Gender diversity and decision-making in teams

Maddalena Ronchi Viola Salvestrini*

August 10, 2026

Abstract

We study how team gender composition shapes the nature and quality of decisions. Our setting is Italian criminal courts, where cases and judges are quasi-randomly assigned to three-judge panels, and appellate review provides a benchmark for decision quality. This allows us to separate gender differences in leniency from differences in diagnostic skill. All-women panels convict more often than all-men panels and issue more appeal-proof convictions - convictions not appealed or upheld on appeal. Mixed-gender panels convict at rates similar to all-men panels, but issue convictions nearly as appeal-proof as all-women panels. Leniency alone cannot explain these patterns. Instead, they point to gender differences in diagnostic skill and coordination frictions in mixed teams. Trial duration, written opinions, and entry-exam scores suggest the accuracy gap reflects effort rather than ability. A structural model traces these gaps to gender differences in effort - amplified by men's free-riding alongside more diligent women - and to coordination frictions.

JEL codes: J16, J82, M5, K4, D81

*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, Will Dobbie, Florian Englmaier, Elisa Facchetti, Alex Frenkel, Yana Gallen, Alfonso Gambardella, Matthew Gentzkow, Andrea Ichino, Ambar La Forgia, Niko Matouschek, Mikkel Mertz, Elia Moracci, Massimo Morelli, Barbara Petrongolo, Nicola Persico, Fabiano Schivardi, as well as seminar participants at the 2026 Morton Schapiro-ERC Gender Economics Workshop at Northwestern University, the 2025 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, the Institute for Human Studies, and from the Ghazala Azmat Award 2025.

1 Introduction

Teams are central to how work is organized in modern economies. In the United States, for example, nearly 80% of employment is in occupations where teamwork is *very* or *extremely* important (O*NET, 2022). At the same time, the rise in women's labor force participation and the decline in gender segregation across occupations have made mixed-gender teams increasingly common. Together, these trends make it both natural and important to ask whether and how team gender composition affects decision-making and performance. The answer is ex-ante ambiguous. Because men and women have been shown to differ in preferences, attitudes, values, and skills (e.g. Chattopadhyay and Duflo, 2004; Croson and Gneezy, 2009; Kim and Starks, 2016; Gallen, 2024), gender diversity may improve team performance through complementarities in knowledge and perspectives, but the same individual-level differences may also increase coordination and communication costs (Prat, 2002). Which force dominates in practice is an empirical question, yet causal evidence remains scant.

Three main challenges explain this knowledge gap. First, team composition – and the allocation of tasks across teams – is typically endogenous to both an organization and its workers, making it hard to credibly compare teams with different gender mixes. Second, in most real-world settings, the quality of team decisions is difficult to measure because objective benchmarks for performance are rarely observed. Third, most occupations remain sufficiently gender-segregated that there is often too little variation to compare all-men, mixed-gender, and all-women teams performing the same job.

In this paper, we overcome these challenges by studying the Italian criminal justice system. Three of its institutional features let us provide causal evidence on how gender composition shapes the nature and quality of team decisions. First, serious criminal offenses are tried by collegial panels of three judges whose convictions are treated as requiring unanimity, and both judges and cases are randomly assigned to panels within courts.[1] This generates within-court variation in panel gender composition that is independent of case characteristics, and lets us estimate the causal effect of gender diversity on judicial rulings. Second, the judiciary's hierarchical structure provides a natural benchmark for decision quality. Rulings that are appealed are reviewed by a higher court, whose patterns of sentence reversal provide an empirical proxy for the accuracy of first-instance rulings. Third, in many developed countries, including Italy, the judiciary is a gender-balanced profession.[2] This provides enough variation to study all-men, all-women, and mixed-gender panels separately, and to test whether gender homogeneity masks differences between all-men and all-women teams.

Leveraging these features, we answer three main questions. First, we examine the causal

[1] The Italian Constitution mandates these rules, ensuring assignment does not depend on gender. Section 2.1 describes the assignment rules, and Section 5.2 shows that (i) team gender composition is unrelated to case characteristics, (ii) judges in mixed- and same-gender teams do not differ on observables, and (iii) the observed distribution of teams types is consistent with random formation.

[2] For instance, the most recent data from the Federal Judicial Center show that nearly 40% of US District Court judges are female. Comparable data for other countries (UNECE) indicate that women represent about 50% of judges in Canada, Germany, Italy, and Norway, and 65% in France.

relationship between the gender composition of judicial teams and the *type* of decisions they take – measured by conviction rates – as well as their *accuracy* – measured by "appeal-proof" decisions, i.e., rulings that are either unappealed or upheld on appeal.[3] Second, we ask what drives these differences. In the standard judges-design framework, judges share the same assessment of cases and differ only in where they place the conviction threshold, i.e., in their *leniency*. This assumption has a sharp prediction: when signal quality is the same across teams, the team that convicts *more often* should have a *higher share* of wrongful convictions, because its marginal convictions rest on weaker evidence of guilt. The joint distribution of conviction rates and appeal-proof rates thus lets us test this benchmark against the alternative that judicial panels also differ in *diagnostic skill*, i.e., in how well they separate guilty from innocent defendants. Conditional on finding skill differences, we then ask whether they reflect innate ability or endogenous effort. Third, guided by these findings, we develop and estimate a model of team diagnostic decision-making to disentangle individual-level drivers from team-level coordination frictions that arise when diverse members must reach consensus, and to trace how each shapes different dimensions of team performance.

To answer these questions, we construct a novel database containing the universe of collegial rulings issued by 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 our analysis. Specifically, we obtain data from four courts in Central and Southern Italy and the corresponding Courts of Appeal.[4] The presidents of each court made the sentences available in paper format, which we digitized and processed using OCR and text-parsing algorithms. We further augment these records with judges' entry-exam scores, digitized from nationwide public competitions for judicial appointments, and with age and place of birth collected from public sources. The resulting dataset covers more than 65,000 judicial decisions. For each one, we observe the composition and characteristics of the judicial panel, the characteristics of the defendant and the crime, the duration of the trial, the outcomes of the first-instance ruling and, where applicable, its appeal. For three of the four courts, we additionally have the full text of the written reasoning that accompanies each verdict.

Using a specification with court fixed effects and crime-by-year fixed effects – thus comparing teams within the same court ruling on comparable crimes at the same time – we first show that judicial teams reach different sentencing outcomes depending on their gender composition. Relative to all-women teams, all-men teams are 7.1 percentage points (14.6%)

[3] 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 Section 5.

[4] 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 heterogeneity in courts' characteristics reassures us that our results should not be specific to a particular type of court or area.

less likely to convict, and mixed-gender teams with one or two women are 9.3 to 5.7 percentage points less likely to convict, respectively. Conviction rates for all-men and mixed-gender teams are statistically indistinguishable – as are those for mixed teams with one versus two women. 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 also find that conditional on conviction, all-women teams 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. Italian appellate review has two features worth highlighting. First, the appellate panel reviews the same evidentiary record as the first-instance court, so a reversal reflects a different reading of the same facts rather than the emergence of new evidence. Second, unlike in common-law systems where appeals are confined to questions of law, Italian appellate review is a *giudizio di merito*: the appellate panel re-weighs the evidence and can overturn the first-instance ruling whenever it concludes that the evidence was interpreted illogically or incorrectly. A reversal can therefore signal a substantive error – not only a procedural one – in the first-instance ruling.

We start from the *perceived accuracy* of a ruling, proxied by its likelihood of being appealed. All-men teams' decisions 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 result persists when we control for sentence length, so it does not reflect harsher sentencing in all-men panels. We then measure teams' *appeal-proof decision rates* by examining whether a ruling is either not appealed or appealed and confirmed by an independent panel of appellate judges. All-men teams are 17.7 percentage points (22.6%) less likely than all-women teams to issue appeal-proof convictions. The presence of at least one woman substantially improves decision accuracy. While all-women teams achieve the highest appeal-proof conviction rate, mixed-gender teams trail by only 5-6.7 percentage points, with no significant difference between the two mixed-gender configurations. We show that these results on conviction rates and decision accuracy reflect the impact of gender rather than other judicial characteristics correlated with gender (e.g. age, experience, or region of origin), and are not driven by particular crime types or by defendant gender.

The combination of higher conviction rates and fewer reversals for all-women teams rules out leniency as the only driver of our results. Under a leniency-only mechanism, a team that convicts more admits weaker marginal cases and should therefore face more reversals, not fewer. Instead, all-women teams convict more and are more accurate, pointing to the presence of differences in teams' ability to distinguish guilt from innocence.[5] Although we

[5]Thus, a standard research design that uses the random assignment of cases to judicial panels as an instrument for conviction (the so-called "judge IV") 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.

only observe accuracy for convictions, random case assignment extends this conclusion to acquittals as well: because the random assignment of cases to teams implies that all teams face the same distribution of guilty and innocent defendants on average, a team that convicts more and is more accurate must also produce fewer wrongful acquittals.[6] Finally, mixed-gender teams do not fall between same-gender teams. Their convictions are nearly as likely to be appeal-proof as those of all-women teams, but their conviction rates are closer to those of all-men teams, suggesting that team performance is not a simple linear aggregation of members' characteristics. Instead, it points to coordination frictions: gender-diverse teams appear less able to translate similarly accurate evidence into unanimous convictions. The presence of at least one woman therefore improves conviction accuracy relative to all-men teams, while mixed-gender frictions reduce the frequency of convictions.

Establishing that teams differ in diagnostic skill raises a natural next question: do these differences reflect fixed ability or endogenous effort? The answer has direct policy implications: ability gaps imply that each judge's contribution to team performance is largely predetermined, so improving outcomes would require changing team composition. Effort gaps, by contrast, are malleable, leaving scope for organizational design to improve outcomes. We present evidence consistent with differences in effort, and further corroborate the presence of coordination frictions in mixed-gender teams. First, using digitized rankings from nationwide public exams for entry into the judiciary we find no gender differences in performance, indicating that women are not positively selected at entry.[7] This interpretation is reinforced by specifications with judge fixed effects, which absorb judge ability and other time-invariant individual characteristics and leave the results unchanged. Second, trials overseen by all-men teams are shorter than those overseen by mixed-gender and all-women teams, whose trials are 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 trials; instead, all-women teams achieve better outcomes *and* take longer to reach a verdict, consistent with greater effort invested in each case. Moreover, despite the same trial length, mixed-gender teams reach different conviction rates than all-women teams – consistent with frictions in mixed teams, where part of that time is absorbed by difficulties in reaching agreement. Finally, we construct several measures of writing quality from the written reasoning of each verdict. Across all measures, all-men teams produce less clear and concise judgments than other teams, while mixed-gender and all-women teams perform similarly – suggesting comparable effort devoted to judicial reasoning.[8]

Motivated by this evidence, we propose and structurally estimate a model of team decision-making in the spirit of Chan et al. (2022) that formalizes two mechanisms through which gender composition shapes team performance: effort-driven differences in diagnostic

[6] This same argument is made by Chan et al. (2022) in the context of radiologists diagnosing pneumonia.

[7] As discussed in Section 2, the possibility to measure baseline ability through standardized entry exams is a distinctive feature of our setting, unavailable in the U.S. where judges are typically elected.

[8] Italian appellate review is a *giudizio di merito*: the appellate panel reassesses the evidence itself, looking for flaws in its interpretation. Differences in the clarity of the written motivation thus cannot, on their own, account for differences in reversal rates across teams.

skill, which govern information production, and coordination frictions in the translation of that information into collective decisions.[9] 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, reflecting the long-standing Blackstone principle that convicting an innocent person is worse than acquitting a guilty one. We define judges' overall aversion to mistakes as their diligence, and the additional weight placed on wrongful convictions as their leniency. 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 conviction quality and – if the prior probability of guilt is sufficiently high (as is plausible for cases that have survived charging and reached trial) – also conviction rates. When both judges reach the same assessment of the evidence, the team's decision follows directly: they convict if the evidence is sufficiently precise to meet the standard of proof, and acquit otherwise. When they receive different assessments, they enter a deliberation process. With some probability they reach agreement and one judge's view prevails; if instead they do not, the presumption of innocence implies that the panel acquits by default. Higher frictions therefore reduce convictions but, because they do not affect the effort judges exert and consequently the precision of the common information environment, they do not reduce the quality of convictions.

We structurally estimate the model and find sizable gender gaps in diligence – higher for women – and no significant gaps in leniency. Women's higher diligence leads them to exert more effort, generating the higher conviction rates and more appeal-proof convictions in all-women teams documented in the reduced form. Team diversity has a differential impact on men's and women's effort. Because a judge's optimal effort decreases with teammates' diligence, men in mixed-gender teams free-ride and exert less effort than in all-men teams, while women over-compensate by working harder relative to 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 those of all-men teams and close to those of all-women teams. Moreover, we estimate greater coordination frictions in mixed-gender panels, which explains why mixed teams convict less frequently than all-women teams, despite similar levels of accuracy.

Finally, we use the model to study two policy-relevant counterfactuals. First, since performance differences across teams operate through an effort-based channel, we ask how far organizational levers that affect behavior on the bench – e.g. through scrutiny or incentives – could narrow the observed gaps. We find that if men were as diligent as women, the accuracy

[9]The idea that individual contributions to team output reflect endogenous and costly effort choices is standard in two related literatures: the committee-design literature, in which members acquire costly information before voting (e.g., Persico, 2004; Gerardi and Yariv, 2008); and the team-production literature, where effort responds strategically to team composition (e.g., Holmstrom, 1982; Itoh, 1991; Kandel and Lazear, 1992; Mas and Moretti, 2009; Bandiera et al., 2010).

gap would largely disappear; by contrast, removing coordination frictions in mixed-gender teams would close the conviction-rate gap with little effect on accuracy. Second, we ask what the Italian judiciary's ongoing feminization implies for aggregate court performance. Women, long a minority in the judiciary, have steadily increased their representation and now make up the majority of new entrants.[10] We find that rising female representation raises aggregate decision accuracy at the cost of longer trials – an accuracy-speed trade-off that will grow more salient as this trend continues. Aggregate conviction rates instead follow a U-shape: they are lowest at the current, most mixed composition, where exposure to coordination frictions is greatest, and rise toward either gender extreme.

This paper contributes to three strands of the literature. First, it advances research on gender diversity and team performance. A large body of work documents that men and women differ in preferences, skills, and social values (e.g., Andreoni and Vesterlund, 2001; Croson and Gneezy, 2009; Kim and Starks, 2016; Gallen, 2024), and that these differences can reshape outcomes when the gender composition of professions and institutions changes (e.g., Schaede and Mankki, 2026; Truffa and Wong, 2025; Matsa and Miller, 2013). Whether and how these differences matter when men and women work together *as a team* is less well understood. Theory highlights competing forces – complementarities in skills and perspectives versus coordination costs (Lazear, 1999; Prat, 2002) – while empirical evidence comes largely from laboratory or classroom experiments with students (e.g., Apesteguia et al., 2012; Hoogendoorn et al., 2013; Calder-Wang et al., 2021; Hardt et al., 2025), with few exceptions from professional settings (La Forgia and Singh, 2026; Gupta et al., 2025).[11,12] We advance this literature by providing novel causal evidence on both the effects of gender composition and the channels through which it shapes team performance. Our high-stakes setting features random assignment and an external benchmark for decision quality, letting us study both the nature and the accuracy of team decisions. This allows us to separate decision thresholds from diagnostic skill and to distinguish both from coordination frictions. We find that gender composition does not affect team leniency: men and women apply similar standards in deciding cases. Instead, gender composition shapes team performance through two channels: the diligence teams devote to producing information – which determines their

[10] 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. See Figure A1 of the Supplemental Appendix.

[11] La Forgia and Singh (2026) study two-physician obstetric teams, finding that women-only teams achieve the best maternal outcomes and that team gender mix matters beyond individual physicians' C-section preferences. Our setting differs by offering an external benchmark for decision quality, allowing us to separate decision thresholds from diagnostic accuracy. Gupta et al. (2025) study U.S. federal rulemaking teams and show that female-inclusive teams deliberate more but complete fewer final rules, attributing the gap to differences in project assignment, incentives, and evaluation. Our random-assignment setting abstracts from these organizational frictions and isolates how gender composition affects decision-making within teams assigned to comparable tasks.

[12] Two related literatures study gender composition in group settings but address different questions. One examines how women's numerical presence in a group shapes their influence on collective decisions (e.g., Born et al., 2022; Karpowitz et al., 2024; Radbruch and Schiprowski, 2024; Rebolledo et al., 2024); the other investigates whether gender-diverse evaluating bodies benefit women among those evaluated (Bagues and Esteve-Volart, 2010; Bagues et al., 2017; Bertrand et al., 2019; Deschamps, 2024). Unlike both, we focus on how gender composition affects the quality of the team's joint output.

diagnostic skill – and the frictions they face in reaching agreement.

Second, this paper relates to research on team production and the determinants of individual effort within teams. When effort benefits the team as a whole, each member has an incentive to free-ride (Holmstrom, 1982). How organizational design and team composition shape these incentives is a central question in personnel economics. On the first dimension, prior work shows that organizational design can mitigate moral hazard through team-based pay, mutual monitoring, or hierarchical roles that facilitate information aggregation (Chan, 2016; Friebel et al., 2017; Chan, 2021). Turning to composition, existing work emphasizes how coworkers shape effort through peer pressure, social ties, or ethnic rivalry (Mas and Moretti, 2009; Bandiera et al., 2010; Hjort, 2014). We document a new mechanism linking team composition to performance: judges strategically adjust their effort to teammates' intrinsic diligence, which varies systematically by gender. Men free-ride when paired with more diligent women, while women over-compensate – a pattern distinct from peer pressure or inter-group conflict, and one that our structural estimates identify as the main driver of cross-team differences in decision accuracy. These findings imply that team performance cannot be inferred from a linear aggregation of individual traits, since the same judge exerts different effort depending on the panel's gender mix.[13]

Finally, this paper relates to the large literature on differences in judicial decision-making. One strand documents disparities in case outcomes across demographic groups, such as gender or ethnicity (e.g., Anwar et al., 2012; Alesina and La Ferrara, 2014; Knepper, 2018; Arnold et al., 2018; Abrams et al., 2022; Ash et al., 2024). These papers ask whether outcome gaps reflect bias, without decomposing judicial heterogeneity into preferences and diagnostic skill. A second strand uses random assignment of judges as an instrument for individual decisions in courts and beyond (e.g., Maestas et al., 2013; Galasso and Schankerman, 2015; Dobbie et al., 2018). This "judges design" relies on the idea that judges differ in their propensity to convict while sharing the same underlying assessment of case quality. A smaller literature on Italian civil and labor courts takes the opposite approach, allowing judges to differ in skill and effort while holding preferences over case outcomes fixed (Coviello et al., 2015; Coviello et al., 2019). We advance this line of work by allowing both margins to vary: using appeal outcomes as an external benchmark for decision accuracy, we distinguish differences in judges' preferences from differences in diagnostic skill and study how both interact within teams.[14] This distinction has broader implications for the "judges design." As emphasized by Chan et al. (2022) and Mungan (2024), monotonicity holds only when heterogeneity in decision-making is preference-based rather than skill-based. We extend this single-agent insight to teams: diagnostic skill is not only heterogeneous across decision-

[13]Evidence that women exert more effort than men has been documented in comparable professional tasks, including medicine, academia, and corporate boards (Coviello et al., 2025; Chu et al., 2022; Hengel, 2022; Gupta et al., 2019; Adams and Ferreira, 2009). These studies establish individual-level gender differences but do not examine how those differences interact with team incentives.

[14]Using appeal outcomes as a performance benchmark is typically infeasible in the U.S.: about 90% of convictions arise from guilty pleas (Motivans, 2025), often with appeal waivers, so few convictions arise from trials and even fewer receive appellate review; juries render most criminal verdicts, leaving little scope to isolate judges' diagnostic ability; and case-level linkage between trial and appellate records is rarely available.

makers, but also endogenous to team composition through strategic effort provision, calling into question designs that assume uniform skill across judges and other decision-makers.

The remainder of this paper is organized as follows. Section 2 describes the institutional setting. Section 3 sets up a high-level conceptual framework for our analysis. Section 4 outlines the data and the main outcomes of interest. Section 5 presents the empirical strategy and discusses its validity. Section 6 presents the reduced-form results. Section 7 discusses the plausible mechanisms underlying our findings and presents our structural analysis. Finally, Section 8 concludes.

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. Our analysis focuses on the criminal proceedings adjudicated by criminal courts (*Tribunali ordinari*).[15] Within criminal tribunals, the mode of adjudication depends on the severity of the charge: less serious offenses are heard by a single judge, while more serious ones are tried by a panel of three judges.[16] We focus on the decisions of these three-judge panels.

Crucially for our purposes, Italy's civil-law tradition provides a 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 the limited scope of appellate review over factual findings can complicate the mapping from judicial composition to observable case outcomes.[17]

Similar to other developed countries, Italy has witnessed a transformation in the gender composition of its workforce, including in the judiciary. 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.[18] 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.[19] This provides the necessary variation to assess whether

[15] 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.

[16] Offenses under collegial jurisdiction are regulated by Art. 33 bis c.p.p. and include mafia-type associations, arms and drugs trafficking, sexual violence, and crimes against public safety and order.

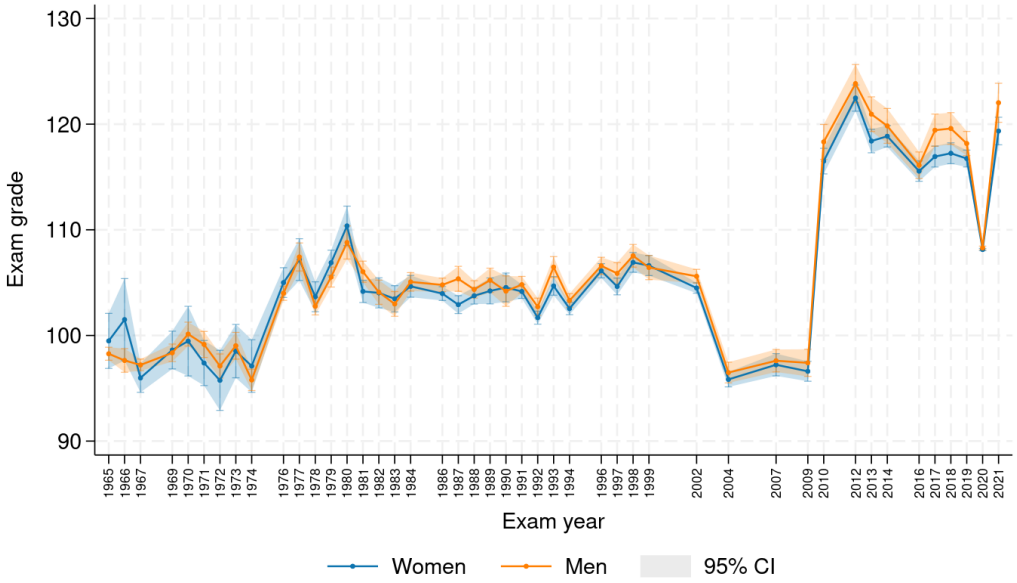
[17] Supplemental Appendix B compares the Italian and U.S. criminal justice systems and discusses why the Italian setting is especially well suited to our research question.

[18] Such shift in gender composition is illustrated in Figure D1 of the Supplemental Appendix.

[19] A potential concern is that the gender composition effects we estimate are driven by a cohort of older male judges who entered the profession when it was overwhelmingly male-dominated, and whose training or attitudes may differ systematically from later cohorts. As our data cover years from 2004 onwards, only 6% of decisions in our sample involve judges who entered the judiciary before 1980, limiting the relevance of such cohort effects in our analysis.

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 (e.g. Bagues et al., 2017; Radbruch and Schiprowski, 2024; Rebolledo et al., 2024), where all-women teams are too rare to be analyzed separately. Although the gender composition of the judiciary has changed over time, scores on the national entry examination do not indicate differential selection by gender during this period. Figure 1 shows that the distribution of grades over time is nearly identical for men and women. We return to this point in Section 7.1.

Figure 1: Grades differences between men and women over time



Notes. The figure shows the distribution of the judiciary national exam grades over exam year for men and women, with 95% confidence intervals.

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 5.2.

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.[20] These criteria must comply with the constitutional requirements of objectivity, predetermination, and impartiality (Art. 25 of the Italian Constitution).

[20]The CSM is a constitutional body 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.

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 specifies how long teams remain together, either fixing them for the full four-year term or reshuffling composition at pre-established intervals. In either case, any substitutions (e.g., due to illness or transfer) follow the same predetermined, objective rules.[21]

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 teams 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.[22] Judges cannot decline assigned cases except under narrow statutory grounds (e.g., illness, extended leave, or circumstances that might compromise impartiality), in which case substitution follows pre-approved random procedures.

Decision-making process. Within the team, judges assume different formal roles. The presiding judge – appointed by the court president based on seniority and experience – directs the trial and coordinates deliberation, while the judge rapporteur drafts the written reasoning. Interviews with judges indicate that these formal distinctions do not translate into meaningful differences in influence or task allocation: deliberation is effectively collegial, and the drafting of the written opinion is often a shared effort rather than the rapporteur's sole responsibility. Finally, although decisions are formally adopted by majority vote, no formal vote is held in practice, 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. The Court of Appeal hears all cases in three-judge panels, whose members are independent from the judges involved in the original proceedings. Both judges and cases are allocated to panels randomly, under the same constitutional principles described in Section 2.1. Appeals may be initiated by either the defendant or the public prosecutor. However, when only the defendant appeals, the Court

[21] Substitutions and reshuffles apply only to new cases: once a team begins hearing evidence, its composition is locked in until the final ruling (Art. 525 c.p.p.). If a judge *must* be replaced mid-trial, the trial must start over and the evidentiary phase repeated.

[22] 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 based on case complexity, ensuring a fair and balanced distribution of workload.

of Appeal can only uphold, reduce, or overturn the initial sentence in the defendant's favor, creating a strong incentive to appeal. As a result, appeals are overwhelmingly advanced by defendants.[23]

The Court of Appeal may confirm, modify, or overturn the original ruling, based on a full review of both the *evidence* and the application of the *law*. Crucially for our empirical strategy, appellate judges decide the case based on the same information available to the trial court: new evidence is generally not admissible, and witnesses are not retried.[24] Moreover, appellate courts are composed of judges with greater seniority and experience than those on first-instance courts, a point we return to in Section 5.2. Cases may subsequently reach the Supreme Court of Cassation, but its review is limited to matters of law; hence, the Court of Appeal is the last venue where errors in fact-finding can be corrected. For this reason, we focus on the first round of appeal.

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. Unlike in the United States, judges are not elected; instead, entry into the judiciary is obtained through a nationwide public examination.[25] 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. Successful entrants are initially appointed as judicial trainees for an 18-month program. At the end of the traineeship, the 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.[26]

Once in service, judges receive a fixed salary that increases with seniority according to a structured pay scale. They are subject to performance evaluations conducted every four years by the CSM. These evaluations assess three core dimensions: (i) productivity; (ii) timeliness; and (iii) quality of legal reasoning (based on a sample of the judge's written judgments).[27] Together with seniority, they determine promotion to higher courts and to administrative roles, as well as eligibility for transfers. Judges cannot be dismissed or demoted except through formal disciplinary proceedings, safeguarding their independence from political or administrative pressures.

[23] Between 2012-2016, public prosecutors file only 5% of all appeals (Della Torre, 2022).

[24] The introduction of new evidence is permitted only in exceptional cases, when the appeal panel deems it absolutely necessary for the decision. Such requests are rare and granted infrequently.

[25] 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. Candidates must hold a law degree and a two-year specialization diploma.

[26] Judges must remain in their first seat for at least three years before requesting a transfer. Transfers are granted by the CSM based on seniority, evaluation scores, family circumstances, and staffing needs.

[27] The evaluations can be positive, non-positive, or negative. A non-positive evaluation delays advancement to the next salary class by one year, while a negative evaluation delays it by two; a second negative evaluation results in dismissal. In practice, these scenarios are rare: according to CSM Statistical Office data, over 98% of evaluations have a positive outcome.

3 Conceptual Framework

Consider a judicial panel deciding whether a defendant is guilty or innocent. Following Chan et al. (2022), we conceptualize the team's decision-making process as consisting of two stages. The first is an *evaluation* stage, in which the panel develops a shared information base from the available evidence. The quality of this information base depends on the team's *diagnostic skill* – its ability to distinguish guilty from innocent defendants – which we denote by θ. A higher θ means that a truly guilty defendant is more likely to be recognized as guilty, and a truly innocent defendant is more likely to be recognized as innocent. Each team member then draws on this shared information base to independently form an overall judgment of how likely the defendant is to be guilty, which we refer to as the defendant's *assessed guilt*.

Second, a *decision* stage in which, given their assessments, team members decide whether to convict, weighing the cost of wrongful conviction against that of wrongful acquittal. We refer to the relative weight placed on wrongful convictions as the team's *leniency*, β. A more lenient team (higher β) requires stronger evidence before convicting – it sets a higher conviction threshold along the scale of assessed guilt. Because conviction requires unanimity, the team can only convict when all members' assessments clear this threshold. Conviction requires establishing guilt beyond a reasonable doubt. When team members' independent assessments point in different directions, that standard is not met, and the team must acquit unless deliberation resolves the disagreement. We denote q the probability that the team resolves such disagreements. A lower q implies greater *coordination frictions*.

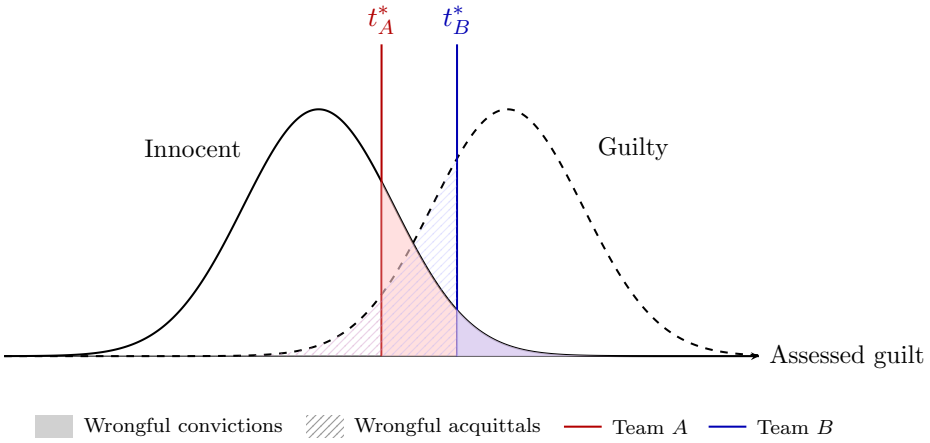
These three forces – diagnostic skill, leniency, and coordination frictions – generate distinct observable patterns depending on which force varies across teams, each of which yields a testable prediction.

Differences in leniency. Consider two teams with the same diagnostic skill θ but different levels of leniency β. Because they rank defendants in the same way by assessed guilt, differences in conviction rates arise only from differences in the conviction threshold. The less lenient team applies a lower threshold and therefore convicts more often. Because the quality of information is the same for both teams, its additional convictions are drawn from marginal cases where guilt is most uncertain and where guilty and innocent defendants are hardest to distinguish. As a result, these additional convictions are more likely to involve innocent defendants, implying that the less lenient team has both a higher conviction rate and a higher rate of wrongful convictions (Figure 2). This is the logic of standard models of judicial decision-making (e.g., Galasso and Schankerman, 2015; Dobbie et al., 2018), and corresponds to the monotonicity assumption underlying the "judges design" literature, where random assignment to more or less lenient judges is used as an instrument for conviction.

> **Hypothesis 1** (Leniency)**.** *If teams differ only in their leniency β, a team that convicts more frequently should have a higher share of wrongful convictions.*

Differences in diagnostic skill. Now consider two teams that are equally lenient, applying the same conviction threshold β, but differ in diagnostic skill θ. The more skilled team

Figure 2: Convictions and mistakes under differences in leniency



Notes. Both teams observe signals of the same precision. Team A applies a lower conviction threshold ($t_A^* < t_B^*$), convicting more often. Because diagnostic skill is identical, the additional convictions come from the overlap region where guilt is uncertain, increasing the share of wrongful convictions. Under this scenario, teams that convict more frequently necessarily have lower conviction accuracy.

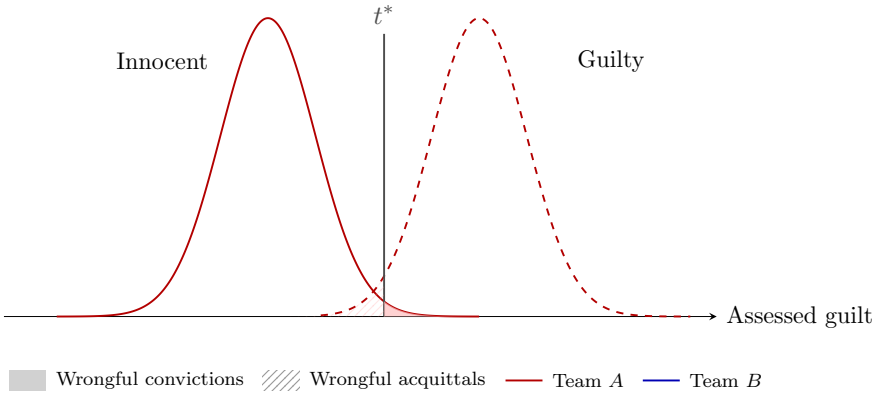
evaluates evidence more precisely, placing truly guilty defendants higher, and truly innocent defendants lower, on the scale of assessed guilt. With better-separated distributions, more guilty defendants clear the conviction threshold while fewer innocent defendants do. The more skilled team can therefore convict more often *and* make fewer errors (Figure 3). This pattern breaks the monotonicity assumption discussed above. Because teams differ in the precision of their assessments rather than in the threshold they apply, case rankings can cross: the same defendant, facing the same conviction standard, can be convicted by one team and acquitted by another.

Hypothesis 2 (Diagnostic skill)**.** *If teams differ only in their diagnostic skill θ, a team that convicts more often can simultaneously have a lower share of wrongful convictions.*

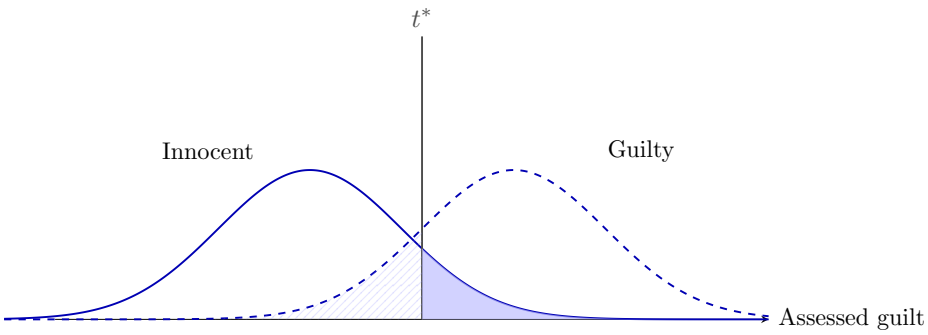
Coordination frictions. Finally, consider two teams with the same diagnostic skill θ and leniency β, but different ability to resolve disagreement among their members. Because convictions require the team to agree that guilt has been established beyond a reasonable doubt, unresolved disagreement leads to acquittal. Teams facing greater coordination frictions therefore convict less often, not because they evaluate the evidence less accurately or apply a higher threshold, but because they are less able to translate their assessments into a unanimous decision. The convictions that do occur are unaffected in quality: frictions prevent some convictions from being reached, but do not alter the accuracy of the underlying evaluation.

Hypothesis 3 (Coordination frictions)**.** *If a team faces greater coordination frictions, its conviction rate will be lower than predicted by its diagnostic skill, without a corresponding change in the share of wrongful convictions.*

Figure 3: Convictions and mistakes under differences in diagnostic skill



(a) Team A: high diagnostic skill



(b) Team B: low diagnostic skill

Notes. The figure shows a scenario where both teams apply the same conviction threshold t^*. Team B (top) has low diagnostic skill: its broad, overlapping distributions produce substantial wrongful convictions (solid shading) and wrongful acquittals (hatched area). Team A (bottom) has higher diagnostic skill: its tighter distributions place defendants more accurately along the scale of assessed guilt, reducing both types of error. With the same evidentiary standard, higher diagnostic skill simultaneously raises conviction rates and improves accuracy.

Section 6 uses these predictions to interpret the reduced-form evidence. Section 7.2 then develops the full structural model, which microfounds and quantifies the three channels.

4 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 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, helping ensure that our estimates are not driven by the idiosyncrasies of a single court or a single geographic area. Specifically, 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.

The sentences were provided in paper format by the presidents of each court.[28] We digitized the sentences issued in collegial rulings and processed them using OCR and text-parsing algorithms to extract the following information: the name and location of the court, the start and end dates of the trial, the identities of the judicial team members, the crime under prosecution, and the final verdict. Since Italian first names unambiguously indicate an individual's gender, we infer the gender of judges based on their first names as collected during the scraping process.[29] Supplemental 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 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, as measured by the judiciary's own selection criteria.[30] Finally, for 96.2% of our judges we retrieved additional personal information – such as judges' age, municipality, and region of birth – through online sources.

4.1 Descriptive statistics

Our sample covers the years 2004 to 2021.[31] 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.

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 just below 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) and less experienced (19 vs 22 years of service). They do not differ from male judges, however, in entry examination performance or in region of origin.

Among the 1,134 unique teams we cover in our dataset, the majority are mixed-gender

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

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

[30] We digitized all judicial selection exams from 1965 to 2007; exams from 2008 onward are publicly available on the Ministry of Justice website. We were unable to retrieve the rankings for three national exams. Overall, we were able to match 92.5% of all judges in our sample who undertook a national exam.

[31] 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.

Table 1: Descriptive statistics – judges

	All	Women	Men	Diff.
Share women	0.561			
	(0.497)			
Age	48.588	46.875	50.747	3.872***
	(9.200)	(8.521)	(9.589)	
Experience	20.429	19.039	22.156	3.118**
	(10.091)	(9.035)	(11.060)	
Entry exam grade	106.376	106.252	106.531	0.280
	(10.079)	(9.423)	(10.874)	
From central Italy	0.148	0.170	0.121	-0.049
	(0.356)	(0.377)	(0.327)	
From northern Italy	0.040	0.031	0.051	0.020
	(0.196)	(0.174)	(0.221)	
From southern Italy	0.812	0.799	0.828	0.029
	(0.391)	(0.402)	(0.379)	
From a diff region	0.196	0.160	0.241	0.081
	(0.397)	(0.367)	(0.428)	
N. judges	371	208	163	371

Notes. This table reports descriptive statistics for the judges in our sample. *Female* is an indicator for whether the judge is a woman. *Experience* is the number of years working as a judge. *Entry exam grade* is the grade obtained in the national examination to enter the judiciary. *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. Mean and standard deviation (in parentheses) are reported for the full sample and separately by gender. * p< 0.1, ** p< 0.05, *** p< 0.01.

(Table 2, Panel A): 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%).[32] Panel B of Table 2 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 C of the same table summarizes the 65,051 decisions issued by judicial teams in our dataset. A guilty verdict is issued in 43% of decisions. 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 (94%, not reported in the Table). This implies that 42% of convictions are appealed.

The types of crimes observed in our sample are summarized in Supplemental Appendix Table A1. The four most frequent categories – accounting for nearly 80% of all offenses – are crimes against property (e.g., theft, robbery, fraud), public order (e.g., mafia-type association, drug and arms trafficking), public trust (e.g., document forgery), and the person (e.g., personal injuries, manslaughter, threats, and crimes targeting women and children).

[32] In Section 5.2, we show that the observed distribution of team gender composition is consistent with random assignment from a pool with the same female share as our sample.

Table 2: Descriptive statistics – teams, cases and decisions

	Mean	St.dev
A. Panels		
All-men team	0.063	0.244
2M1F team	0.335	0.472
1M2F team	0.417	0.493
All-women team	0.184	0.388
N. panels	1134	
B. Case characteristics		
Number of defendants	2.257	4.657
Number of crimes	5.515	19.402
Defendant is foreign	0.184	0.380
Defendant is female	0.108	0.270
N. Cases	10676	
C. Decisions characteristics		
Guilty	0.426	0.494
Jail time (cond.)	5.215	4.797
Length of trial (years)	2.316	1.704
Appeal	0.190	0.392
Appeal (cond.)	0.422	0.494
N. Decisions	65051	

Notes. Panel A 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 B describes case characteristics. Panel C 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.

5 Empirical strategy

Identifying the causal effect of team gender composition on decision-making is challenging because team composition and task assignment are typically endogenous, and the intrinsic quality of team decisions is difficult to measure in most settings. The Italian judiciary addresses both challenges: judges and cases are randomly allocated to panels within courts, and a hierarchical appellate structure provides a clear benchmark for decision accuracy. This allows us to compare decisions across teams operating in the same court and ruling on the same type of crime in the same year, but differing in gender composition. We focus on two outcomes: the *type* of decisions, measured by conviction rates, and its *accuracy*, measured by the rate of "appeal-proof" rulings – decisions that are either not appealed or upheld on appeal.

The validity of this strategy rests on two assumptions. First, that unobservable characteristics of cases and judges do not vary systematically across team types within a court. Second, that appellate review provides a meaningful benchmark for decision accuracy; in particular, this requires that (i) appellate panels do not treat decisions differently based on the gender composition of the trial panel, and (ii) appellate judges are at least as accurate as first-instance judges, so that a reversed decision is, on average, less likely to be correct

than those that are not. We provide empirical support for both assumptions in Section 5.2.

5.1 Regression model

We estimate the following specification:

$$Y_{isct} = \alpha_0 + \alpha_1 \text{All-men}_{isct} + \alpha_2 \text{2M1F}_{isct} + \alpha_3 \text{1M2F}_{isct} + \delta X_{ict} + \eta Z_{isct} + \gamma_c + \gamma_{s,t} + \epsilon_{isct} \quad (1)$$

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 All-men_{isct}, 2M1F_{isct}, and 1M2F_{isct}, which are indicators for teams with three men, two men and one woman, and one man and two women, respectively. The omitted category is all-women teams. Because of random allocation of judges and cases to teams within courts, α_1, α_2, and α_3 describe the causal effect of having zero, one, or two women on the judicial team – each relative to a team with three women.

The coefficients γ_c and $\gamma_{s,t}$ denote court fixed-effects and type of crime-by-year fixed effects, respectively. Court fixed effects absorb time-invariant differences across courts; crime-by-year fixed effects ensure that we compare teams ruling on similar crimes at the same time, absorbing crime-specific trends such as shifts in legal interpretation or public sensitivity around specific offenses.[33] Finally, we include a vector of controls for team characteristics X_{ict} and 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.[34]

5.2 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 is whether sufficient variation in team gender composition remains after conditioning on court and crime type-by-year fixed effects. Supplemental Appendix Table A2 shows that over 93% of the original variation in each team type survives the inclusion of fixed effects, confirming that the gender composition of a team is not concentrated within specific courts or crime types, but varies idiosyncratically within them.

Random assignment of judges and cases to teams. A causal interpretation of our estimates requires that variation in team gender composition within courts reflects random fluctuations rather than systematic selection. We provide three pieces of evidence supporting

[33] Estimating more restrictive models that include court-by-year fixed effects or court-by-crime-by-year yields virtually identical results (Supplemental Appendix Table A9).

[34] 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.

Table 3: 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.

this assumption. First, we test for selection on observables by regressing case characteristics on team composition indicators, controlling for court and year fixed effects. Table 3 shows that only one coefficient is significant at the 5% level, as would be expected by chance, indicating that cases are well balanced across team types. Second, Table 4 shows that judges assigned to mixed- and same-gender teams do not differ in observable characteristics. Third, we verify that the gender composition of judicial teams we observe is consistent with random formation. We repeatedly draw triads of judges without replacement from each court's pool, preserving, within each court, both the number of panels and the number of male and female judges. Comparing the simulated and observed distributions, we find no statistically significant difference: the Chi-square test shown in Table 5 fails to reject equality between them.

Table 4: Balancing tests - judge characteristics

	Women		Men	
	Estimate	Std.Err.	Estimate	Std.Err.
Dependent variables:				
Age	-0.025	(0.011)	-0.064	(0.055)
Experience	-0.020	(0.015)	-0.093	(0.102)
Exam grade	0.029	(0.033)	0.025	(0.033)
From different region	-0.006	(0.013)	0.045	(0.032)
From northern Italy	0.000	(0.006)	0.028	(0.017)
From central Italy	-0.025	(0.019)	-0.056	(0.049)
From southern Italy	0.024	(0.015)	0.041	(0.040)

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.

Validity of appellate review as a benchmark. Using appeal outcomes to measure decision accuracy requires two conditions: (i) appellate panels do not treat decisions differently based on the gender composition of the trial panel, and (ii) appellate judges are at least as accurate as first-instance judges.[35] Condition (i) would fail if appellate panels confirmed

[35] A similar assumption appears in prior work. Alesina and La Ferrara (2014) use higher-court reversals

Table 5: Observed and simulated distribution of teams

	All-men	2M1F	1M2F	All-women	Total
Observed distribution	72	380	473	209	1134
Simulated distribution	89	348	488	209	1134
Total	161	728	961	418	2268
Pearson $\chi^2(3) = 3.4358$, 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.

decisions more often when their own gender composition matches that of the trial panel, so that upheld decisions reflect compositional affinity rather than underlying quality. We test for this by interacting indicators for each trial-panel type with an indicator for whether the appellate panel's gender composition matches. Supplemental Appendix Table A3 shows that a matching gender mix does not affect the probability that a conviction withstands appeal.[36] Condition (ii) would fail if appellate judges had lower diagnostic skill than trial judges. We note that our strategy does not require appellate courts to be strictly more accurate: even comparable accuracy, combined with a second independent review, makes an affirmed decision more likely to be correct than a reversed one. Supplemental Appendix Table A5 shows that appellate judges in our sample have greater experience and performed better on the entry exam than first-instance judges, making it unlikely that they are systematically less accurate.

Correlated effects. Random assignment ensures that our estimates are causal, but a remaining concern is whether they capture the effect of gender composition itself or of other team characteristics that mechanically co-vary with it. Because male and female judges differ systematically in age and experience (Section 4.1), these attributes may partly account for differences across teams with different gender compositions (Manski, 1993). We address this concern in two ways. First, in equation (1) we control for each judge's characteristics, including age and experience.[37] Second, we re-estimate equation (1) replacing gender composition with alternative measures of team diversity based on the characteristics that differ significantly by gender. As shown in Section 6.4, none of these alternative dimensions of diversity is systematically associated with team outcomes, suggesting that these correlated

as a benchmark of accuracy in their study of racial bias in capital sentencing, a setting where appeals are mandatory; Coviello et al. (2015) and Coviello et al. (2019) instead use the probability of appeal as a quality indicator in Italian labor courts. Our appeal-proof measure combines both margins: a ruling counts as appeal-proof only if it is not appealed or, if appealed, upheld on substantive review. This is more informative than appeal incidence alone, as parties may appeal for reasons unrelated to ruling quality, so the substantive-review outcome filters out appeals without merit.

[36] Interpreting appeal outcomes as a measure of decision accuracy would also be invalid if appellate panels were systematically biased in favor of a particular trial-panel composition, regardless of their own composition. To generate our results, however, such bias would need to run *in favor* of women. This runs counter to previous evidence from several professional settings, *including the judiciary*, which documents gender bias against women rather than in their favor (e.g., Benson et al., 2026; Ash et al., 2024; Sarsons, 2024; Egan et al., 2022; Boring, 2017; Goldin and Rouse, 2000).

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

team characteristics cannot account for the effects we attribute to gender composition.

6 Results

This section presents our baseline results. The next two subsections discuss them in turn and report the full estimates in Tables 6 and 7. Figure 4 provides a graphical summary of the main results.

6.1 Type of decisions

We first ask whether judicial teams reach different decisions in comparable cases depending on their gender composition. Column (1) of Table 6 reports estimates from Equation (1). The estimates are also visually summarized in panel (a) of Figure 4. All-women teams are the most likely to convict. Relative to them, all-men teams are 7.1 percentage points (14.6%) less likely to issue a conviction, and mixed-gender teams are 5.7-9.3 percentage points (11.8-19.2%) less likely to convict. In contrast, conviction rates among all-men and mixed-gender teams are statistically indistinguishable, with no meaningful variation within mixed-gender teams depending on whether they include one or two women.[38]

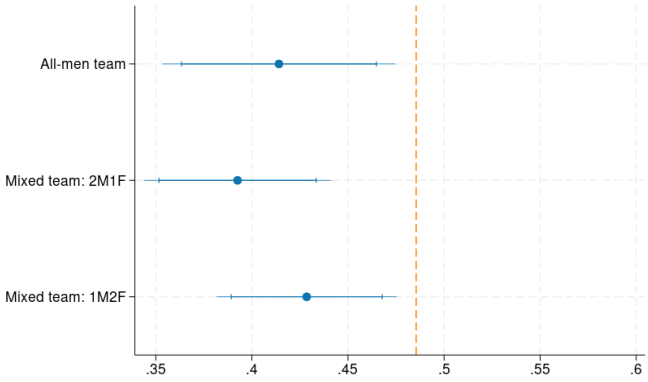
We next examine whether team gender composition affects sentencing decisions, focusing on the length of prison sentences. Column (2) of Table 6 reports unconditional estimates, which combine differences in both the probability of conviction and sentence length conditional on conviction. All-women teams impose sentences that are, on average, 3-4 months longer than those imposed by mixed-gender teams, with no significant difference relative to all-men teams. Column (3) isolates the intensive margin by conditioning on conviction and shows that all-women teams impose prison sentences that are about 12 months shorter than all-men teams and 6 months shorter than mixed-gender teams.[39] As with conviction decisions, sentencing severity does not differ significantly between all-men and mixed-gender teams, nor within mixed-gender teams depending on whether they include one or two women.

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 does not distinguish among the channels discussed in Section 3. Higher conviction rates by all-women teams could reflect lower leniency: if they apply a lower conviction threshold, they would bring in more marginal cases where the evidence of guilt is weaker (Hypothesis 1). Alternatively, higher conviction rates could reflect greater diagnostic skill: if all-women teams are better able to identify guilty defendants, they may convict more often by reducing wrongful acquittals rather than by lowering the evidentiary standard (Hypothesis 2). Similarly, the lower conviction rates of mixed-gender teams relative to all-women teams could reflect either differences in leniency, differences in diagnostic skill, or coordina-

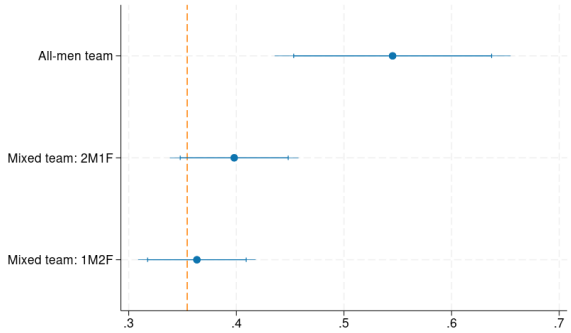
[38] The difference between teams with one woman and two women is significant only at the 10% level and small relative to the difference with all-women teams.

[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).

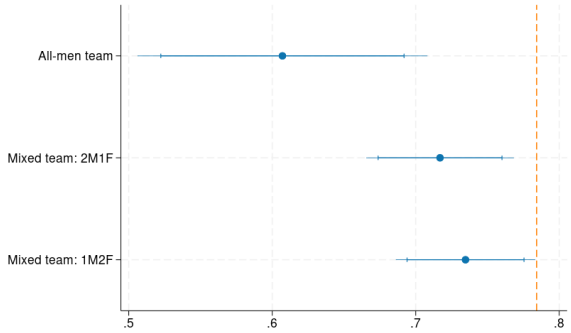
Figure 4: Effect of team gender composition on conviction rate and conviction accuracy



(a) Conviction rate



(b) Probability of appeal (cond.)



(c) Appeal-proof conviction rate (cond.)

Notes. The figure summarizes the estimated effect of team gender composition on three outcomes. In panel (a), the dependent variable is an indicator for whether the team convicts the defendant. In panel (b), the dependent variable is an indicator for whether the decision is appealed against. In panel (c), 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 panel (a), the analysis is done on the whole sample; in panels (b) and (c), the analysis is restricted to conviction decisions only. In each panel, the blue dots equal the raw mean for all-women teams plus the regression coefficient for that team type (*All-men team*, *2M1F team*, *1M2F team*; the omitted category is all-women teams). The orange dashed line marks the raw mean for all-women teams. All regressions include court and crime type-by-year fixed effects, 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. The figure also reports the 90% and 95% confidence intervals based on standard errors clustered at the team level.

tion frictions that make unanimity harder to reach (Hypothesis 3). We distinguish among these channels in the next subsection by leveraging the appellate structure of the Italian judiciary, using appeal outcomes as proxies for the accuracy of first-instance convictions.

6.2 Accuracy of decisions

To study how a panel's gender composition affects decision accuracy, we proceed in two steps. First, we examine whether gender composition affects the *perceived* accuracy of rulings, measured by the probability that a decision is appealed. The maintained assumption is that decisions perceived as less accurate are more likely to be challenged. As discussed in Section 2, although both defendants and prosecutors can in principle appeal any decision, appeals are overwhelmingly concentrated among convictions.[40] Second, because appeals may be

[40]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 (Supplemental Appendix Table A4).

Table 6: 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. for All-women team	0.485	2.534	5.220

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.

partly strategic and may not perfectly track ruling accuracy, we examine appeal outcomes directly and define a ruling as *appeal-proof* if it is either not appealed, or if it is appealed and the appellate court confirms both the verdict and, in the case of conviction, the length of the prison sentence.[41]

Two features of this measure are important to highlight. First, because appeal-proof status is defined for all rulings – not only those that are appealed – , it does not suffer from the selection bias that would arise from conditioning on the endogenous decision to appeal. Second, because appeals are almost entirely concentrated among convictions, our measure of accuracy is one-sided: it captures whether convictions withstand appellate review, but does not directly assess the accuracy of acquittals. This limitation, however, does not prevent us from using appeal outcomes to capture teams' overall accuracy. Because cases are randomly assigned to judicial teams, different team compositions face, on average, comparable distributions of guilty and innocent defendants. As a result, systematic differences in conviction rates and appeal-proof conviction rates cannot be attributed to differences in the underlying case pool. Instead, they reflect differences in how teams evaluate cases and translate those evaluations into decisions.

Perceived accuracy of decisions. Panel A of Table 7 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 (46.1%) more likely to be

[41] According to this definition, 87% of rulings, and 71% of convictions are considered "appeal-proof". Supplemental Appendix Table A6 shows that our results are robust to alternative definitions that focus only on the verdict or allow for some variation in sentence length between the first-instance and appellate decisions. Specifically, the results are unchanged when we classify a decision as appeal-proof if (i) it is unappealed or, if appealed, the verdict is confirmed; or (ii) it is unappealed or, if appealed, the verdict is confirmed and the absolute percentage difference in prison time is below the sample median (7.6%).

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 convictions 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 of convictions. Consistent with this, once we restrict the sample to convictions, the estimated gap between all-men and all-women teams grows to 19.1 percentage points (54%). Mixed-gender teams display a distinct pattern: despite convicting at similar rates to all-men teams, their convictions face appeal rates as low as those of all-women teams (see Panel (b) of Figure 4).

Finally, since all-men teams are more likely to impose longer prison sentences upon conviction (see Table 6), one might be concerned that their higher appeal rates simply reflect harsher punishment at the intensive margin. Column (3) addresses this by controlling for sentence length. The results are virtually unchanged, suggesting that the higher appeal rate of convictions issued by all-men teams is not driven by sentence severity.

Appeal-proof decisions. As appeal decisions are made by the parties and may reflect strategic considerations in addition to the underlying accuracy of the ruling, in Panel B of Table 7 we turn to appeal-proof decisions, which provide a closer proxy for the effective accuracy of first-instance rulings. The results mirror those found for 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 become more pronounced when we focus on conviction decisions only (column (2) of Table 7 and panel (c) of Figure 4). All-men teams are 17.7 percentage points (22.6%) 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 higher than those of all-men teams, with no statistically significant difference between teams with one versus two women; both mixed-gender configurations fall only 5-6.7 percentage points below all-women teams, who display the highest level of accuracy.

Finally, because all-women teams impose shorter prison sentences conditional on conviction (Table 6), one might worry that their higher appeal-proof rates mechanically reflect more cautious sentencing rather than greater accuracy. If sentences are shorter, appellate courts have less room to reduce them, mechanically raising the share of convictions that are confirmed.[42] Column (3) addresses this concern 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 sentence severity.

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

Table 7: Effect of gender composition of the team on the accuracy of decisions

	(1) Unconditional	(2) Conditional	(3) Controlling for jailtime
A. Probability of appeal			
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. for All-women team	0.178	0.354	0.354
B. Appeal-proof decisions			
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. for All-women team	0.890	0.784	0.784

Notes. In panel A, the dependent variable is an indicator for whether the team's decision is appealed against. In panel B, 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.

6.3 Interpreting the results

Figure 4 summarizes the evidence on conviction rates and decision accuracy. We now use these patterns to evaluate the hypotheses laid out in Section 3. All-women teams convict more often, yet their convictions remain more likely to survive appellate review, even after controlling for sentence length. This pattern is inconsistent with Hypothesis 1: if all-women teams were simply applying a lower conviction threshold, they should be convicting more marginal cases, whose convictions would be less robust on appeal, not more. The pattern is instead consistent with Hypothesis 2: higher diagnostic skill simultaneously raises conviction rates and improves conviction quality.[43]

[43] An alternative interpretation is that all-men teams write less persuasive motivations while being equally accurate in assessing guilt. However, Italian criminal appeals are reviews on the merits, reassessing the factual basis of the decision rather than only the quality of the written motivation. If drafting quality was driving our results, we would expect higher appeal rates but not lower appeal-proof rates. Instead, we find

Mixed-gender teams convict at similar rates to all-men teams, yet their convictions are almost as likely to be appeal-proof as those of all-women teams. This points to two distinct forces at work. Diagnostic skill shapes conviction quality independently of how often teams convict: the presence of at least one woman is sufficient to raise accuracy close to the all-women benchmark. At the same time, coordination frictions suppress conviction rates without lowering the quality of the convictions that do occur, consistent with Hypothesis 3. More broadly, the two outcomes shift at different points in panel composition: conviction rates rise only when the panel is all-women, while quality improves once at least one woman is present. Such a pattern is inconsistent with team behavior being a linear aggregation of individual judges' characteristics, and instead points to behavioral interactions among panel members. We return to these mechanisms in Section 7.2, where we develop a structural model to disentangle them quantitatively.

6.4 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 5.2, a key concern is that our estimates might reflect not gender composition itself but other characteristics correlated with it. We address this in three ways. First, all baseline specifications control for judge characteristics that vary significantly by gender (Section 4.1) – so that differences across team compositions are not driven by these observable dimensions. Second, we re-estimate equation (1) replacing gender composition with team diversity based on the characteristics that differ significantly by gender (namely age and experience).[44] If our main results were driven by these correlated characteristics, we would expect similar coefficients to those reported in Tables 6 and 7 – for instance, since women tend to be younger, all-young teams should resemble all-women teams. Table A7 shows no significant relationship between any of these alternative dimensions and either conviction rates or appeal-proof conviction rates, supporting the conclusion that our results are driven by gender composition rather than by correlated traits. Third, since judges may differ along dimensions we do not observe, controlling for observable characteristics alone may be insufficient to rule out that gender proxies for some unobserved trait correlated with decision-making. To address this concern, we estimate the following specification with judge fixed effects, separately for men and women:

$$Y_{j,isct} = \alpha_0 + \alpha_1 SameGender_{j,isct} + \delta Z_{ict} + \gamma_c + \gamma_{s,t} + \gamma_j + \epsilon_{isct}$$

where $SameGender_{j,isct}$ equals one if judge j sits in a gender-homogeneous team, γ_j denotes judge fixed effects and the remaining notation follows from equation (1). The judge fixed

that their convictions are more likely to be appealed *and* less likely to be appeal-proof, which points to differences in underlying accuracy.

[44] Teams are classified as all-young/all-old/mixed (based on court median age), and all-experienced/all-inexperienced/mixed (based on median years of service).

effects then absorb all time-invariant individual characteristics – observed and unobserved – so that identification comes entirely from within-judge variation as the same judge appears on teams with different gender compositions. Because we estimate this specification separately for men and women, $SameGender_{j,isct}$ identifies an all-men panel for male judges, an all-women panel for female judges, with mixed-gender teams as the omitted group in both subsamples. Table A8 shows results consistent with our main findings: even holding the judge fixed, team composition continues to shift outcomes. Stable unobserved differences between judges therefore cannot account for the gender patterns we document.

Court-specific time trends: Our baseline specification absorbs 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 shifts in local judicial culture or caseload composition. If these shocks were correlated with team gender composition, our baseline estimates could be biased. We address this concern by running two more demanding specifications. The first replaces court fixed effects with court-by-year fixed effects, comparing teams of different compositions within the same court and year. To address the further concern that courts may experience crime-specific shocks within a year, we additionally consider a court-by-year-by-crime-type specification, which restricts the comparison to teams handling the same crime type in the same court and year. Supplemental Appendix Table A9 reports the results and shows that the estimates are virtually unchanged across both specifications.

Different classifications of crimes: Our crime classification follows the 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. Supplemental 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 our results is that men and women hold different attitudes, beliefs, or knowledge regarding specific types of crimes. For instance, we might expect the presence of women on judicial teams to have a stronger impact on "gender-sensitive" offenses – such as violence against women or children – than for more neutral ones like theft.[45] To test this, we sequentially exclude one crime category at a time and re-estimate equation (1) on the remaining cases. Supplemental Appendix Table A11 shows that the estimated coefficients remain stable in sign and magnitude across all specifications, indicating that no single crime category is driving our results.[46]

Gender and nationality of the defendant. An alternative explanation is that our results

[45] Violence against women or children and theft fall under "crimes against the person" and "crimes against property", respectively.

[46] Results are similar when we estimate separate regressions by crime type, indicating that the effect is robust across all case types. We report the leave-one-out specification as it is more precisely estimated.

reflect homophily between judges and defendants – for instance, if judges treat same-gender defendants differently, or if judges' responses to defendants' nationality vary by judge gender. We test this by re-estimating equation (1) separately for male and female defendants, and for Italian and foreign defendants. Supplemental Appendix Table A12 shows that the patterns documented in the previous sections hold across all subsamples, ruling out homophily-based explanations.

7 Mechanisms: Effort, Skill, and Team Dynamics

In this section, we ask what generates the performance gaps documented above. Establishing that teams differ in diagnostic skill raises a natural question: do these differences reflect fixed ability or endogenous effort? The answer matters for policy. If the gaps reflect fixed ability – traits such as cognitive skill or legal judgment – then each judge's contribution to team performance is largely predetermined, and improving team outcomes would require changing who sits on the panel. If instead the gaps reflect effort – the time and care judges devote to preparing and evaluating cases – then team performance is more malleable, and institutional design could improve outcomes by shaping incentives, accountability, or other features of the work environment. Separately, we revisit the distinctive pattern of mixed-gender teams – convicting less but with similar accuracy to all-women teams – which previous Sections attributed to coordination frictions, and present further evidence consistent with that interpretation. We first present reduced-form evidence consistent with two mechanisms: that the accuracy gap reflects effort rather than fixed ability, and that coordination frictions shape the conviction-rate gap of mixed-gender teams. Motivated by these results, we develop and estimate a structural model of team decision-making that microfounds the primitives introduced in Section 3. The model incorporates two channels: heterogeneity in diagnostic skill, modeled through endogenous effort provision, and within-team frictions arising from diversity. It allows us to decompose the contribution of each channel to different dimensions of team outcomes and to evaluate how the growing presence of women in the judiciary is likely to affect aggregate court performance.

7.1 Is diagnostic ability fixed or effort-driven?

Higher diagnostic accuracy could reflect stable differences in skill, differences in the time and care judges devote to evaluating cases, or an interaction between the two, if ability lowers the cost of effort or increases its returns. We use the richness of our data to assess which explanation better fits the empirical patterns. Below, we present three pieces of evidence that point to effort and are difficult to reconcile with fixed differences in ability alone.[47]

[47]Evidence from several real-world settings, including medicine, academia, and corporate boards, suggests that women often exert greater effort and work more diligently than men on comparable tasks (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. They may reflect differences in intrinsic motivation or risk aversion, as well as the fact that women face harsher penalties for identical errors and receive less institutional support despite similar performance (Gagliarducci and Paserman, 2012; Egan et al., 2022; Sarsons, 2024;

Entry examination results. First, we examine whether our findings could be driven by gender differences in selection into the judiciary. Because 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 assess this, we collected and digitized the official rankings from all national entry examinations for the judiciary held between 1965 and 2022. If women were more positively selected than men, we would expect them to perform better on the entry exam. Figure 1 provides no support for this hypothesis: men's and women's average grades closely track each other across entry cohorts, following a common upward trend with largely overlapping confidence intervals. In the years where the difference is statistically significant, men score on average 1.8 points higher than women – a small gap relative to mean scores, and in the opposite direction from what positive selection of women would predict. While women may still be 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 when evaluating candidates 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.[48] Column (1) of Table 8 reports the results. All-men teams complete cases substantially faster than any other composition: their trials last, on average, eight months (28%) less than those of all-women teams and nearly six months less than those of mixed-gender teams. The difference between all-women and mixed-gender teams, by contrast, is not statistically significant. Two considerations emerge from these results. First, all-women teams achieve better outcomes while also devoting more time to each case. If the accuracy gap reflected only fixed differences in ability, better outcomes need not be accompanied by systematically longer trials. The fact that higher accuracy coincides with longer proceedings is therefore more consistent with greater effort. Second, all-women and mixed-gender teams take equally long to reach decisions despite showing different conviction rates. This is consistent with coordination frictions in mixed teams (Hypothesis 3): part of their deliberation time is absorbed by difficulties in reaching consensus, reducing conviction rates without affecting the quality of convictions that do occur.

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.[49] While they do not create binding precedent, these texts are subject to

Dupas et al., 2026), which could raise the perceived cost of mistakes.

[48]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.

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

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

		Writing quality		
	(1)	(2)	(3)	(4)
	Length of trial (days)	Length (words)	Clarity index	N. maxims
All-men team	-249.393***	25937.369***	-1.236***	8.256***
	(88.067)	(6182.655)	(0.287)	(2.396)
2M1F team	-47.420	-341.315	-0.021	0.836
	(62.092)	(2805.348)	(0.155)	(0.810)
1M2F team	-106.677*	-775.294	-0.012	0.496
	(57.070)	(2252.154)	(0.132)	(0.699)
Observations	17042	17042	17042	17042
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var. for All-women team	888.291	10171.990	0.131	3.481

Notes. The dependent variable in column (1) is the length of the trial (in days); in column (2) is length (in words) of the written opinion; in column (3) 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 (4), 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 trial duration and 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 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.

appellate scrutiny and must clearly set out the logic and evidentiary evaluation underlying the decision. We examine 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).[50] 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.[51] Finally, our third measure of drafting quality captures the frequency of citation of Supreme Court of Cassation 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 is explicitly criticized by the Italian Ministry of Justice as weakening judicial reasoning and making arguments appear mechanically constructed, rather than reflecting a thoughtful and case-specific analysis (Ministero della Giustizia, 2016).[52] Results are reported in columns (2)-(4) of Table 8.[53] All-men teams produce judgments that are more than twice as long as those

details.

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

[51]The index combines average sentence length, average word length, average root-word length, and the frequency of gerunds, demonstrative adjectives, modal verbs, and contingency clauses; higher values indicate longer, more convoluted prose and thus less clear drafting (see Giommoni et al., 2024). The index is standardized to mean 0 and standard deviation 1.

[52]The Italian system is not based on common law: judges' reasoning is grounded in legal codes, with occasional reference to court precedents (see Supplemental Appendix B).

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

of other teams and cite Supreme Court of Cassation maxims nearly three times as often as mixed-gender or all-women teams. Their written reasoning 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.

7.2 A model of team diagnostic decision-making

The evidence presented above points to effort as a key driver of cross-team performance gaps. Standard models of judicial decision-making, which 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 these patterns: they imply that teams who convict more should have weaker convictions – the opposite of what we find. Building on the two-stage framework introduced in Section 3, we propose a model that extends single-agent models of diagnostic decision-making (e.g. Mungan, 2024, 2026; Chan et al., 2022) to a team setting. The extension departs from the single-agent benchmark in two ways. First, diagnostic skill is heterogeneous across decision-makers and endogenous to team composition through strategic effort provision: a judge's effective skill depends not only on her own characteristics but on who else sits on the panel. Second, the team setting introduces coordination frictions at the decision stage, which can prevent precise individual signals from translating into a unanimous conviction.[54]

7.2.1 Model

States and prior. Consider a team of two judges $i \in \{1, 2\}$ making a binary decision under uncertainty.[55] 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

[54]Two considerations inform our modeling choice. First, modeling individual contribution as an endogenous, costly choice is the standard primitive in two related literatures: the committee-design literature, in which members acquire costly information that determines the precision of the panel's signal and the quality of collective decisions (e.g., Mukhopadhaya, 2003; Persico, 2004; Gerardi and Yariv, 2008); and the team-production literature, which establishes that effort within teams is an endogenous strategic choice and responds to peer composition (e.g., Holmstrom, 1982; Itoh, 1991; Kandel and Lazear, 1992; Mas and Moretti, 2009; Bandiera et al., 2010). Second, our empirical patterns are difficult to rationalize with a pure fixed-ability model. Under standard aggregation rules – average, median, or majority – mixed-gender teams should fall between all-men and all-women teams on every outcome. Instead, they resemble all-men teams in conviction rates and all-women teams in appeal-proof rates. Matching this pattern with fixed-ability would require one aggregation rule for conviction – with men being decisive at this margin – and a different one for accuracy – with women being decisive – even though both outcomes are produced by the same panel, in the same case, through a single decision process, which is difficult to motivate institutionally. Endogenous effort instead generates these outcome-specific effects directly.

[55]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.

$c(e_i)$ (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 *common* information environment, captured by $\theta(e_i, e_j) \in \left(\frac{1}{2}, 1\right)$, increasing and concave in each judge's effort and exhibiting decreasing returns across judges ($\theta_{e_i e_j} < 0$).[56] Based on this shared evidentiary record, each judge independently forms a privately observed 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.[57] 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 $2\theta(1-\theta)$.[58]

Disutility from mistakes. Judicial teams can make mistakes. Let $d \in \{g, i\}$ denote the team's decision (g = convict, i = acquit). Despite the decision being a team one, each judge can differ in the disutility they experience from mistakes. A wrong conviction generates a disutility for judge i of $\phi_i \beta_i$; a wrong acquittal generates a disutility of ϕ_i; correct decisions yield no disutility:[59]

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

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 criminal law regards convicting an innocent person as more costly than acquitting a guilty one (*Blackstone Principle*), we assume $\beta_i > 1$.[60] 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).[61]

Stage 2: Voting and deliberation. The model is solved by backward induction. After observing their individual signals, each judge forms a posterior belief about the defendant's guilt

[56] 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.

[57] 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.

[58] This 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 different assessments.

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

[60] In criminal law, *Blackstone's ratio* – articulated by the English jurist William Blackstone in his *Commentaries on the Laws of England* (1765–69) – holds that "it is better that ten guilty persons escape than that one innocent suffer". The principle reflects an asymmetric concern about convicting the innocent and informs the modern standard of *guilt beyond a reasonable doubt*, under which any reasonable doubt about the defendant's guilt requires acquittal.

[61] Women have been shown to face harsher penalties for identical errors (Egan et al., 2022; Sarsons, 2024; Dupas et al., 2026), which could raise their perceived cost of mistakes.

and votes accordingly.[62] Given signal X_i and precision θ, judge i's posterior is

$$\Pr(G \mid X_i, \theta) = \begin{cases} \dfrac{p\,\theta}{p\,\theta + (1-p)(1-\theta)} & \text{if } X_i = G \\ \dfrac{p\,(1-\theta)}{p\,(1-\theta) + (1-p)\,\theta} & \text{if } X_i = I \end{cases}$$

Judge i votes to convict whenever 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 judge: $\beta/(1+\beta)$, where $\beta = \max(\beta_i, \beta_j)$. The Bayesian voting rule therefore implies two conditions for conviction, depending on the signal received. After a guilty signal, the posterior exceeds the threshold if and only if precision is high enough for the evidence to be informative:

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

After an innocent signal, the posterior can still exceed the threshold if precision is so low that the signal is essentially uninformative and the prior probability of guilt dominates; this occurs if and only if

$$\theta \;\leq\; \frac{p}{p + \beta(1-p)} = 1 - \bar{\tau} \tag{3}$$

In what follows, we focus on the regime in which $\theta > 1 - \bar{\tau}$, so that only guilty signals can trigger conviction.[63] In this regime, three outcomes are possible. First, if both signals indicate guilt ($X_i = X_j = G$) and condition (2) holds, the panel convicts. Second, if both signals indicate innocence ($X_i = X_j = I$), or both indicate guilt but the evidence threshold in (2) is not met, 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 agreement: one judge is selected at random and their assessment prevails. With probability $1-q$, they fail to reach agreement, and the panel acquits by default.[64] 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 function

[62] We assume sincere voting. This is natural in our institutional setting, where deliberation is private, votes are not individually recorded, and judges face no incentive to deviate from their true assessment.

[63] We verify in Section 7.2.2 that, at the estimated parameters, this condition holds for all team types. Note that since $\theta > 1/2$ by construction, the only alternative regime is one in which both conditions (2) and (3) hold simultaneously, so each judge votes to convict regardless of the signal received. This scenario is inconsistent with the principles of the Italian criminal justice system, where defendants are presumed innocent and conviction must rest on evidence proving guilt beyond a reasonable doubt, not on priors alone. Nevertheless, we present derivations under such regime ($\theta \leq 1 - \bar{\tau}$) in Supplemental Appendix C for completeness.

[64] 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.

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

$$\theta_{e_i}(e_i,e_j) = \frac{c'(e_i)}{\phi_i\left[p+\beta_i(1-p)\right]} \tag{4}$$

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. Denote by $e_i^*(e_j)$ judge i's optimal effort given partner effort e_j. An equilibrium is a fixed point of the joint map $(e_1^*(e_2), e_2^*(e_1))$; jointly solving the first-order conditions for both judges pins down equilibrium efforts $e_i^* = f(\phi_i, \phi_j, \beta_i, \beta_j, p)$.[65]

We obtain the following comparative statics.[66] 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 precision. 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 preference parameters ϕ and β.

Conviction rates (C). When condition (2) holds, the unconditional conviction rate is:[67]

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

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

[65] This first-order condition characterizes an interior solution in which equilibrium precision satisfies $\theta > 1-\bar{\tau}$, so that only guilty signals trigger conviction. As already discussed, in the alternative regime ($\theta \leq 1-\bar{\tau}$), judges convict after any signal realization and the marginal return to effort is zero, yielding a corner solution at $e_i = 0$. In Section 7.2.2, we verify that at the estimated parameters the equilibrium lies in the interior region for all team types. The interior equilibrium is unique; see Supplemental Appendix C for a proof.

[66] Proofs easily follow by differentiating equation (4) and applying the Implicit Function Theorem. Full derivations are in Supplemental Appendix C.

[67] When condition (2) fails, no signal clears the conviction threshold and $C = 0$.

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:

$$C = \left[p\theta^2 + (1-p)(1-\theta)^2\right] + q\,\theta(1-\theta) \tag{5}$$

Higher signal precision θ raises conviction rates when the prior probability of guilt is sufficiently high. Sharper signals act on C through three channels: more correct convictions of guilty defendants (pushing C up), fewer wrongful convictions of innocent ones (pushing C down), and fewer convictions reached through deliberation (also pushing C down), since disagreements become rarer. The first effect dominates when most defendants at trial are in fact guilty: the gain in correct convictions then exceeds the offsetting falls in wrongful convictions and in deliberation outcomes. Formally, $\partial C/\partial\theta = 2(\theta + p - 1) + q(1-2\theta) > 2p - 1 > 0$ when $p > 1/2$. Since cases reach a collegial panel only after a preliminary hearing has established sufficient grounds for trial, $p > 1/2$ is plausible in our setting, so 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 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 (5), $\partial C/\partial q = \theta(1-\theta) > 0$. 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.[68] That is,

$$APC = \Pr(S = G \mid d = g) = \frac{p\,\theta\left[\theta + q(1-\theta)\right]}{C} \tag{6}$$

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 $pq\theta(1-\theta)$).

[68] 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 5, 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 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 model's primitives map into cross-team comparisons of conviction rates and appeal-proof conviction rates.

(i) *Diligence*: teams with more diligent members (higher ϕ) exert more effort and thus achieve higher diagnostic skill θ, leading to both higher conviction rates C and higher appeal-proof conviction rates APC than less diligent teams. In teams with heterogeneous ϕ, efforts are strategic substitutes: the more diligent judge raises effort, while the less diligent one free-rides by lowering it. Whether C and APC are higher than in a homogeneous team depends on which response dominates.

(ii) *Leniency*: teams with more lenient members (higher β) exert more effort, raising θ, but also apply a tighter conviction threshold. Hence the net effect on C is ambiguous compared to a less lenient team, and will depend on the values of parameters. APC is instead unambiguously higher. In teams with heterogeneous β, the sign of the difference in both C and APC depends on whether the more lenient judge's additional effort dominates the teammate's free-riding.

(iii) *Frictions*: teams that face greater coordination frictions (lower q) have lower C but higher APC. Importantly, the effect on C exceeds the effect on APC: frictions operate only at the decision stage and leave effort and signal precision unchanged.[69]

7.2.2 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

[69] From the derivatives, $|\partial APC/\partial q|/|\partial C/\partial q| = (APC - p)/C$. Since $APC - p < 1 - p$ and $C > p\theta^2 + (1-p)(1-\theta)^2$, this ratio is well below one for reasonable parameter values.

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).[70]

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 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.[71] 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.[72]

Table 9 reports the results. Panel A displays the estimated parameters. Women are substantially more diligent than men (ϕ_f is estimated to be 6.6 times as large as ϕ_m), while leniency does not differ significantly across gender. 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.[73] Panel B shows that the model fits the data well, with a mean absolute percentage error of 1.55% 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 exert less effort than they would in an all-men team, while women exert more effort than they would in an all-women team. Diversity thus affects the two genders asymmetrically, reducing men's effort and increasing women's. Mixed-gender teams' total effort falls between that of all-men and all-women teams, but because of women's over-compensation, it 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

[70] To mirror the model structure, the targeted moments are obtained from a regression that pools together 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 Supplemental Appendix Table A13. The findings and their interpretation are qualitatively unchanged relative to those in Tables 6 and 7.

[71] 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.

[72] 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.

[73] At the estimated parameters, equilibrium precision satisfies $\theta > 1 - \bar{\tau}$ for all team types, confirming that the equilibrium lies in the interior regime where only guilty signals trigger conviction.

Table 9: Structural parameter estimate and model fit

Panel A: Parameter Estimates			
Parameter		Estimate	
β_m (Men's leniency)		1.345	
		(0.046)	
β_f (Women's leniency)		1.481	
		(0.075)	
ϕ_f (Women's diligence, rel. to men)		6.583	
		(0.304)	
p (Prior probability of guilt)		0.556	
		(0.007)	
q_0 (Agreement prob., same-gender teams)		0.726	
		(0.014)	
q_{mix} (Agreement prob., mixed-gender teams)		0.563	
		(0.015)	
Panel B: Model Fit			
Moment	Target	Simulated	Error
Conviction rates			
All-women teams	0.484	0.463	−0.021
All-men teams	0.418	0.433	0.016
Mixed-gender teams	0.412	0.412	0.000
Appeal-proof conviction rates			
All-women teams	0.767	0.766	−0.000
All-men teams	0.592	0.598	0.006
Mixed-gender teams	0.710	0.712	0.001
Mean absolute percentage error			1.55%
Panel C: Implied Equilibrium Effort			
Team type	Judge 1	Judge 2	Total
All-women (F–F)	0.841	0.841	1.682
All-men (M–M)	0.143	0.143	0.287
Mixed-gender (M–F)	0.134	0.931	1.066

Notes. The table reports results from the structural estimation described in Section 7.2.2. 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.

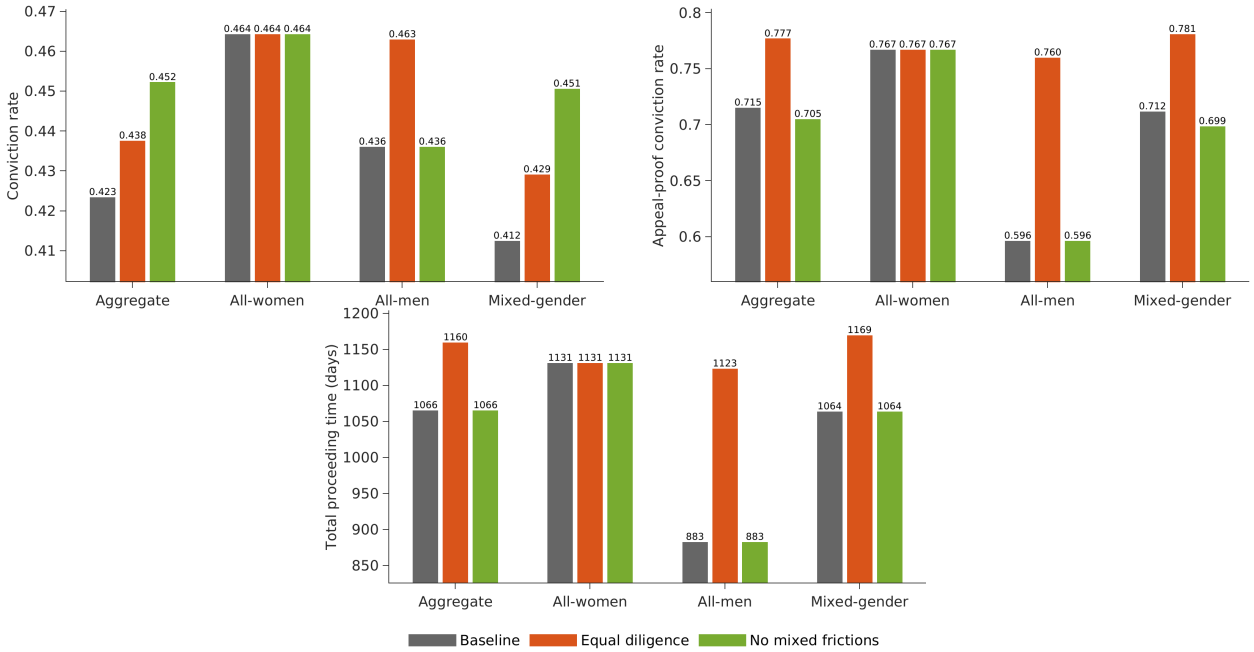
panels, women's over-compensation largely offsets men's free-riding, so total precision – and thus APC – nearly matches that of all-women teams. However, their lower q makes it harder to convert disagreements into convictions, pulling C close to that of all-men teams. Since q operates only at the decision stage and leaves precision unchanged, lower agreement in mixed-gender teams primarily acts by reducing 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.

7.2.3 Counterfactual Analysis

The estimation results show that women are substantially more diligent than men, and that mixed-gender teams face greater coordination frictions. This raises two natural questions. First, how much does each channel contribute to the observed performance gaps across teams? Second, as the share of women in the judiciary continues to grow, what are the

implications for aggregate court performance?[74] We address these questions through two sets of counterfactual exercises. First, we vary one structural parameter at a time, holding the others fixed and re-solving for equilibrium effort, to identify the main drivers of the gaps in team performance. Second, we vary the share of women in the judiciary and recompute aggregate outcomes under each implied gender composition. For each counterfactual scenario, we estimate conviction rates, appeal-proof conviction rates and total proceeding time, defined as the number of days between the assignment of a case to a team and its final resolution (the first-instance verdict for cases that are not appealed and the appeal ruling for cases that are).

Figure 5: Counterfactual analysis: equalizing structural parameters



Notes. The figure shows the results of three counterfactual exercises. The top panel reports conviction rates; the middle panel, appeal-proof conviction rates; and the bottom panel, total proceeding time, defined as the number of days between the assignment of a case to a team and its final resolution (the first-instance verdict for cases that are not appealed; the appeal ruling for cases that are). In each exercise, one parameter is set to its counterfactual value while all others remain at their estimates, and equilibrium effort is re-solved. Team composition is held at current shares (18.4% AW, 5.9% AM, 75.7% MG). Baseline team-level proceeding times are regression-adjusted data means; counterfactual times are obtained by adding the estimated marginal effect of effort on trial duration to the data means.

Parameter counterfactuals. Figure 5 reports the results of the parameter counterfactuals, where each structural parameter is varied in turn 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 6.2 percentage points, driven mostly by all-men teams, whose conviction accuracy rises by 16.4 percentage points. Notably, the appeal-proof conviction rate of mixed-gender teams – which increases by 6.9 percentage points – slightly exceeds that of all-women teams under equal diligence, because

[74] As discussed in Section 2, women now represent the majority of new entrants into the profession, a trend that is progressively shifting the gender composition of collegial panels. As mentioned in Section 1, comparable shifts are underway in many other developed countries, making the counterfactuals below relevant beyond Italy.

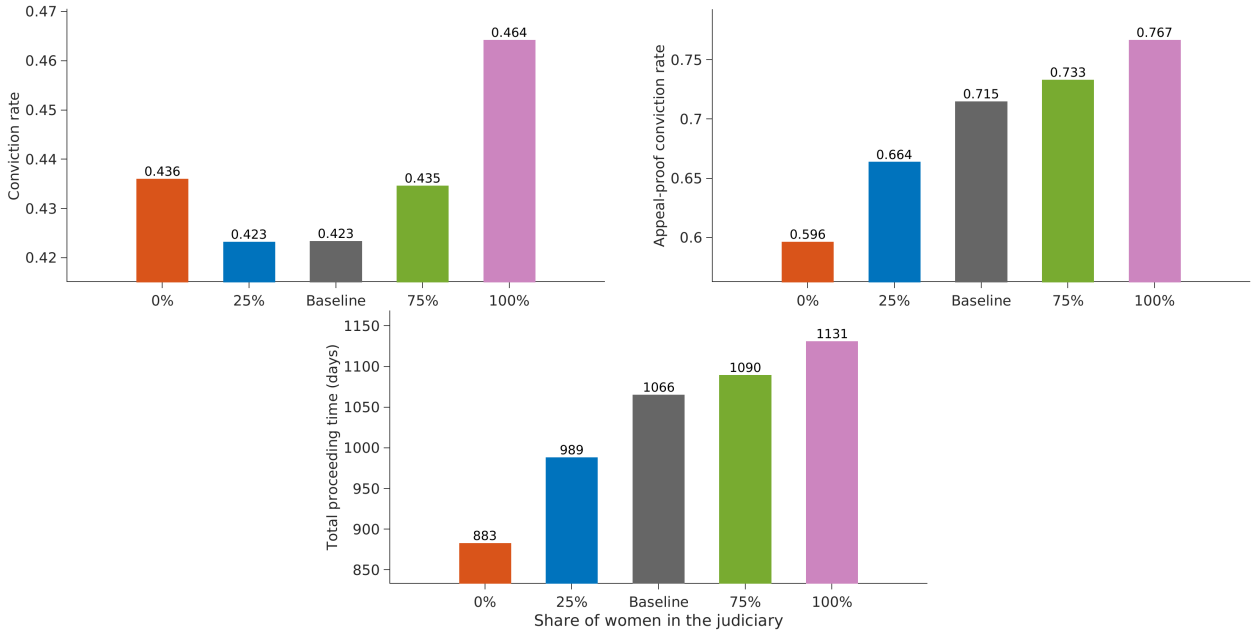
greater coordination frictions in mixed panels filter out marginally lower-quality cases from the conviction pool due to disagreement. Conviction rates also rise modestly (+1.5 pp) as a consequence of higher effort. This exercise highlights the central role of the diligence gap: if men were as diligent 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 94 days.

No mixed-gender frictions ($q_{mix} = q_0$). Removing coordination frictions in mixed-gender teams largely closes their conviction-rate gap with all-women teams, raising the aggregate conviction rate by 2.9 percentage points, with only a small reduction in aggregate *APC* (−1.0 pp). This asymmetry arises because frictions operate only at the deliberation stage and do not affect effort or precision: lower q reduces how often disagreement cases result in conviction, but since these convictions are of lower average accuracy, filtering them out slightly raises the accuracy of the remaining convictions. Consistent with no change in effort, trial length is unchanged.

Composition counterfactuals. Figure 6 examines how aggregate outcomes change as women represent a larger share of the judiciary, holding estimated parameters fixed.[75] We vary the share of women in the judge pool from 0% to 100% and, for each share, compute the probability that a randomly formed three-judge panel is all-women, all-men, or mixed-gender. Aggregate outcomes are then obtained by weighting the model-predicted outcomes for each team type by these probabilities. The results show a clear trade-off: as the share of women rises, decision quality increases monotonically, but so does the time required to clear cases. Relative to the current composition of the judiciary (56% women), a shift to 75% women raises *APC* by 1.8 percentage points and total proceeding time by 24 days; an all-women judiciary raises them by 5.2 percentage points and 65 days, respectively. These gains reflect the higher diligence of women, which translates into more precise signals and thus more accurate convictions, at the cost of more time spent on each case. Reducing the share of women has the opposite effect: at 25% women, *APC* falls by 5.1 percentage points and proceeding time shortens by 77 days; at 0%, by 11.9 percentage points and 183 days, respectively. Conviction rates, by contrast, exhibit a U-shaped pattern: they are lowest near the current composition and rise as the judiciary becomes either more male- or more female-dominated. This pattern reflects the fact that mixed-gender teams, which have the lowest conviction rates due to coordination frictions, are most prevalent at intermediate female shares; moving toward either extreme replaces them with same-gender teams, which convict more often. Taken together, these exercises suggest that as the judiciary becomes increasingly female-dominated, aggregate decision accuracy should improve, accompanied by an increase in total proceeding time. This trade-off 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

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

Figure 6: Counterfactual analysis: varying the share of women in the judiciary



Notes. The figure shows how aggregate court outcomes vary with the share of women in the judiciary. The top panel reports conviction rates; the middle panel, appeal-proof conviction rates; and the bottom panel, total proceeding time (defined as in Figure 5). Each bar corresponds to a different share of women in the pool (0%, 25%, 75%, 100%), with the baseline (56%) shown in gray. Team shares are computed combinatorially for three-judge panels. Structural parameters are held at their estimated values. Baseline proceeding time is the regression-adjusted data mean; counterfactual times are obtained by adding the estimated marginal effect of effort on trial duration to the data mean.

will become increasingly relevant as the ongoing feminization of the judiciary continues to reshape the composition of the bench.

8 Conclusions

This paper investigates how the gender composition of teams shapes decision-making and decision quality. The Italian judiciary offers a rare empirical laboratory for studying this question. Cases are decided by collegial panels, and the quasi-random allocation of judges and cases generates plausibly exogenous variation in team gender composition. Moreover, the appeals process allows us to move beyond observed decisions and study their quality, proxied by whether a sentence is either not appealed or, if appealed, confirmed by a higher court. 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 despite trials of similar length, 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.

Our findings have implications for the design of professional teams that evaluate complex information under uncertainty and must reach a collective decision, a common feature of many expert organizations. First, accuracy and speed are linked through effort: greater effort raises accuracy but lengthens trials. Policies that act on effort therefore face an accuracy–speed trade-off, which our counterfactuals show to be sizable. Second, to the extent that quality gaps across teams are driven by effort rather than fixed individual traits, organizational levers – workload management, accountability mechanisms, and incentive structures that reward quality alongside speed – can help narrow them without changing team composition. Third, calls for greater diversity should account for coordination frictions, which 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.

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

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

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., Chen, D. L., and Ornaghi, A. (2024). Gender attitudes in the judiciary: Evidence from U.S. 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.

Bandiera, O., Barankay, I., and Rasul, I. (2010). Social incentives in the workplace. The Review of Economic Studies, 77(2):417–458.

Benson, A., Li, D., and Shue, K. (2026). "Potential" and the gender promotion gap. American Economic Review, 116(2):375–417.

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.

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.

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

Calder-Wang, S., Gompers, P. A., and Huang, K. (2021). Diversity and performance in entrepreneurial teams. NBER Working Paper No. 28684.

Chan, D. C. (2016). Teamwork and moral hazard: Evidence from the emergency department. Journal of Political Economy, 124(3):734–770.

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.

Chu, B., Handel, B., Kolstad, J., and Malmendier, U. (2022). Gender differences in non-promotable tasks: The case of clinical note-taking. Working paper.

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

Coviello, D., Ichino, A., and Persico, N. (2015). The inefficiency of worker time use. Journal of the European Economic Association, 13(5):906–947.

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.

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., and Collective, S. D. (2026). Gender differences in economics seminars. American Economic Review, 116(2):749–789.

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.

Friebel, G., Heinz, M., Krueger, M., and Zubanov, N. (2017). Team incentives and performance: Evidence from a retail chain. American Economic Review, 107(8):2168–2203.

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.

Gallen, Y. (2024). Motherhood and the gender productivity gap. Journal of the European Economic Association, 22(3):1055–1096.

Gerardi, D. and Yariv, L. (2008). Information acquisition in committees. Games and Economic Behavior, 62(2):436–459.

Giommoni, T., Guiso, L., Michelacci, C., and Morelli, M. (2024). The economic cost of ambiguous laws. CESifo Working Paper No. 11929.

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

Gupta, A., Kalmenovitz, J., Ranjan, R., and Xiao, K. (2025). When diversity rules. Available at SSRN 5389250.

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. (2025). Who does the talking here? the impact of gender composition on team interactions. Management Science, 71(5):4131–4152.

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.

Holmstrom, B. (1982). Moral hazard in teams. The Bell Journal of Economics, 13(2):324–340.

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.

Itoh, H. (1991). Incentives to help in multi-agent situations. Econometrica, 59(3):611–636.

Kandel, E. and Lazear, E. P. (1992). Peer pressure and partnerships. Journal of Political Economy, 100(4):801–817.

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.

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. (2026). Gender mix and team performance: Evidence from obstetrics. NBER Working Paper No. 35084.

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

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.

Mas, A. and Moretti, E. (2009). Peers at work. American Economic Review, 99(1):112–145.

Matsa, D. A. and Miller, A. R. (2013). A female style in corporate leadership? Evidence from quotas. American Economic Journal: Applied Economics, 5(3):136–169.

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.

Mukhopadhaya, K. (2003). Jury size and the free rider problem. Journal of Law, Economics, and Organization, 19(1):24–44.

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

Mungan, M. C. (2026). Symmetry, presumptions, and the judges design. PLoS One, 21(1):e0340446.

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.

Persico, N. (2004). Committee design with endogenous information. The Review of Economic Studies, 71(1):165–191.

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. (2026). Quota vs quality? Long-term gains from an unusual gender quota. American Economic Review, Forthcoming.

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

Supplemental Appendix to

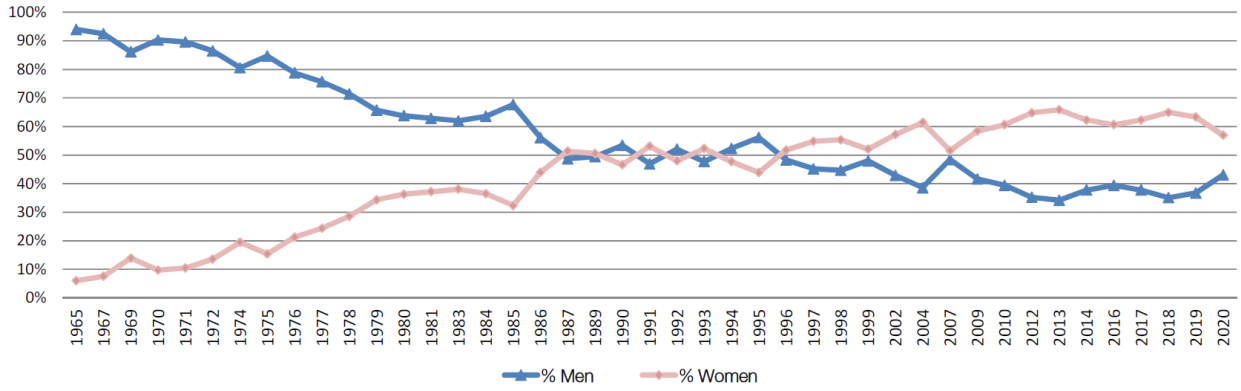
"Gender diversity and decision-making in teams"

Maddalena Ronchi Viola Salvestrini

This appendix is supplemental to the manuscript "Gender diversity and decision-making in teams."

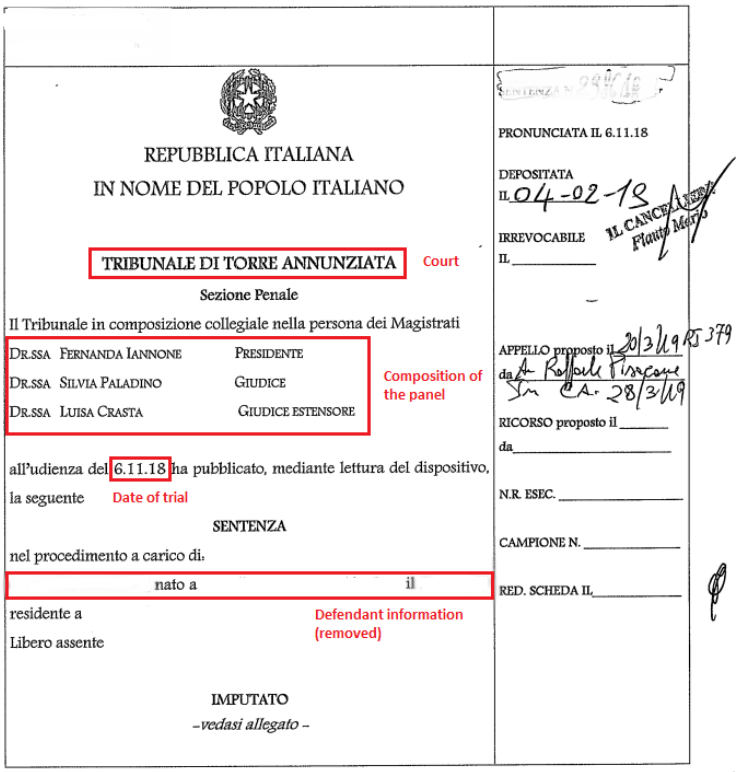
A Other Tables and Figures

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 — Court

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

Composition of the panel

all'udienza del 6.11.18 ha pubblicato, mediante lettura del dispositivo, la seguente — Date of trial

SENTENZA

nel procedimento a carico di:

nato a — il

residente a

Libero assente

Defendant information (removed)

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 Fiorenza Sost. CA. 28/3/19

RICORSO proposto il

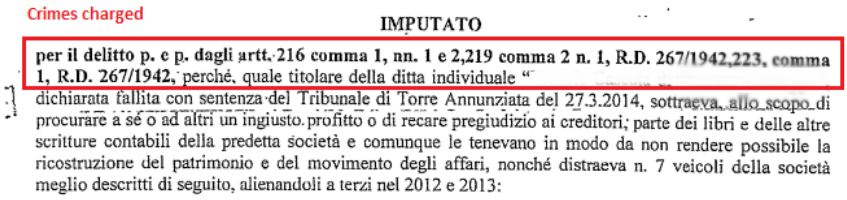
da

N.R. ESEC.

CAMPIONE N.

RED. SCHEDA IL

(a) Court, date, team, defendant



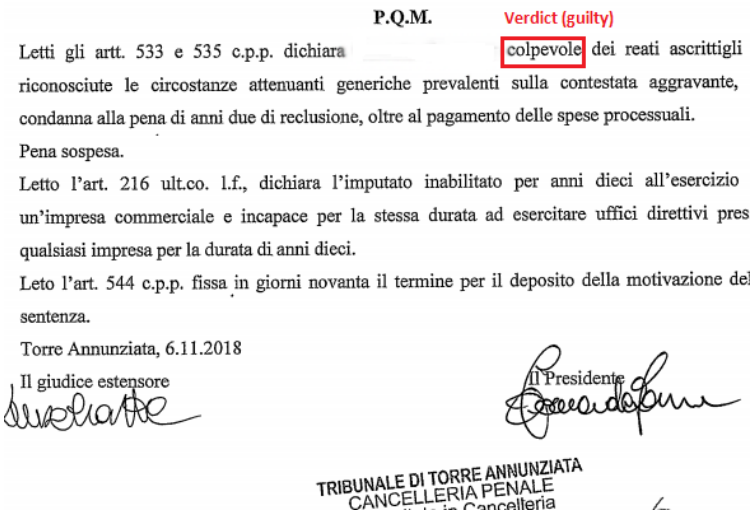
Crimes charged

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. — Verdict (guilty)

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 gender composition of the appellate court

	(1) Quality (cond.)
All-men team	0.031
	(0.060)
Matching appeal team	0.027
	(0.067)
All-men team x Matching appeal team	-0.011
	(0.109)
2M1F team	-0.009
	(0.040)
2M1F team team x Matching appeal team	0.039
	(0.083)
1M2F team	-0.034
	(0.037)
2M1F team team x Matching appeal team	0.008
	(0.075)
Observations	10395
Court and crime x year FEs	X
Controls	X
Mean dep.var. for All-women team	0.784

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 A4: Effect of conviction and jail time on appeal probability

	(1) Prob. appeal	(2) Prob. appeal	(3) Prob. appeal (cond)
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 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

	(1) Unconditional	(2) Conditional	(3) Controlling for jail time
A. Without jailtime condition			
All-men team	-0.041**	-0.065***	-0.072***
	(0.018)	(0.024)	(0.025)
2M1F team	-0.005	-0.044***	-0.049***
	(0.008)	(0.017)	(0.017)
1M2F team	-0.012	-0.042**	-0.045**
	(0.009)	(0.018)	(0.018)
Observations	65051	27686	27686
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var.	0.960	0.919	0.919
B. With tolerance in jail time difference			
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	-0.044*	-0.040*
	(0.013)	(0.024)	(0.023)
Observations	65051	27686	27686
Court and crime x year FEs	X	X	X
Controls	X	X	X
Mean dep.var. for All-women team	0.902	0.799	0.799

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 and experience 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. for All-young team	0.369	0.698
B. Experience diversity		
High-experience team	-0.055	0.038
	(0.059)	(0.064)
Mixed-experience team	-0.028	0.017
	(0.039)	(0.044)
Observations	46738	20475
Court and crime x year FEs	X	X
Controls	X	X
Mean dep.var. for All-inexperienced team	0.361	0.693

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, while Panel B focuses on experience 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. *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. The omitted category in Panels A, and B is an all-young team and a low-experience 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 judge fixed effects

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

Notes. The dependent variable in columns (1) and (2) is an indicator for whether the team convicts the defendant; in columns (3) and (4) is an indicator for whether the conviction is appeal-proof. Columns (1) and (3) focus on women only, columns (2) and (4) focus on men only. *Same-gender team* is an indicator for whether the team is gender-homogeneous. Controls include proxies of case complexity (number of defendants and crimes), gender and nationality of the defendant. Standard errors clustered at the judge level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A9: Controlling for court-by-year and court-by-year-by-crime fixed effects

	Guilty		Appeal-proof (cond.)	
	(1)	(2)	(3)	(4)
All-men team	-0.089***	-0.104***	-0.106**	-0.086**
	(0.029)	(0.029)	(0.041)	(0.038)
2M1F team	-0.096***	-0.092***	-0.051**	-0.048**
	(0.024)	(0.024)	(0.024)	(0.023)
1M2F team	-0.057**	-0.056**	-0.047**	-0.045**
	(0.024)	(0.023)	(0.022)	(0.021)
Observations	65050	65035	27685	27666
Court x year and crime x year FEs	X		X	
Court x year x crime FEs		X		X
Controls	X	X	X	X
Mean dep.var. for All-women team	0.485	0.485	0.784	0.784

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 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. Regressions in column (1) and (3) include court-by-year and crime type-by-year fixed effects. Regressions in column (2) and (4) include court-by-year-by-crime fixed effects Controls include proxies of case complexity (number of defendants and of crimes), gender and nationality of the defendant, and judges' age, region of birth, experience and ability. Standard errors clustered at the team level in parentheses. * p< 0.1, ** p< 0.05, *** p< 0.01.

Table A10: Different classification of crimes

	(1) Guilty	(2) Quality (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. for All-women team	0.485	0.784

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. for All-women team	0.485	0.485	0.485	0.485	0.485	0.485	0.485	0.485	0.485	0.485
B. Quality (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. for All-women team	0.784	0.784	0.784	0.784	0.784	0.784	0.784	0.784	0.784	0.784

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) Male	(2) Female	(3) Italian	(4) Foreign
A. Guilty				
All-men team	-0.077**	-0.070	-0.060*	-0.115**
	(0.030)	(0.056)	(0.033)	(0.053)
2M1F team	-0.091***	-0.129***	-0.092***	-0.091**
	(0.024)	(0.040)	(0.026)	(0.038)
1M2F team	-0.051**	-0.129***	-0.064**	-0.030
	(0.024)	(0.036)	(0.025)	(0.033)
Observations	57447	7224	56005	8580
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var. for All-women team	0.485	0.485	0.485	0.485
B. Appeal-proof (cond.)				
All-men team	-0.163***	-0.306***	-0.191***	-0.104*
	(0.051)	(0.079)	(0.057)	(0.062)
2M1F team	-0.068**	-0.028	-0.068**	-0.047
	(0.027)	(0.047)	(0.030)	(0.041)
1M2F team	-0.053**	-0.024	-0.055*	-0.018
	(0.025)	(0.047)	(0.029)	(0.042)
Observations	24627	2808	22882	4611
Court and crime x year FEs	X	X	X	X
Controls	X	X	X	X
Mean dep.var. for All-women team	0.784	0.784	0.784	0.784

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

Table A13: Conviction and appeal-proof conviction rates for estimation

	(1) Guilty	(2) Appeal-proof (cond.)
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. for All-women team	0.485	0.784

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.

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 in-

quisitorial powers), in which the judge (or team of judges) plays an active role during the trial.[76] 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 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.[77]

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.

[76] 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.

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

C Model derivations

This appendix collects proofs and derivations for the structural model of Section 7.2. We use the same notation as in the main text. In addition, we let $\bar{e}_i(e_j)$ denote the minimum own effort at which signals become decision-relevant, $F_i(e_i, e_j)$ the implicit equation from the first-order condition (4), and D_i the associated second-order condition.

A. Alternative voting regime

In the main text, we focus on the interior regime in which equilibrium precision satisfies $\theta > 1 - \bar{\tau}$, where $\bar{\tau} \equiv \beta(1-p)/[p + \beta(1-p)]$ and $\beta = \max(\beta_i, \beta_j)$. In this regime, only a guilty signal triggers a vote to convict. Here we characterize the alternative regime in which $\theta \leq 1 - \bar{\tau}$.

Note that $1 - \bar{\tau} = p/[p + \beta(1-p)]$. For the alternative regime to be compatible with $\theta > 1/2$ (which holds by construction), we need $1 - \bar{\tau} > 1/2$, i.e., $p > \beta/(1+\beta)$. When $\beta > 1$, this requires the prior to be sufficiently high relative to leniency. When this condition holds and $1/2 < \theta \leq 1 - \bar{\tau}$, we also have $\theta \geq \bar{\tau}$ (since $\bar{\tau} < 1/2$), so both conditions (2) and (3) are satisfied simultaneously: the posterior exceeds the conviction threshold after *both* guilty and innocent signals. Each judge therefore votes to convict regardless of the signal received.

Outcomes. Since both judges always vote to convict, $C = 1$ and $APC = p$: the team convicts every defendant indiscriminately, and the share of correct convictions equals the prior probability of guilt.

Effort. When the decision does not depend on the signal realization, improving signal precision θ has no effect on the judge's expected payoff. The expected utility of judge i reduces to

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

which is strictly decreasing in e_i. The unique optimum is the corner solution $e_i^* = 0$.

Best-response structure. The corner and interior regimes generate a discontinuity in the best-response function. For a given e_j, define the threshold effort $\bar{e}_i(e_j) \geq 0$ as the minimum own effort such that signals become decision-relevant:

$$\theta(\bar{e}_i(e_j),\, e_j) = 1 - \bar{\tau}.$$

For $e_i < \bar{e}_i(e_j)$, precision is too low for signals to affect votes, and the judge votes to convict after any signal (alternative regime). For $e_i \geq \bar{e}_i(e_j)$, only guilty signals trigger conviction (interior regime). Since θ is increasing and concave in each argument, $\bar{e}_i(e_j)$ is decreasing in e_j: a harder-working partner lowers the effort threshold at which signals become informative.

The optimal effort function therefore takes the form:

$$e_i^*(e_j) = \begin{cases} 0 & \text{if } \mathbb{E}[u_i(0, e_j)] \geq \mathbb{E}[u_i(\hat{e}_i(e_j), e_j)] \\ \hat{e}_i(e_j) & \text{otherwise} \end{cases}$$

where $\hat{e}_i(e_j) = \max\{\bar{e}_i(e_j),\, e_i^{int}(e_j)\}$ and $e_i^{int}(e_j)$ denotes the interior solution to the first-order condition (4). That is, judge i either exerts zero effort (remaining in the alternative regime) or exerts effort at or above the threshold that makes signals informative. Effort levels in the open interval $(0, \bar{e}_i(e_j))$ are never optimal, since the marginal return to effort is zero in this region while costs are strictly increasing.

An equilibrium is a fixed point of the joint map $(e_1^*(e_2), e_2^*(e_1))$. Because of the discontinuity, existence is not guaranteed by standard continuity arguments but can be verified numerically for any given parameter vector. At estimated parameters, we verify in Section 7.2.2 that the interior equilibrium obtains for all team types.

B. Uniqueness of the interior equilibrium

We show that in the interior regime ($e_i \geq \bar{e}_i(e_j)$ for both judges, so that $\theta > 1 - \bar{\tau}$ and only guilty signals trigger conviction), the equilibrium is unique. We maintain the following regularity condition on the precision function:

Assumption. $|\theta_{e_i e_j}(e_i, e_j)| \leq |\theta_{e_i e_i}(e_i, e_j)|$ for all $(e_i, e_j) > 0$.

This assumption states that the cross-curvature of θ does not exceed its own-curvature: the marginal return to judge i's effort is at least as sensitive to i's own effort as it is to the partner's. It is satisfied, in particular, by any precision function of the form $\theta = h(e_i + e_j)$ with h concave, since in that case $\theta_{e_i e_i} = \theta_{e_i e_j} = h''$.

Proof. The first-order condition (4) implicitly defines each judge's interior best-response function. Define

$$F_i(e_i, e_j) \equiv \theta_{e_i}(e_i, e_j) - \frac{c'(e_i)}{\Phi_i} = 0,$$

where $\Phi_i \equiv \phi_i[p + \beta_i(1-p)] > 0$. By the implicit function theorem,

$$\frac{de_i^*}{de_j} = -\frac{\theta_{e_i e_j}}{\theta_{e_i e_i} - c''(e_i)/\Phi_i}.$$

The denominator $D_i \equiv \theta_{e_i e_i} - c''(e_i)/\Phi_i$ is strictly negative (by concavity of θ in own effort and strict convexity of c). Since $\theta_{e_i e_j} < 0$, the slope $de_i^*/de_j < 0$: best responses are downward-sloping, confirming strategic substitutability.

To establish uniqueness, it suffices to show that the product of the best-response slopes is strictly less than one:

$$\left|\frac{de_i^*}{de_j}\right| \cdot \left|\frac{de_j^*}{de_i}\right| = \frac{\theta_{e_i e_j}^2}{D_i \cdot D_j} < 1.$$

Under the assumption above, $\theta_{e_i e_j}^2 \leq \theta_{e_i e_i} \cdot \theta_{e_j e_j}$. Since $D_i = \theta_{e_i e_i} - c''(e_i)/\Phi_i$ and $c''(e_i)/\Phi_i >$

0, we have $|D_i| > |\theta_{e_ie_i}|$ and analogously $|D_j| > |\theta_{e_je_j}|$. Therefore

$$\frac{\theta_{e_ie_j}^2}{D_i \cdot D_j} \leq \frac{\theta_{e_ie_i} \cdot \theta_{e_je_j}}{D_i \cdot D_j} < 1.$$

The best-response map is a contraction, so the interior equilibrium is unique. □

Verification for the logistic specification. The functional form used in estimation is $\theta(e_i, e_j) = \sigma\big((e_i + e_j)/2\big)$, where $\sigma(x) = e^x/(1+e^x)$ is the logistic function. Since θ depends on $e_i + e_j$, we have $\theta_{e_ie_i} = \theta_{e_ie_j} = \sigma''(\cdot)/4$, so the assumption holds with equality and the result applies.

C. Comparative statics of equilibrium effort

We derive the comparative statics stated in the main text by differentiating the first-order condition (4). Using the notation from Part B, $F_i(e_i, e_j; \phi_i, \beta_i, p) = \theta_{e_i}(e_i, e_j) - c'(e_i)/\Phi_i = 0$, where $\Phi_i = \phi_i[p + \beta_i(1-p)]$. Let $D_i = \theta_{e_ie_i} - c''(e_i)/\Phi_i < 0$ (the second-order condition).

Own diligence ($\partial e_i^*/\partial \phi_i > 0$).

$$\frac{\partial e_i^*}{\partial \phi_i} = -\frac{1}{D_i}\frac{\partial F_i}{\partial \phi_i} = -\frac{1}{D_i}\frac{c'(e_i)}{\phi_i^2[p + \beta_i(1-p)]} > 0,$$

since $D_i < 0$ and all remaining terms are positive.

Partner's diligence ($\partial e_i^*/\partial \phi_j < 0$). An increase in ϕ_j raises e_j^* (by the result above), which affects e_i^* through the best-response function. Since $de_i^*/de_j < 0$ (strategic substitutes), $\partial e_i^*/\partial \phi_j < 0$.

Own leniency ($\partial e_i^*/\partial \beta_i > 0$).

$$\frac{\partial F_i}{\partial \beta_i} = \frac{c'(e_i)(1-p)}{\phi_i[p + \beta_i(1-p)]^2} > 0,$$

so $\partial e_i^*/\partial \beta_i = -(\partial F_i/\partial \beta_i)/D_i > 0$.

Partner's leniency ($\partial e_i^*/\partial \beta_j < 0$). By the same logic as for partner's diligence: higher β_j raises e_j^*, which lowers e_i^* through strategic substitutes.

Prior probability of guilt ($\partial e_i^*/\partial p < 0$ when $\beta_i > 1$).

$$\frac{\partial F_i}{\partial p} = \frac{c'(e_i)(1-\beta_i)}{\phi_i[p + \beta_i(1-p)]^2}.$$

When $\beta_i > 1$, this is negative, so $\partial e_i^*/\partial p = -(\partial F_i/\partial p)/D_i < 0$.

References

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.