision as ultimately revealed to be either "correct" or "incorrect": this corresponds to a scenario where all decisions are appealed, and the appellate court observes the defendant's true state - so an upheld conviction implies guilt and a reversal implies innocence. In practice, not all decisions are appealed, and an upheld decision might not coincide with the truth. As discussed in Section 4, our empirical analysis relies on the weaker assumption that the appellate court is unbiased and (weakly) more accurate than the trial panel, so that decisions that are reversed on appeal are, on average, less likely to be correct than those that are not. Allowing for a small, effort-invariant probability of appellate error does not alter the model's predictions.

the comparison team, the sign of the difference in C and APC depends on the net change in precision: because efforts are strategic substitutes, the low-ϕ judge tends to free-ride while the high-ϕ judge works more; C and APC are higher only if the latter's increase in effort dominates, and lower otherwise.

(ii) Leniency: teams with more lenient members (higher β) face two opposing forces: a *threshold* effect that lowers C, and a *precision/effort* effect that raises C. Hence the net effect on C is ambiguous compared to a less lenient team, and will depend on the values of parameters. Since APC depends on θ but not directly on β, a more lenient team will have higher APC than a less lenient team only if its greater leniency translates into higher total precision – that is, only if the additional effort induced by higher β more than offsets any free-riding by the teammate. If, relative to another team, one judge's β is higher and the other's lower, whether APC is higher likewise hinges on the net change in θ.

(iii) Frictions: Teams that experience larger frictions (lower q) have lower C but higher APC. The effect on C is direct: lower q reduces the probability that disagreement cases result in conviction. The effect on APC works through a composition channel: because disagreement cases are of prior quality p, which is below the quality of unanimous convictions, admitting fewer of them into the conviction pool raises the average quality of convictions. However, because q does not affect effort or precision, differences in q across teams produce larger variation in conviction rates than in conviction quality.[66]

In Section 6.2.2, we estimate gender-specific values of ϕ, β, and q and show that the model can match the reduced-form patterns documented in Section 5.

6.2.2 Estimation

For any given parameter vector, the model is solved by backward induction. We adopt the following functional forms: the signal precision is $\theta(e_i, e_j) = \exp(\frac{e_i+e_j}{2})/[1+\exp(\frac{e_i+e_j}{2})]$ and the effort cost function is $c(e_i) = e_i^2/2$.

We assume that agents can be of two types – men and women – and we allow both ϕ_i and β_i to differ by type. Specifically, we denote by ϕ_m and β_m the parameters for men, and by ϕ_f and β_f those for women. We also let the agreement probability q differ between same-gender teams (q_0) and mixed-gender teams (q_{mix}). The resulting parameter vector is $\Theta = (\beta_m, \beta_f, \phi_m, \phi_f, p, q_0, q_{mix})$. We estimate Θ to match the following six empirical moments:

$$\hat{m} = [\hat{C}_{\text{AW}}, \hat{C}_{\text{AM}}, \hat{C}_{\text{MG}}, \hat{APC}_{\text{AW}}, \hat{APC}_{\text{AM}}, \hat{APC}_{\text{MG}}],$$

where $\hat{C}_t$ and $\hat{APC}_t$ denote, respectively, the estimated conviction rate and the appeal-proof conviction rate of team type $t \in \{\text{AW}, \text{AM}, \text{MG}\}$ (all-women, all-men, mixed-gender).[67]

Since we have six empirical moments and seven structural parameters, we require one normalization. We set $\phi_m = 1$, so that ϕ_f is interpreted as women's diligence *relative* to

[66] From the derivatives, $|\partial APC/\partial q|/|\partial C/\partial q| = (APC - p)/C$. Since $APC - p$ is bounded above by $1-p$ while C is bounded below by $p\theta^2 + (1-p)(1-\theta)^2$ (the unanimous-conviction component), this ratio is well below one for reasonable parameter values.

[67] To mirror the model structure, the targeted moments are obtained from a regression that pools together

men's. This is a natural choice because ϕ enters the model as a scale parameter on the cost of errors: only the ratio ϕ_f/ϕ_m – not the individual levels – determines the difference in effort choices across types, making the normalization innocuous.[68] We then estimate the remaining six parameters via minimum distance, choosing the parameter vector that minimizes $Q(\Theta) = \big(m(\Theta) - \hat{m}\big)' W \big(m(\Theta) - \hat{m}\big)$, where $m(\Theta)$ denotes the vector of model-implied moments for a given Θ; $\hat{m}$ is the vector of targeted empirical moments; and W is a diagonal weighting matrix with entries given by the inverse of the variance of the corresponding empirical moments.[69]

Table 10 reports the results. Panel A displays the estimated parameters. Women are both (a) more diligent ($\phi_f > \phi_m$) and (b) more lenient ($\beta_f > \beta_m$), with the gender gap in diligence substantially larger than that in leniency: ϕ_f is estimated to be 6.2 times as large as ϕ_m, while β_f is estimated to be just 1.2 times β_m. Same-gender teams reach agreement more easily than mixed-gender teams ($q_{mix} < q_0$, with q_0 being approximately 1.3 times larger than q_{mix}, consistent with greater coordination frictions in gender-diverse panels. All parameters are precisely estimated. Panel B shows that the model fits the data well, with a mean absolute percentage error of 1.7% across the six targeted moments. Panel C reports the equilibrium effort levels implied by the estimates. As predicted by the comparative statics, women exert substantially more effort than men in every team configuration. In mixed-gender panels, men free-ride relative to all-men teams, while women over-compensate relative to all-women teams, so that mixed-gender teams' total effort falls between that of all-men teams and all-women teams, but is much closer to the latter. These estimated parameters generate the cross-team patterns documented in the reduced-form analysis. Women's substantially higher diligence leads to higher precision θ, which raises both C and APC in all-women teams relative to all-men teams. In mixed-gender panels, women's over-compensation largely offsets men's free-riding, so total precision – and thus APC – nearly matches that of all-women teams. However, their lower q makes it harder to convert disagreements into convictions, pulling C close to that of all-men teams. Since q operates only at the decision stage and leaves precision unchanged, lower agreement in mixed-gender teams reduces how often they convict but not the quality of convictions when they do – rationalizing the wedge between conviction rates and conviction quality that characterizes mixed-gender teams.

---

teams with one and two female judges:

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

where Y_{isct} is either an indicator for conviction or for appeal-proof conviction. This specification is consistent with our reduced-form evidence, which shows that for our outcomes of interest the estimated coefficients for the two types of mixed-gender teams are similar in magnitude and not statistically different from each other. Results from the pooled regression are reported in Appendix Table A13. The findings and their interpretation are qualitatively unchanged relative to those in Tables 5 and 7.

[68] Results are virtually unchanged under alternative normalizations, such as fixing $\beta_m = 1$, fixing p at 0.5 or 0.55 or fixing $q_0 = 1$, and estimating both ϕ_m and ϕ_f freely.

[69] Numerically, we implement a global search method on a Sobol quasi-random low-discrepancy grid of candidate parameter vectors, paired with a local search algorithm around the three best candidates for a global minimum.

6.2.3 Counterfactual Analysis

The estimation results show that women are both more diligent and lenient than men, and that mixed-gender teams face greater coordination frictions. A natural question is how much each of these channels contributes to the observed performance gaps, and – in light of this – what would happen to aggregate court outcomes as the share of women in the judiciary continues to grow. We address both questions through two sets of counterfactual exercises, reported in Table 11. First, we change one structural parameter at a time, holding the others fixed and re-solving for equilibrium effort, to identify the main drivers of the gap in team performance (Panel A). Second, we vary the share of women in the judiciary and re-computes aggregate outcomes, to shed light on the implications of different gender compositions for aggregate court performance (Panel B). For each counterfactual scenario, we estimate conviction rates, appeal proof conviction rates and total proceeding time – defined as the number of days from between the assignment of a case to a team and its final resolution, where final resolution is the first-instance verdict for cases that are not appealed and the appeal ruling for cases that are. The baseline outcomes are reported in Panel C for comparison.

Parameter counterfactuals. Panel A of Table 11 reports the results of the first set of exercises. In each scenario, one parameter is changed while all others remain at their estimated values, and equilibrium effort is re-solved.

Equal diligence ($\phi_m = \phi_f$). Raising men's diligence to women's level has the largest effect on conviction quality. The aggregate APC rate increases by 5.3 percentage points, driven mostly by all-men teams, whose conviction accuracy rises by 14.1 percentage points. Notably, the appeal-proof conviction rate of mixed-gender teams – which increases by 5.9 percentage points – slightly exceeds that of all-women teams under equal diligence, because greater coordination frictions in mixed panels filter out lower-quality disagreement cases from the conviction pool. Conviction rates also rise modestly (+1.3 pp) as a consequence of higher effort. This exercise highlights the central role of the diligence gap: if men were as careful as women, the quality gap between team types would largely disappear. However, this accuracy gain comes at the cost of longer trials: predicted trial length increases by 80 days.

Equal leniency ($\beta_m = \beta_f$). Setting men's leniency to women's level – making men more averse to wrongful acquittals – lowers the aggregate APC by 3.3 percentage points. The intuition follows from the first-order condition (equation (3)): a higher β increases the cost of wrongful acquittals, which encourages effort but also raises the conviction threshold. With men's diligence low, the additional effort is too small to meet this higher threshold: at the estimated parameters, precision remains below the conviction threshold and, as a result, all-men teams stop convicting entirely. This finding reinforces the conclusion that the gap in team performance is primarily driven by differences in effort rather than in preferences for leniency alone.

No mixed-gender frictions ($q_{mix} = q_0$). Removing coordination frictions in mixed-gender

teams raises the aggregate conviction rate by 3.5 percentage points, with the aggregate APC falling slightly (-1.1 pp). As discussed in Section 6.2.1, frictions act as a filter: by making it harder to convert disagreement cases into convictions, lower q keeps lower-quality cases out of the conviction pool, raising the average accuracy of convictions that are made. Eliminating frictions therefore increases convictions but slightly reduces their average quality. The effect is smaller than in the diligence counterfactual, indicating that frictions play only a limited role in explaining cross-team performance differences. Finally, since frictions operate only at the deliberation stage and do not affect effort, trial length is unchanged.

Composition counterfactuals. Panel B of Table 11 examines how changes in the share of women affect aggregate outcomes, holding estimated parameters fixed.[70] We vary the share of women in the pool from 0% to 100% and compute the implied team composition of three-judge teams. This exercise is motivated by the ongoing feminization of the Italian judiciary documented in Appendix Figure A1: women now represent the majority of new entrants into the profession, a trend that is progressively shifting the gender composition of collegial panels.

The results show a clear tradeoff between decision quality and trial duration. As the share of women rises, the appeal-proof conviction rate increases monotonically. Relative to the current composition (56% women), moving to 75% women raises aggregate APC by 1.8 percentage points, while an all-women judiciary raises it by 5 percentage points. Conviction rates also increase by 1.3 and 4.7 percentage points, respectively. These gains reflect the higher diligence of women, which translates into more precise signals and thus more and more accurate convictions. However, more diligent teams also take longer to process cases. At 75% women, aggregate trial length increases by 24 days; at 100%, the increase is 65 days relative to the current composition. Reducing the share of women has the opposite effect. When 25% of judges are women, aggregate APC falls by 5 percentage points; with no women, it falls by 11.7 percentage points. At the same time, these scenarios yield substantially shorter trials: 77 fewer days at 25% women, and 183 fewer days at 0%.[71]

Taken together, these exercises suggest that as the judiciary becomes increasingly female-dominated, aggregate decision accuracy should improve, accompanied by an increase in trial duration. This tradeoff is inherent to the diligence channel: more careful case preparation produces better outcomes but requires more time. Whether the quality gains justify the additional time is ultimately a policy question, and one that will become increasingly relevant as the demographic trends documented in Appendix Figure A1 continue to reshape the composition of the bench.

[70] This exercise abstracts from general equilibrium effects that a large change in the gender composition of the judiciary might induce, such as selection into the profession.

[71] Note that the aggregate conviction rate slightly rises in the 0% women scenario (+1.9 pp) because the large mass of low-C mixed teams (75.7% of panels) is replaced by all-men teams, whose conviction rate is slightly higher, more than offsetting the disappearance of the highest-C all-women teams.

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 either not appealed or appealed and confirmed.

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-women 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-men teams convict less frequently and have the lowest rate of appeal-proof convictions. Mixed-gender teams resemble all-men teams in conviction rates but achieve conviction quality nearly as high as all-women teams.

A diagnostic-skill framework with endogenous effort helps interpret these patterns. The combination of higher conviction rates and fewer reversals for all-women teams is inconsistent with our results being driven exclusively by gender differences in leniency and points to the presence of gender differences in diligence. This could, in principle, reflect gender differences in either innate talent or effort, and we provide a series of tests that are more consistent with women exerting more effort. Because each judge's optimal effort decreases with teammates' diligence and leniency, men in mixed-gender teams free-ride (exerting less effort than in all-men teams) while women over-compensate. This mechanism helps us rationalize the fact that mixed-gender teams reach conviction quality above that of all-men teams and similar to all-women teams. At the same time, mixed teams convict less often than all-women teams, consistent with coordination frictions: for similar levels of total effort, reaching consensus to convict is harder when views must be reconciled across a more diverse panel.

These findings have important implications. First, higher-quality decisions come with time costs: the teams whose convictions are most likely to be appeal-proof are also slower. Second, incentives inside the team matter: when members attach different personal costs to making mistakes, they contribute effort unevenly. Third, there is no single optimal team design: placing more diligent members – or ensuring at least one highly diligent member – can raise decision quality, but at the expense of speed and with the extra work concentrating on some teammates. Finally, calls for greater diversity should also account for potential coordination frictions that can make consensus harder even when effort is high.

References

Abrams, D., Galbiati, R., Henry, E., and Philippe, A. (2022). When in rome... on local norms and sentencing decisions. Journal of the European Economic Association, 20(2):700–738.

Adams, R. B. and Ferreira, D. (2009). Women in the boardroom and their impact on governance and performance. Journal of financial economics, 94(2):291–309.

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.

Aparicio Fenoll, A. and Zaccagni, S. (2022). Gender mix and team performance: Differences between exogenously and endogenously formed teams. Labour Economics, 79:102269.

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.

Ash, E., Chen, D. L., and Ornaghi, A. (2022). Gender attitudes in the judiciary: Evidence from us circuit courts. American Economic Journal: Applied Economics, *conditionally accepted*.

Ash, E., Chen, D. L., and Ornaghi, A. (2024). Gender attitudes in the judiciary: Evidence from us circuit courts. American Economic Journal: Applied Economics, 16(1):314–350.

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.

Benson, A., Li, D., and Shue, K. (2021). Potential" and the gender promotion gap. Unpublished working paper.

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.

Borghans, L., Heckman, J. J., Golsteyn, B. H., and Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. Journal of the European Economic Association, 7(2-3):649–658.

Boring, A. (2017). Gender biases in student evaluations of teaching. Journal of public economics, 145:27–41.

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.

Burbano, V., Padilla, N., and Meier, S. (2024). Gender differences in preferences for meaning at work. American Economic Journal: Economic Policy, 16(3):61–94.

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

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are referees and editors in economics gender neutral? The Quarterly Journal of Economics, 135(1):269–327.

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.

Chu, B., Handel, B., Kolstad, J., and Malmendier, U. (2022). Gender differences in non-promotable tasks: The case of clinical note-taking. Technical report, Tech. rep., University of California, Berkeley.

Coviello, D., Deserranno, E., and Persico, N. (2025). Gender disparities in the welfare effect of the minimum wage.

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.

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.

Dupas, P., Handlan, A., Modestino, A. S., Niederle, M., Seré, M., Sheng, H., Wolfers, J., Collective, S. D., et al. (2025). Gender differences in economics seminars. American Economic Review, forthcoming.

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.

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.

Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. American economic review, 90(4):715–741.

Gupta, K., Murray, S. G., Sarkar, U., Mourad, M., and Adler-Milstein, J. (2019). Differences in ambulatory ehr use patterns for male vs. female physicians. NEJM Catalyst, 5(6).

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

Hengel, E. (2022). Publishing while female: Are women held to higher standards? evidence from peer review. The Economic Journal, 132(648):2951–2991.

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

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

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.

Kim, D. and Starks, L. T. (2016). Gender diversity on corporate boards: Do women contribute unique skills? American Economic Review, 106(5):267–271.

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.

La Forgia, A. and Singh, M. (2025). Gender mix and team performance: Evidence from obstetrics. Working Paper.

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.

Motivans, M. A. (2025). Federal justice statistics, 2023. Technical Report NCJ 309946, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

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 influence. Working paper.

Rebolledo, E. B., Bruhn, J., Choon, T. H., and Weber, E. A. (2024). Gender composition and group behavior: Evidence from U.S. city councils. NBER Working Paper No. 33223.

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

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

Truffa, F. and Wong, A. (2025). Undergraduate gender diversity and the direction of scientific research. American Economic Review, 115(7):2414–2448.

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 in mixed- vs same-gender teams		Men in mixed- vs same-gender teams	
	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)
Exam grade	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 jail time (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 jail time
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

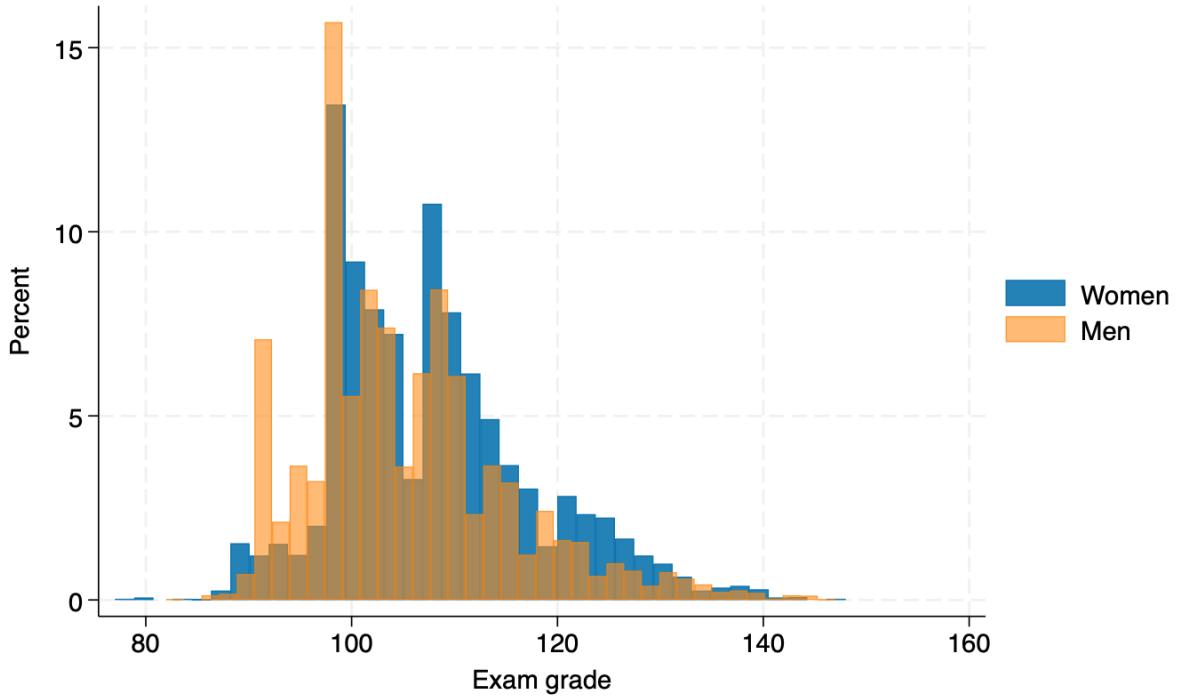
Notes. In all columns, the dependent variable is an indicator for whether the team's decision is appealed against. In column (1), the analysis is done on the whole sample; in columns (2) and (3), the analysis is restricted to conviction decisions only. Column (3) also controls for the length of jail 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 7: Effect of gender composition of the team on appeal-proof decisions

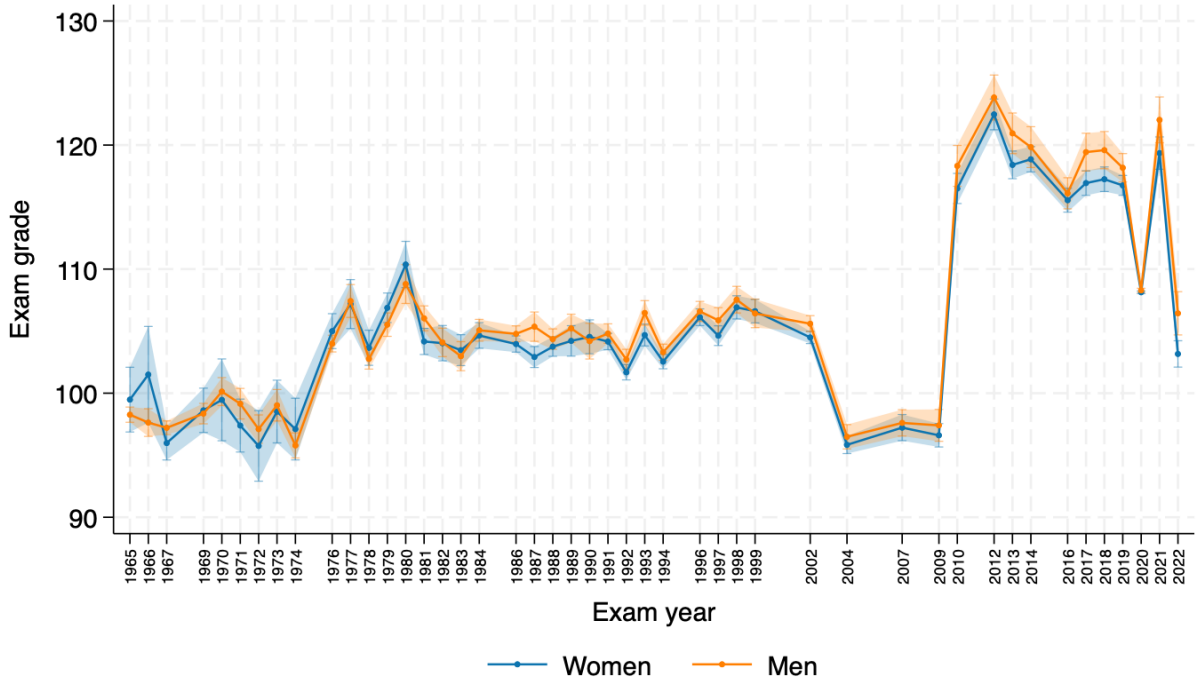
	(1) Unconditional	(2) Conditional	(3) Controlling for jail time
All-men team	-0.087**	-0.177***	-0.164***
	(0.034)	(0.051)	(0.050)
2M1F team	-0.009	-0.067**	-0.057**
	(0.014)	(0.026)	(0.025)
1M2F team	-0.014	-0.050**	-0.045*
	(0.013)	(0.025)	(0.024)
Observations	65051	27686	27686
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.867	0.714	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 or appealed and confirmed. 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 jail 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.

Figure 1: Grades differences between men and women



(a) Distribution of grades by gender



(b) Grades over exam years 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).

Table 8: Effect of gender composition of the team on decision-making time

	(1) Length of trial (days)
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 9: 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 column (1) is length (in words) of the written opinion; in column (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.

Table 10: Structural parameter estimate and model fit

Panel A: Structural Parameter Estimates			
Parameter		Estimate	
β_m (Men's leniency)		1.026	
		(0.030)	
β_f (Women's leniency)		1.575	
		(0.142)	
ϕ_f (Women's diligence, rel. to men)		6.160	
		(0.440)	
p (Prior probability of guilt)		0.563	
		(0.010)	
q_0 (Agreement prob., same-gender teams)		0.760	
		(0.029)	
q_{mix} (Agreement prob., mixed-gender teams)		0.563	
		(0.023)	
Panel B: Model Fit			
Moment	Target	Simulated	Error
Conviction rates			
All-women teams	0.484	0.472	−0.012
All-men teams	0.418	0.441	0.023
Mixed-gender teams	0.412	0.414	0.002
Appeal-proof conviction rates			
All-women teams	0.767	0.765	−0.002
All-men teams	0.592	0.600	0.008
Mixed-gender teams	0.710	0.712	0.002
Mean absolute percentage error			1.71%
Panel C: Implied Equilibrium Effort			
Team type	Judge 1	Judge 2	Total
All-women teams (F-F)	0.818	0.818	1.637
All-men teams (M-M)	0.126	0.126	0.252
Mixed-gender teams (M-F)	0.119	0.903	1.022

Notes. The table reports results from the structural estimation described in Section 6.2.2. Panel A displays the estimated parameters, with bootstrap standard errors in parentheses (500 replications). Men's diligence ϕ_m is normalized to 1; all other parameters are estimated via minimum distance. Panel B compares the targeted empirical moments with those implied by the model at the estimated parameters. Panel C reports the equilibrium effort levels implied by the estimated parameters for each team type; effort levels are deterministic functions of the estimated parameters.

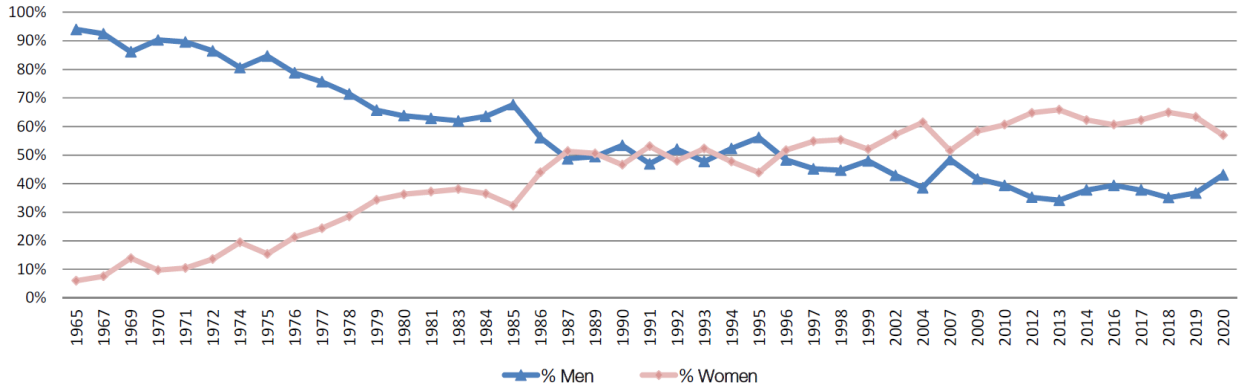
Table 11: Counterfactual Analysis

	(1) Conviction rate	(2) Appeal-proof conviction rate	(3) Total proceeding time (days)
Panel A: Parameter counterfactuals			
Equal diligence ($\phi_m = \phi_f$)			
Aggregate	0.439	0.768	1146
All-women teams	0.473	0.765	1131
All-men teams	0.468	0.738	1089
Mixed-gender teams	0.428	0.771	1153
Equal leniency ($\beta_m = \beta_f$)			
Aggregate	0.401	0.682	1070
All-women teams	0.473	0.765	1131
All-men teams	0.000	0.000	894
Mixed-gender teams	0.414	0.716	1068
No mixed frictions ($q_{mix} = q_0$)			
Aggregate	0.461	0.704	1066
All-women teams	0.473	0.765	1131
All-men teams	0.445	0.598	883
Mixed-gender teams	0.460	0.697	1064
Panel B: Composition counterfactuals			
0% women	0.445	0.598	883
25% women	0.428	0.665	989
75% women	0.439	0.733	1090
100% women	0.473	0.765	1131
Panel C: Baseline			
Aggregate	0.426	0.715	1066
All-women teams	0.473	0.765	1131
All-men teams	0.445	0.598	883
Mixed-gender teams	0.413	0.712	1064

Notes. The table reports the results of two counterfactual exercises. The outcome in column (1) is conviction rate; in column (2) the appeal-proof conviction rate, and in column (3) the total proceeding time, defined as the number of days between the assignment of a case to a team and the first-instance verdict for cases that are not appealed and the appeal ruling for cases that are. Panel A modifies one parameter at a time while holding all others at estimated values, with team composition held at current shares (18.4% AW, 5.9% AM, 75.7% MG). Panel B varies the share of women in the pool, and recomputes team shares combinatorially for 3-judge panels. Panel C reports the baseline outcomes for a comparison. Baseline team-level lengths are regression-adjusted data means; counterfactual lengths are obtained by adding the estimated marginal effect of effort on trial duration to the data means.

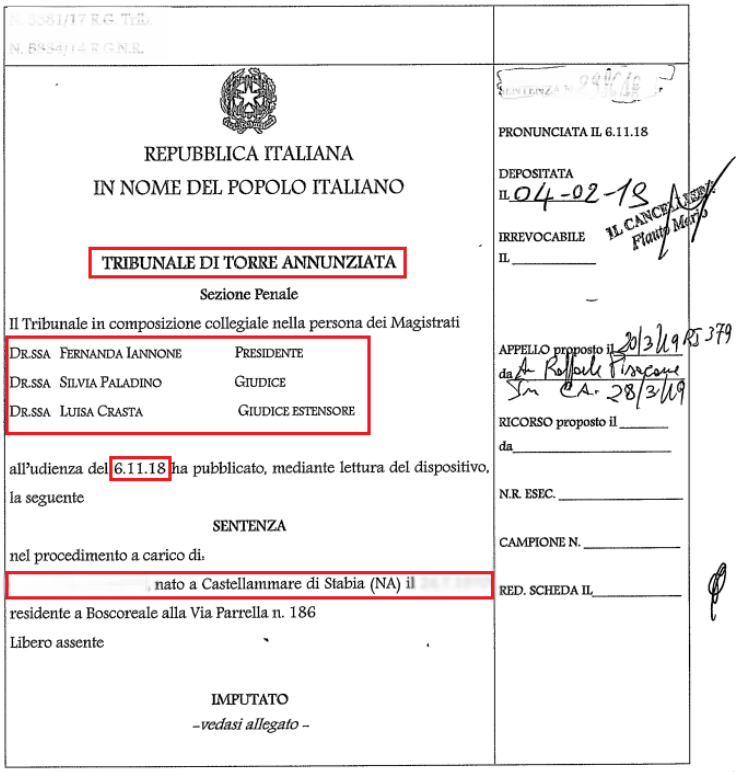
A Other Tables and Figures

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 Av. Raffaele Pinsone

Sin CA. 28/3/19

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 gender composition of the appellate court

	Quality (cond.)
All-men team	0.030
	(0.062)
Matching appeal team	0.026
	(0.068)
All-men team x Matching appeal team	-0.017
	(0.110)
2M1F team	-0.013
	(0.041)
2M1F team x Matching appeal team	0.039
	(0.083)
1M2F team	-0.039
	(0.039)
1M2F team x Matching appeal team	0.009
	(0.076)
Observations	10395
Court and crime x year FEs	X
Controls	X
Mean dep.var.	0.367
Effect of matching appeal team for all-men teams	0.009
Effect of matching appeal team for 2M1F teams	0.065
Effect of matching appeal team for 1M2F teams	0.035
Effect of matching appeal team for all-women teams	0.026

Notes. The dependent variable 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. *Mixed-gender team* is an indicator of whether the team consists of two men and one woman, or one man and two women.*Matching appeal team* is an indicator of whether the gender composition of the appellate team is the same as the one that ruled in the first instance. At the bottom of the table we report the effect of having an appeal team with the same gender composition for all team types. 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 A5: Characteristics of first-instance and appellate judges

	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)	
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 A6: Different definition of appeal-proof decisions

	Adding tolerance in jail time difference		
	(1) Unconditional	(2) Conditional	(3) Controlling for jail time
All-men team	-0.089**	-0.184***	-0.173***
	(0.036)	(0.051)	(0.050)
2M1F team	-0.003	-0.067***	-0.059**
	(0.013)	(0.026)	(0.025)
1M2F team	-0.008	-0.043*	-0.040*
	(0.013)	(0.024)	(0.024)
Observations	65051	27686	27686
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.882	0.729	0.729

Notes. In all columns, the dependent variable is an indicator for whether the team takes an *appeal-proof decision*. In Panel A, this is defined as a decision that is either not appealed or it is appealed and the verdict is confirmed. In Panel B, this is defined as a decision that is either unappealed or, if appealed, confirmed with a percentage difference (in absolute value) in the prison time between the first-instance and appellate court below the sample median (7.6%). 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 jail 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 A7: 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
Court and crime x year FEs	X	X
Controls	X	X
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
Court and crime x year FEs	X	X
Controls	X	X
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
Court and crime x year FEs	X	X
Controls	X	X
Mean dep.var.	0.438	0.707

Notes. The dependent variable in column (1) is 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 an 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 A8: Controlling for court-by-year fixed effects

	(1) Guilty	(2) Appeal-proof (cond.)
All-men team	-0.090***	-0.106**
	(0.029)	(0.041)
2M1F team	-0.097***	-0.052**
	(0.024)	(0.024)
1M2F team	-0.057**	-0.046**
	(0.024)	(0.022)
Observations	65050	27685
Court x year 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-by-year 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 A9: Controlling for judge fixed effects

	Guilty		Appeal-proof (cond.)	
	(1) Women	(2) Men	(3) Women	(4) 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 A10: Different classification of crimes

	(1) Guilty	(2) 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 A11: 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 crime type 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 A12: 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
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
Mean dep.var.	0.712	0.710	0.698	0.784
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X

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 A13: 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 A14: 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. Additionally, the US system is dominated by guilty pleas – about 90% of convictions are pleas (Motivans, 2025) – often with appeal waivers, so few convictions arise from trials and even fewer receive appellate review. 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. Due to these differences, our strategy to use appeal outcomes as performance benchmark to extract information on judges' diagnostic skills is typically infeasible in the U.S. setting. 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.[72] 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 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 limited to minor offenses, and is not considered an admission of guilt: the judge, after examining the case, may reject the agreement and acquit the defendant. Instead, plea-bargaining is a dominant feature of the U.S. system, where nearly 90% of convictions are from guilty pleas (Motivans, 2025). This

[72]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.

implies that very few convictions are the result of trials at all, and guilty pleas often come with appeal waivers, sharply limiting appellate review.

Appellate 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.[73]

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*).Instead, in the U.S. criminal trials, juries decide the facts and pronounce the verdicts, so that the trial decision aggregates lay fact-finding with limited scope to isolate judges' diagnostic ability.

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.

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