When Diversity Rules*

Abhinav Gupta, Joseph Kalmenovitz, Ravi Ranjan, and Kairong Xiao

January 15, 2026

We study whether gender diversity influences regulatory outcomes using a new dataset covering 15,000 federal rulemakers and 38,000 rulemaking projects. Teams that include women deliberate longer and revise rule drafts more extensively, yet their projects are less likely to be approved and become a final rule. We develop a model in which output gaps arise from four channels: sorting, cost of effort, incentives, and managerial evaluation. Reduced-form evidence, combined with a machine-learning C-BERT model, shows that women sort into lower-probability projects and face harsher evaluations by agency leadership, but receive stronger incentives and exert greater effort. Structural estimates confirm that disadvantages in sorting and evaluation account for much of the gender gap, partially offset by incentive effects. Overall, gender representation materially shapes regulatory performance.

Keywords: Economics of Regulation, Gender Diversity, Performance Incentives, Structural Estimation, Machine Learning

JEL Codes: G18, G34, C55, K23, J16

*Kalmenovitz is from the University of Rochester (jkalmeno@simon.rochester.edu, corresponding author); Gupta is from the University of North Carolina (Abhinav_Gupta@kenan-flagler.unc.edu); Ranjan is from the University of North Carolina (Ravi_Ranjan@kenan-flagler.unc.edu); Xiao is from Columbia University (kx2139@columbia.edu). We received excellent comments from Charles Hadlock and David Yermack, and from seminar and conference participants at RCF-ECGI Corporate Finance and Governance Conference, the University of Texas Austin, Florida State University, and Fordham University.

Introduction

Does diversity enhance or hinder organizational performance? This question is at the center of a growing literature in finance and economics, and it has motivated a flurry of policy initiatives on gender quotas in corporate boards and executive positions.[1] While many studies explore how gender diversity affects firms, far less is known about its impact in the public sector. This omission is unfortunate, as the government offers a compelling setting to study how diversity shapes performance. Thousands of career professionals work in teams to produce complex outputs which regulate key parts of the economy, from financial markets to public health. Importantly, the public sector provides unusually rich data on personnel, team composition, and project outcomes. This transparency enables a detailed analysis of how diversity within decision-making teams correlates with performance.

These questions are not merely academic, as diversity in government has become a highly controversial issue. On the one hand, an elected government is expected to reflect the diversity of its electorate. On the other hand, efforts to promote diversity may yield inefficiencies. In January 2025, the Trump administration issued an Executive Order banning the promotion of DEI in the federal workforce since it undermines "hard work [and] merit", thereby reversing the Biden administration's earlier expansion of DEI across federal agencies to improve operational effectiveness.[2] These opposing positions reflect a deeper tension between representation and performance in public service, with far-reaching implications for how public institutions are designed and evaluated.

In this paper, we use the U.S. federal rulemaking process to study whether and how gender diversity affects performance. Rulemaking teams are responsible for transforming preliminary proposals into finalized regulations, working within structured timelines and under public scrutiny. We assemble a novel dataset that tracks 15,000 individual rulemakers working on 38,000 regulatory projects over more than two decades. We find that female-inclusive teams

[1] See, for instance, California's SB 826 and the European Union's Directive 2022/2381.
[2] See Executive Order 14035 (86 FR 34593) and Executive Order 14148 (90 FR 8237).

deliberate more extensively and revise initial drafts more aggressively, yet are ultimately less likely to complete their projects. To interpret these patterns, we develop a simple model in which performance gaps can arise from four channels: selection (if women are disproportionately assigned to lower-quality projects), costs (if effort is more burdensome), incentives (if the return to effort varies), and evaluation (if female-authored work is judged more harshly). Reduced-form and structural estimates show that disparities in selection, costs, and evaluation account for most of the observed gender gap, while stronger performance incentives for women partially offset these disadvantages.

Our findings contribute to both policy and academic debates. For policymakers, they offer evidence on how female representation may influence the productivity of regulators. For economists, they provide rare micro-level insight into how diversity affects performance in a high-stakes, data-rich environment. While grounded in the public sector, the underlying mechanisms – assignment, cost of effort, incentives, and evaluation – apply broadly to organizations seeking to reconcile equity with effectiveness.

To carry out the analysis, we assemble a novel employee-level dataset on rulemaking activities and career outcomes from two primary sources. The first source is the Unified Agenda, a twice-annual compilation of all federal rulemaking activities. We scrape every Agenda edition since 1996, collecting information on each rulemaking project and its outcome. Crucially, we collect the names of all individuals working on the project, and we assign gender to individuals whose first names exceeds a 75% likelihood of being male or female (based on Social Security gender frequency tables). The second source is the civilian federal payroll, obtained via Freedom of Information Act requests. This employee-level dataset provides detailed employee records such as agency, experience, rank, and pay. We match rulemakers from the Agenda to their respective payroll records, using a structured fuzzy-matching procedure and manual verification. Our final dataset covers 38,000 rulemaking projects and 15,000 individual rulemakers. The average rulemaking team comprises 1.6 rulemakers. Across all rule proposals, the completion rate is 70% and the remaining projects are officially

terminated or practically abandoned. Slightly less than half the rulemakers (45%) are females. The average rulemaker has 18 years of experience, earns $156,600 in inflation-adjusted annual compensation, and works on 4 rulemaking projects.

With the new dataset at hand, we compare the performance of all-male teams, where no female is included, to that of female-inclusive teams, where at least one female is included. We find that inclusive teams are more likely to publish a draft rule and to issue an interim rule. However, all-male teams are more likely to complete their project and publish a final rule. Those differences emerge from a set of regressions, with various rulemaking milestones as the outcome and an indicator for a female-inclusive team as the independent variable. In our preferred specification we add the interaction of year, agency, and tier[3] fixed effects, to account for known determinants of rule progression: macro-level shocks, agency practices, and rule tiers. Effectively, we compare two equally-important projects initiated by the same agency at the same time, but assigned to two different teams. All the results are statistically significant and the magnitudes range from 2.1% to 2.9% of the sample mean.

Our preliminary findings are puzzling: despite appearing more active in the drafting and deliberation stages, female-inclusive teams are less likely to complete their rulemaking projects. To understand this dynamic, we develop a simple structural model of the rulemaking process. The rulemaking project has an initial perceived value, determined by its preliminary features such as topic and broad goals. In the first stage, the principal selects a rulemaker to work on the project.[4] In the second stage, the team collectively chooses how much effort to invest in the project, balancing the marginal cost of effort against its benefit (such as pay-for-performance incentives). In the final stage, the team presents the draft rule and the principal evaluates the draft. The evaluation includes a systematic assessment of the draft's substance, and importantly can be influenced by the team's characteristics.[5] The rule is

[3]We use the three-tier system (based on Executive Order 12866) to differentiate between significant, substantive, and routine or administrative rules.

[4]We do not distinguish between the self-selection by the rulemaker (bottom-up) and the assignment by the principal (top-down). In practice, a team of rulemaker can be assigned, which we interpret as a representative rulemaker.

[5]In our model, the evaluation can also be influenced by an idiosyncratic shock such as industry pushback

enacted only if the total net benefit – determined by the project's initial value, the team's effort, and the principal's evaluation – is positive.

Our model highlights four potential channels that can contribute to the observed gender gap in completion rates: the selection channel, whereby the principal selects males for projects with a higher perceived net benefit; the costs channel, if males have lower costs of effort; the incentives channel, if males have stronger incentives to exert effort; and the evaluation channel, whereby the principal consistently overvalues drafts presented by males. We combine reduced-form and structural tools to investigate the four channels and quantify their importance.

The selection channel implies that males systematically work on projects with higher ex-ante perceived value. To study this possibility, we use a BERT model to convert the title and abstract of each rulemaking proposal – available at the time of assignment – into a high-dimensional vector representation. We then train a gender-blind neural network to predict whether a project will be completed, using only the initial project description and other pre-assignment characteristics. This process yields our object of interest, *Completability*: the predicted probability of project completion based solely on the proposal's ex-ante quality, without incorporating any team effort or gender-based outcomes.

Using our predicted measure of project completability, we find that proposals assigned to female-inclusive teams have a significantly lower average probability of completion (73.7%) compared to those assigned to all-male teams (75.8%). This 2.1 percentage point gap persists in OLS regressions that control for agency, cohort, and project importance, with estimated differences ranging from 0.9 to 2.1 percentage points, or 1.2–2.8% of the sample mean. These results suggest that female-inclusive teams are more likely to be assigned lower-quality projects based on observable characteristics at the time of assignment.

While the selection channel helps explain part of the gender gap in outcomes, we next turn to examine the possibility that females exert less effort post-selection. In the model,

or political hurdles. We allow the evaluation to reflect both societal considerations (such as correcting market failures) and private benefits to the principal (such as career opportunities).

this would occur if the marginal cost of effort for females exceeds the performance incentives. Since effort is unobserved, we turn to a variety of empirical proxies. We start by focusing on projects with at least two drafts, and compute the share of each draft devoted to each of the 100 LDA-derived regulatory topics (based on data from Chang et al. (2023)). Then, we take the absolute difference in those shares for each topic and average them to get the overall change across drafts. This serves as a proxy for the team's effort during the rulemaking process. Using the proxy as a new outcome for our baseline specification, with the strictest set of fixed effects, we find that females conduct significantly larger textual revisions (9.5% above the sample mean). This finding, combined with our finding that females publish more draft rules and interim rules, suggests that females appear to invest more effort in the early stages of the rulemaking process.

Higher effort may result from stronger performance incentives or a lower cost of effort. While measuring the cost of effort is challenging in a reduced-form analysis and is deferred to the structural model, our data enable a more precise estimation of pay-for-performance sensitivity. We estimate employee-level regressions, where the outcome variable is annual pay and the independent variables are the number of completed rulemaking projects and their interaction with gender. The coefficients represent the average pay-for-performance sensitivity (first variable) and the excess sensitivity for females (interaction variable). We control for the rulemaker's total assigned projects, and we add the interaction of year, agency, location, and occupation fixed effects. Intuitively, we compare the pay progression of two rulemakers with opposing genders who hold the same occupation and work at the same time and place and for the same agency. We find that females exhibit a substantial pay-for-performance sensitivity, with performance-related pay being 1.8-3.0 times more sensitive to output than for males. In other words, while the average male salary in our sample is higher, we find no evidence that they receive stronger performance incentives.

The last part of the reduced-form analysis relates to the evaluation channel. Through the lens of the model, the question is whether all-male teams are more likely to receive

approval from their superiors, conditional on the initial project quality and the effort exerted by the team. In an ideal experiment, we would ask the agency to evaluate identical drafts produced by different teams and compare their approval rates. While we cannot conduct this experiment, for obvious reasons, we strive to approximate it using C-BERT: a causal variant of BERT designed to estimate counterfactual outcomes. Intuitively, we quantify the draft's textual features, "swap" identical drafts between teams, and see if males still complete more projects than females.[6] Focusing on projects where at least one draft was produced, we use BERT to convert the full draft text into high-dimensional vectors. Following Hochberg et al. (2023) and Veitch, Sridhar, and Blei (2020), we use a neural network that estimates the likelihood a draft was written by a female-inclusive team, and exclude 2,258 drafts with strongly gendered language (predicted probabilities in the top or bottom fifth percentile). We then train two separate neural networks to predict completion: one using all-male projects and another using female-inclusive projects. We pass each project through the neural network model trained on the opposite gender to generate a counterfactual probability of completion, that is, the likelihood the project would have been completed if presented by a different team. We then compare this counterfactual to the project's actual outcome to measure how much the final decision reflects the tilt in the principal's evaluation of the draft. A positive tilt value indicates that the draft was evaluated favorably relative to expectations for the opposite gender, while a negative value suggests harsher treatment. The further this value is from zero, the stronger the tilt in evaluation – whether positive or negative.

We find that the average tilt score of males (8.3%) is substantially higher than females (3.1%). This means that, on average, drafts from male teams are evaluated more favorably relative to the counterfactual benchmark than those from female teams. Using our baseline specification with the interaction of year, agency, and tier fixed effects and the tilt score as an outcome, we estimate a 4.4-5.3 percentage points gender gap, representing 5-6% of the mean

[6]While we analyze both the evaluation and selection channels using BERT, the estimations are conducted independently and the outcomes have distinct interpretations.

tilt.[7] Furthermore, observable team characteristics such as pay, tenure, location, rank, and occupation do not explain the gap: even after controlling for these variables, female-inclusive teams have a significantly lower tilt (4.1 percentage points, or 4.5% of the mean) than males. We then examine heterogeneity across project cohorts, regulatory topics, and agencies. The gender gap in tilt persists across nearly all cohorts and is especially pronounced in policy areas such as taxation, legal operations, and banking. Major financial regulators such as the SEC, FDIC, and CFPB exhibit particularly large gaps.

In sum, the reduced-form analysis reveals persistent gender gaps in the rulemaking process. Females are assigned to rules with lower ex-ante probability of completion, and the drafts they present exhibit lower tilt scores, indicating systematically less favorable treatment from the agency's leadership relative to comparable work by male teams. However, females appear to be better rewarded for completing their projects and to exert more effort, at least in the earlier stages of the rulemaking process. Combined, these findings support the interpretation that the lower completion rates are driven primarily by worse selection and evaluation, and less so by gaps in effort. While those findings are informative, we recognize the constraints in a reduced-form analysis due to limited observability. For instance, it is possible that females exert less effort, especially in the latter stages of the rulemaking process.

To gain better insight into these dynamics, we turn to the structural estimation of four key parameters in our model: the initial net benefit of the rule, the pay-to-effort sensitivity, the cost of effort, and the principal's evaluation. In particular, we seek to understand whether those parameters vary by gender. We document several important findings. First, females are systematically assigned to projects with lower initial quality: on average, their projects have roughly 7% lower initial perceived net benefit than those given to males. These estimates broaden the scope beyond the BERT algorithm, which can only capture the project's initial quality based on visible textual features. This selection mechanism, as female rulemakers are more likely to be tasked with lower-quality projects, constrains their ability to complete their

[7] We obtain similar results when we replace the binary indicator (any female) with a continuous variable (share of female team members).

rulemaking projects.

We further estimate that females enjoy stronger performance incentives, as each additional unit of work increases their expected payoff by 44.8% more than it does for males. On the other hand, effort is 44.4% more costly for females than it is for males. This creates a genuine trade-off, as females must balance the higher cost against the greater return. Interestingly, the trade-off emerging from the structural estimation is consistent with its empirical counterpart, where we find that female rulemakers are better rewarded than their male counterparts for completing their rulemaking projects. Naturally, those are empirical proxies, while the corresponding structural parameters represent their full extent. On balance, we estimate that the benefit of effort outweighs its costs, resulting in slightly more effort from females. This is consistent with our reduced-form analysis, where we find that females appear to exhibit greater effort in the earlier stages of the rulemaking process.

Most striking, however, are the differences in principal evaluation: once effort and initial project quality are held constant, proposals developed by females are judged less favorably by decision makers. Normalizing the male valuation to zero, we find a sizable discount attached to female output which contributes to the observed gender gap in output. Note that our model does not specify the origins of either selection or evaluation, but this combination can potentially create a doom loop: females are assigned to low-quality rules, yielding lower completion rates, which justifies a worse principal evaluation, leading to even worse assignments. It is also possible that evaluation and selection gaps arise endogenously, as a response to initiatives to provide greater reward per unit of effort to underrepresented groups.

Taken together, selection, cost of effort, and principal evaluation contribute to the observed gender gap, with compensation incentives partially narrowing it. To disentangle the relative importance of each channel, we simulate four counterfactual scenarios. First, by matching men and women to rules of identical baseline quality – for instance, a blind assignment mechanism – we show that differential assignment explains nearly a third of the observed output gap. Second, substituting women's effort costs with their male counterparts' lowers

the gap by 44%, highlighting how steeper effort penalties hinder females' ability to translate work into outcomes. In practice, this policy might involve offering women targeted support programs, such as reduced administrative burdens or research assistants. Third, the evaluative discount on female output accounts for 61% of the output gap, confirming that differential evaluation is a key barrier. In practice, agencies could implement standardized performance rubrics to minimize evaluation gaps. Finally, swapping women's pay-to-effort sensitivity for men's demonstrates that existing stronger compensation incentives for women act to reduce the gender disparity. Eliminating this channel, for instance, by equalizing bonus formulas across genders, could exacerbate the observed rulemaking gap by 70%.

In sum, we investigate how inclusion of females affects the rulemaking process. Despite engaging in more extensive drafting, female-inclusive teams are less likely to complete their projects and issue a final rule. We explore this finding through a simple three-stage model highlighting the selection, cost of effort, incentives, and evaluation channels. Leveraging reduced-form regressions, a machine-learning approach, and structural estimation, we find that females are systematically selected for lower-quality projects, face steeper costs of effort, and produce rule drafts that are evaluated less favorably by their principals. On the other hand, they receive stronger incentives and thus exert greater effort. Note that we do not take a stand on how diversity and inclusion in regulatory agencies affects total welfare. For instance, we focus on the perceived value of the rule, not on its true value, and on the optimization problem of a representative team, not on changes in total productivity or output. We leave those important questions for future follow-on studies.

Our paper makes the following four contributions. First, we add to the literature on regulatory performance where scholars study the role of pay, outside job opportunities, and organizational features.[8] Our paper is the first to highlight the role of gender diversity and to

[8]For studies on pay, see for instance Dal Bó, Finan, and Rossi (2013) and Kalmenovitz (2021). For studies on the revolving door, see for instance deHaan et al. (2015); Lucca, Seru, and Trebbi (2014); Tabakovic and Wollmann (2018); Duchin, Kalmenovitz, and Wang (2024). For studies on organizational features, see for instance Gopalan, Kalda, and Manela (2021) (field offices), Hirtle, Kovner, and Plosser (2020) (supervision hours), and Kalmenovitz and Vij (2025) (risk assessment).

show that female inclusion, a hotly-debated subject among scholars and policymakers around the world, has a substantial impact on regulatory outcomes. Moreover, we deploy various methodologies to uncover the channels connecting diversity to rulemaking, quantifying in particular the role of selection, effort, incentives, and principal evaluation.

Second, we shed new light on the regulatory pipeline, where rule proposals are initiated, revised, and (sometimes) finalized. Recent studies show that firms closely monitor the pipeline and adjust their behavior in advance.[9] We contribute to this nascent literature by studying the incentives and performance of the individuals who work on the regulatory pipeline. We show that the internal organization of the federal workforce can spill over, by changing the dynamics of the regulatory pipeline and the number of enacted regulations. While our focus is on gender diversity, we formalize and estimate broader organizational features such as compensation incentives, project selection, and evaluation by agency leadership.

Third, we add to the robust economic literature on diversity and performance. Studies are exclusively focused on the private sector,[10] and we extend it by studying diversity among federal regulators. Utilizing diverse empirical methods and a comprehensive project-level and employee-level dataset, we demonstrate how – and why – diversity among rulemaking teams relates to the likelihood of completing the project. While our findings may not apply equally to private sector diversity initiatives, at the minimum they contribute to our understanding of how diversity affects an understudied group of employees who perform complex tasks with significant economic impact.

Fourth, we contribute to the growing literature which uses structural estimation to study various corporate governance issues.[11] We extend this literature by developing and estimating a structural model that jointly connects team selection, effort, and evaluation to project outcomes. This framework enables us to go beyond descriptive evidence, providing a

[9] See, for instance, Chang et al. (2023).

[10] See review in Appendix A.2.

[11] See, for instance, Wu (2017) (CEO career concerns); Gao, Wang, and Wu (2024) (career decisions of finance professionals); Coles, Lemmon, and Meschke (2012) (CEO ownership); Taylor (2010, 2013); Page (2018) (CEO turnover and CEO-board dynamics); and Morellec, Nikolov, and Schürhoff (2012); Nikolov and Whited (2014) (capital structure).

unified lens through which to interpret our reduced-form patterns, quantify the magnitude of underlying frictions, and evaluate their implications. In particular, the model allows us to conduct counterfactual policy simulations, offering guidance on how institutional design changes could mitigate inefficiencies. Moreover, we combine structural estimation with C-BERT, a large language model trained for context-specific semantic analysis, to study this issue. This approach enables us to extract rich, high-dimensional measures of proposal content and team output that are difficult to capture with standard text-analysis methods. It is also applicable to other settings where textual data can inform the estimation of structural parameters, offering new opportunities for research at the intersection of applied microeconomics, corporate governance, and machine learning.

1 Background

1.1 Data and institutional setting

Institutional setting. The paper is centered on the federal rulemaking process.[12] The sequence typically begins when a legislative mandate, such as the Dodd-Frank Act, directs the agency to craft a rule on a specific subject. Some rules are initiated by the agency as part of its broad authority, not by a specific act of Congress. The agency will announce its intent to craft a rule, and later on publish a draft in the Federal Register and open a public comment period (typically 30–60 days). After weighing submitted feedback, the agency may adjust the draft and the revision process can take several iterations. At the end, the agency will decide whether to abandon the effort or publish a final rule with a specified effective date in the Federal Register. The new rule will be subsequently incorporated into the Code of Federal Regulations.

[12]The process is primarily outlined in Section 553 of the Administrative Procedure Act of 1946 and Executive Order 12866. We benefited greatly from conversations with a former administrator at the Office of Information and Regulatory Affairs (OIRA), as well as Administrative Conference of the United States (2015) and Kerwin and Furlong (2018).

As an illustrative example, consider the U.S. Securities and Exchange Commission (link). A rulemaking project is typically led by one division, such as Corporation Finance or Investment Management. Staff within the division conduct background analysis, consult relevant stakeholders, and develop a draft rule. The draft is submitted to the five-member Commission for approval. A majority vote by the Commissioners is required to issue the proposed rule for public comment. After receiving and reviewing public feedback, staff may revise the draft, and the final rule is again subject to a vote by the Commission. Only if a majority votes in favor is the rule adopted and published in the Federal Register. Another example is the Federal Communications Commission (link), also headed by five Commissioners. When a new rule is proposed, staff from the relevant Bureau – such as the Wireline Competition Bureau or the Wireless Telecommunications Bureau – are assigned to draft a Notice of Proposed Rulemaking. This draft is reviewed internally and must be approved by a majority vote of the Commissioners before being published for public comment. After the comment period and any necessary revisions, the Commissioners vote again to determine whether to adopt the rule in final form.

Data sources. The pivotal dataset is sourced from the Unified Agenda (UA), widely regarded as a comprehensive repository of federal rulemaking activities. Published twice annually, the UA contains information on rulemaking actions by all federal agencies that are planned or underway.[13] Each agency submits a detailed status update on its rulemaking initiatives, ranging from preliminary considerations to rules that are nearly finalized. These submissions are coordinated and compiled by OIRA, an entity within the Office of Management and Budget. For each regulation, agencies outline their timeline of actions such as issuing a notice of proposed rulemaking, hosting public meetings, or withdrawing a proposal in part or in full. The resulting UA offers a consolidated description of regulatory developments across the federal government. Each rule is assigned a unique Regulatory Identification

[13]Submission to the UA satisfies statutory obligations under Section 602 of the Regulatory Flexibility Act and Section 4(b) of Executive Order 12866.

Number (RIN), a stable identifier throughout the rulemaking process. In what follows, we use the terms "RIN," "rule," and "regulation" interchangeably to refer to a specific regulatory effort.[14] For parts of the analysis we exploit the fact that the Agenda differentiates between three tiers of rules.[15] Tier 1 includes *significant* rules, which raise novel legal issues, and *economically significant* rules, which have an annual economic effect of $100 million or more. Tier 2 includes *substantive* rules that are important but not sufficiently to be labeled as significant. Tier 3 includes *routine* and *administrative* rules which largely pertain to how federal agencies organize themselves and execute their powers. Separately, some rules are labeled *major*, a definition that partially overlaps with the three groups (see 5 U.S.C. §804(2)).

We download all the UA editions between 1996-2022 and scrape all the information for each RIN across the various UA editions. We identify the key milestones during the RIN's lifetime, including whether the agency has published a draft or an interim rule and how long the project has been pending in the pipeline. We also derive the primary outcome variable in this paper, identifying whether the project was completed (issuing a final rule) or terminated. Finally, we collect the names of all team members who worked on each RIN. We extract the first names and match them to Social Security gender frequency tables. We define a threshold of 75% to identify gender: male (female) regulator is one whose first name has 75% or higher chance of being a male (female).[16] At the end, we were able to collect the names of 15,492 individual rulemakers, out of which 14,792 have an identified gender based on the process described above.

The secondary dataset we use is the federal payroll, which we obtained by submitting repeated Freedom of Information Act requests to various federal entities. This dataset covers

[14] In some instances, a single rulemaking action may involve multiple RINs, particularly when more than one agency is involved. Conversely, a single RIN may encompass multiple regulatory amendments or sections within the Code of Federal Regulations.

[15] The system was established in 1993, following Executive Order 12866 which subjected significant regulations to a more rigorous vetting process.

[16] VanHelene et al. (2024) report that common gender classifiers – such as Genderize, Gender API, and the gender R package – are over 95% accurate and typically use a 50% probability cutoff for classification. Namsor (https://namsor.app/features/gender-name/) notes that gender predictions are unreliable when the probability falls between 45% and 55%. To reduce the risk of misclassification, we adopt a more conservative threshold of 75%.

all civilian federal employees since 1996 and includes employee-level information on agency, occupation, hiring date, location (city), salary, and rank. We match the rulemakers from the Agenda to their payroll records in a process which we describe in Appendix A.1, relying on a structured fuzzy-matching procedure and accounting for variations in name formats. The matching required us to exclude agencies with masked employee data, focusing on 29,528 rule proposals from eligible agencies. Of these, 82% included at least one matched team member, and 72% could be fully matched to payroll records. Match rates were especially high in several financial and economic agencies, with over 80% in the Department of Commerce, Federal Housing Finance Board, and Commodity Futures Trading Commission. At the end, we were able to match 10,246 individual rulemakers to their corresponding payroll records.

1.2 Descriptive statistics

Table 1, Panel A, describes the sample of 37,739 teams (rule proposals). The average team is comprised of 1.6 rulemakers, with 52.9% of the teams including at least one female. On average, team composition is 53.0% male and 42.8% female. Average team pay is $176,300 and team tenure is 18.2 years. Slightly more than half of the teams (54.7%) are based in Washington, DC. Nearly a third (29.2%) of the rulemaking projects were terminated while two-thirds (70.8%) were completed, resulting in either a final or an interim rule.[17] On average, a rulemaking project lasts 748 days. A quarter (24.5%) of rulemaking projects are officially suspended for at least 12 months.[18] In terms of importance, a quarter (27.7%) of the rules belong to Tier 1, two-thirds (64.3%) belong to Tier 2, and 8% to Tier 3. Separately, 5.4% of the rules are considered Major.

Panel B describes the sample of 15,492 individual rulemakers. The gender composition is slightly unbalanced, with 45% females. The average rulemaker works on 3.8 rulemaking projects, two-thirds of which (66.3%) are completed and the rest are terminated. The

[17] Specifically, 61.9% were issued as final rules only, 3.3% as interim rules only, and 5.6% as both final and interim rules. The remaining rules are still pending or with indeterminate outcome.

[18] Those numbers are comparable to Chang et al. (2023), who study the impact of the rulemaking pipeline on firm outcomes using data from the Unified Agenda.

average rulemakers has 1.5 coauthors, 43% of which are females. Among rulemakers with available payroll data, the average rulemaker was hired in 1991 and earned $156,600 in total compensation. They receive on average a 3% annual pay raise and 3.8 promotions throughout their career. For reference, Kalmenovitz, Vij, and Xiao (2025) study a larger sample of federal employees with 21.9 million observations, where the average pay and pay raise is $106,100 and 4.6%, respectively. Thus, rulemakers are more senior and earn substantially higher compensation compared to the average federal employee.

2 Model

In this section we introduce a simple model of a rulemaking project. The model seeks to understand how various frictions affect the team's ability to complete the project, that is, to push the rule proposal through the pipeline and convert it into a full rule. We will use the model to investigate how female inclusion affects the likelihood of completion, relying on reduced-form evidence (Section 3) and structural estimation (Section 4).

The principal of the regulatory agency considers a new rulemaking project i. Initially, the perceived net benefit of the project is x_i. It is a function of preliminary features such as the rule's topic and the industry it primarily applies to, before the rulemaking team is selected. The ensuing rulemaking process has three stages. In the first stage, project i is assigned to a team of rulemakers, denoted by T_i. In the second stage, the team exerts collective effort to develop the initial idea and fill in the details. This includes the specific regulatory requirements, potential penalties for violators, and various exemptions. For simplicity, we model team effort as being exerted by a representative rulemaker $j \in T_i$, who chooses effort to maximize their expected salary:

$$\max_{e_j} \omega_j e_j - \frac{\gamma_j}{2} e_j^2,$$

where ω_j is pay-to-effort sensitivity, that is, the sensitivity of government pay to the effort,

and γ_j is the cost of effort. Given this formulation, the optimal effort is given by:

$$e_j^* = \frac{\omega_j}{\gamma_j}, \tag{1}$$

such that effort increases with pay-to-effort sensitivity ω and decreases with cost of effort γ.

In the third and final stage, the principal evaluates the rule draft the team has prepared. The evaluation contains two components. The first is a systematic assessment, where the principal evaluates the project's benefit based on the draft and on the characteristics of the assigned team, such as their experience and past performance. We label the systematic component as $\delta_{i,j}$. The second component is idiosyncratic and cannot be predicted ex-ante, such as unexpected pushback from the public. We label this as $\epsilon_{i,j}$, a random shock drawn from a type I extreme value distribution, with zero mean and unit standard deviation.

At the end of the rulemaking process, the net benefit of the project is given by:

$$u_{i,j} = x_i + e_{i,j} + \delta_{i,j} + \epsilon_{i,j}, \tag{2}$$

where x is the rule's baseline value, e is the rulemaking effort, δ is the systematic component of the principal's evaluation, and ϵ is the random shock. The principal's decision depends on the net benefit of the project, $u_{i,j}$: he or she will approve the rule if and only if $u_{i,j} > 0$, that is, when the rule's benefits outweigh the costs. Note that $u_{i,j}$ includes both the societal costs and benefits of the rule, as well as the principal's private costs and benefits, such as reputation and personal preferences. In other words, we do not take a stand in this paper on how regulatory capture affects the principal's behavior.

In sum, the expected likelihood that a rulemaking project will be completed and result in a full rule is:[19]

$$p_{i,j} = \mathbb{E}\left[u_{i,j} \geq 0\right] = \frac{\exp(x_i + e_j + \delta_{i,j})}{1 + \exp(x_i + e_j + \delta_{i,j})}. \tag{3}$$

[19]The expectation is taken with respect to the idiosyncratic shock $\epsilon_{i,j}$.

3 Empirical evidence

In this section, we study whether including females in the rulemaking process affects completion rates. Early evidence is in Table A.1, where we compare project outcomes among all-male teams versus female-inclusive teams. The former includes teams where no female is present, and the latter includes teams where at least one member is female. We find that the completion rate is 71.6% for all-male teams and 69.9% for female-inclusive teams. Equivalently, the termination rate is 28.1% for all-male teams and 30.1% for female-inclusive teams. Thus, female inclusion is associated with lower completion rates and higher termination rates in the rulemaking process. In what follows we investigate the causal relation between female-inclusion and project completion and potential channels.

3.1 Baseline result

We start by estimating the following OLS regression in the cross-section of rules:

$$Completion_{r,a,t} = \alpha + \beta \cdot Female + \lambda + \epsilon, \tag{4}$$

where $Completion_{r,a,t} = 1$ if rule proposal r, initiated by agency a at time t, was completed, and $Female = 1$ if the rulemaking team includes at least one female. Here, β represents the average difference in completion rates between all-male teams ($Female = 0$) and female-inclusive teams ($Female = 1$). Since female inclusion is not random, an omitted variable could determine the project's outcome and the decision to include a female on the team, leading to a biased estimate of β. For instance, perhaps female regulators are matched to low-quality projects which are less likely to be completed. To rule out some of those concerns, we augment Equation (4) with fixed effects for agency, tier of significance, and cohort (year in which the proposal was initiated).

The results are reported in Table 2. In column 1, the univariate correlation between completion and female-inclusion is -2.0. This means that including a female in the team

reduces the completion rate by 2 percentage points, which is 2.9% of the sample mean. In the next three columns we add fixed effects for agency, tier of significance, and cohort. In the tightest specification we compare rules initiated by the same agency at the same time and with the same level of significance. The magnitude is somewhat lower, 2.1% relative to the mean, and statistically significant at the 1% level. This saturated fixed effects design helps rule out several sources of confounding factors: it accounts for agency-specific practices, cohort-wide shocks, and differential treatment of economically or politically salient rules. In other words, the negative association we observe is not driven by differences in agency culture, macroeconomic or political conditions at the time of initiation, or systematic variation in how important rules are assigned and handled.

In sum, all-male teams are more likely to complete their projects, compared to inclusive teams with females onboard. This relation remains robust when we condition on known contributors to rule completion, suggesting that the gender gaps are not just an artifact of agency culture or differences in rule type. Looking at the model-implied Equation (3), the observed gender gap can be attributed to four non-mutually exclusive sources. First, the principal can select all-male teams for projects with a higher perceived net benefit (x_i), in ways that are not captured by the above fixed effects. We refer to this as the selection channel. Additionally, males may have lower costs of effort (γ) or stronger incentives to exert effort (ω), which we refer to as the costs channel and the incentives channel respectively. Either way, they exert more collective effort (e_i), conditional on the initial project quality. The fourth possibility is that the principal evaluates all-male projects more favorably (δ_i), conditional on the initial project quality and on the ensuing team effort. We refer to this as the evaluation channel. We turn to examine those possibilities in the next sections.

3.2 Selection

In this section, we investigate how selection contributes to the observed gender gaps. Through the lens of the model, the question is whether all-male teams complete more projects in

part because their projects have a higher ex-ante perceived net benefit (x_i).[20] In an ideal experiment, we would assign different teams to projects with identical x_i values, and compare their completion rates. While we cannot conduct this experiment literally, for obvious reasons, we leverage a large language model known as BERT (Bidirectional Encoder Representations from Transformers) to capture contextual meaning in the text. We recognize that this approach does not perfectly capture the rule's initial quality, and we discuss those limitations in the ensuing sections.

3.2.1 Methodology

Our goal is to see if male teams are assigned to more promising projects. We focus on the original project description, before any team had a chance to exert effort and develop a full draft, and test whether females are assigned to less-promising projects. We first convert text into a high-dimensional representation using BERT, a model capturing contextual relationships between words by learning each word's meaning based on its surrounding text. It was pre-trained by Google on a large corpus of unlabeled text, and we fine-tune it using the federal rule proposals from our database. Importantly, we utilize only the text available when the proposal was first announced, which includes its title and abstract. At the end of this step, the rulemaking project description is represented by a high-dimensional vector. Finally, we train a neural network model predicting the probability that the project will be completed. The model is gender blind, meaning that it does not take the identity or gender composition of the assigned team as an input. Instead, it relies solely on the original project description and other observable project characteristics available at the time of assignment. This ensures that any differences we observe in project quality across teams are not mechanically driven by gender differences in observed outcomes during training.

At the end of this process we obtain the object of interest, *Completability*. It represents

[20] Note that the team composition is based on self-selection by rulemakers (bottom-up) and assignment by the principal (top-down). We do not distinguish between those mechanisms and jointly refer to them as the selection channel.

the predicted probability that the rule will be completed, based solely on the initial project description and learned patterns from the historical data. This measure serves as the empirical counterpart of the theoretical x_i, the ex-ante project quality prior to any effort exerted by the assigned team.

3.2.2 Results

With the new variable at hand, we study whether all-male teams are assigned to projects with higher ex-ante likelihood of completion, compared to female-inclusive teams. Initial evidence is in Table A.1, where we report differences in mean outcomes between the two type of teams. Focusing on *Completability*, we find that the average value is 73.7% for female-inclusive teams, compared to 75.8% for all-male teams, and the 2.1% difference is statistically significant. In other words, females appear to be assigned to rules with lower completability prospects.[21] To formally test this, we re-visit our OLS specification (Equation (4)):

$$Completability_{r,a,t} = \alpha + \beta \cdot Female_{r,a,t} + \lambda + \epsilon, \tag{5}$$

where $Completability_{r,a,t}$ is the initial probability of completion estimated before the team has exerted any effort. As before, the focus is on β and we use the same set of fixed effects, to control for the rule's agency, cohort, and level of importance. The results are in Table 3. Across all specifications, we again find significant differences between genders. Concretely, female-inclusive teams are assigned to projects with 0.9-2.1 percentage point lower probability of completion, which is 1.2-2.8% of the mean outcome. In other words, female-inclusive projects have a lower probability of converting into a final rule, based on the initial project description and before any effort has been chosen.

In sum, we find that female-inclusive teams are assigned to rules with lower ex-ante probability of completion. This disparity is based on the initial linguistic characteristics of

[21] Note that the predicted probability is close to the realized outcomes, which is 69.9% for female-inclusive teams and 71.6% for all-male teams. Moreover, the model performs well in terms of accuracy (76.2%), precision (76.3%), recall (91.2%) and F1 Score (83.1%).

the project. These findings do not rule out the possibility that other factors – aside from the project's initial features – may contribute to the observed gender gap. In particular, we will study the role of team effort and principal evaluation (e and δ in our model).

3.3 Effort

In the model, effort is a key contributor to the successful completion of the rulemaking project (Equation (1) and Equation (3)), and it is determined by balancing the marginal cost of effort (γ_j) against the performance incentives (ω_j). Thus, it is possible that male rulemakers face lower cost of effort and/or stronger performance incentives, resulting in more effort. In this section we investigate this channel and our results are summarized in Table 4. Since effort is unobserved, we turn to a variety of empirical proxies.

3.3.1 Textual revisions

We begin with a proxy for effort based on a comparison of the rule drafts in different development stages. We focus on a sample of 12,804 rule proposals where we have at least two drafts, and we use an unsupervised LDA algorithm to identify 100 regulatory topics and obtain the fraction of each draft dedicated to each topic. We then define our proxy for effort as:

$$Revisions_i = \frac{1}{100} \sum_{k=1}^{100} \left| TopicShare_{i,k}^{\text{final}} - TopicShare_{i,k}^{\text{initial}} \right|, \tag{6}$$

where $TopicShare_{i,k}^{\text{stage}}$ denotes the proportion of rule i's text dedicated to topic k in the indicated draft stage (initial or final). *Revisions* captures the average change in topic composition between the early and final versions of the rule proposal. It serves as a proxy for rulemaking effort: greater textual shifts across drafts suggest more substantive work in refining and developing the proposal. By summarizing the mean absolute change across 100 regulatory topics, this measure avoids dependence on any particular subject matter and reflects overall textual development.

We estimate our baseline specification (Equation (4)) with the same set of controls and fixed effects, except that the outcome is *Revisions*. The results are in Table 4, Panel A. In column 1, female-inclusive teams change the textual distribution of LDA topics by 0.078, which is 16.8% of the sample mean. Columns 2-4 gradually add tighter fixed effects, and the final estimates point to 9.5% more textual revisions in female-inclusive teams. These findings suggest that female-inclusive teams exert more effort than all-male teams, based on observed differences in revisions throughout the rulemaking process.

3.3.2 Intermediary milestones

As an alternative approach, we estimate Equation (4) with additional outcomes of the rulemaking process. Concretely, we define $Final = 1$ ($Interim = 1$) if the proposal yielded a final rule (interim rule). We define $Draft = 1$ if a full draft was published for comments, regardless of whether a final or interim rule was approved. We also define $Suspended = 1$ if the rule was temporarily put on hold, and $Cancelled = 1$ if the rule was officially withdrawn from the rulemaking pipeline. Finally, *Duration* measures the time (days) the rule proposal was in the rulemaking pipeline, regardless of how it exited the pipeline. The results are in Table 4, Panel B. We find that female-inclusive teams are 3% less likely to complete a final rule, but more likely to complete an interim rule (14.6%) or a draft (5.7%). They are 23.3% more likely to announce a temporary suspension and 6.2% more likely to be officially terminate the project. Consequently, their projects spend significantly more time in the rulemaking pipeline.

In sum, female regulators do not appear to be less active or less productive. Female-inclusive projects are significantly more likely to reach the draft or interim stage, but also more likely to be paused or withdrawn, and tend to remain in the regulatory pipeline longer. Looking at the shifts in topic composition between initial and final versions of a rule proposal, we find that female-inclusive projects exhibit significantly greater textual revisions. Thus, the observed gap in completion rates seems less likely to be driven by lower effort

among female rulemakers.

3.3.3 Incentives

In the model, stronger incentives (ω_j) yield greater effort. Thus, it is possible that male rulemakers receive better incentives, resulting in greater effort and hence higher completion rates. Since we do not directly observe effort, we will study how completion of rulemaking projects relates to career outcomes. The results are reported in Table 4, Panel C. We begin by looking at the internal rewards for project completion, and we estimate the following panel regression:

$$y_{i,t} = \beta_1 \cdot CompletionRatio_{i,t-1} + \beta_2 \cdot CompletionRatio_{i,t-1} \cdot \mathbb{1}(Female_i) + \lambda, \qquad (7)$$

where the outcome is the rulemaker's pay or log of pay at time t, and $CompletionRatio$ is the fraction of rulemaker i's projects that were completed up to year $t-1$. $Female$ is an indicator for a female rulemaker. The coefficient β_1 represents the average impact of rule completion on pay, and the coefficient β_2 represents the differential impact on females. We control for the total number of projects up to year $t-1$ and for salary in year $t-1$. We add the interaction of year, agency, location, and occupation fixed effects. This ensures we compare two rulemakers with opposite genders who hold the same occupation and work at the same time and place and for the same agency. For instance, two SEC attorneys working in 2005 from the agency's headquarters in Washington, DC. We also include person fixed effects, which subsume the indicator $\mathbb{1}(Female_i)$, thus focusing on pay raises within-employee over time as a function of their rulemaking success.

The results in column (1) show a significant positive association between rule completion and pay. For male rulemakers, a 100% completion rate is associated with an increase in annual pay of \$2,480 relative to a 0% completion rate. For female rulemakers, the premium is substantially larger: the total effect is \$4,489 (\$2,480 + \$2,009). Scaling by

a one-standard-deviation increase in completion rate (45%), this translates into roughly a $1,116 increase for males and a $2,020 increase for females. In column (2), where the dependent variable is the natural log of pay, the coefficients imply that a 100% completion rate is associated with a 1.3% pay increase for males and a 3.9% increase for females. A one-standard-deviation increase in completion rate corresponds to a 0.59% increase for males and a 1.76% increase for females. In both specifications, the pay-for-performance sensitivity is notably higher for females: the implied female-to-male ratio of the completion premium is about 1.8 in column (1) and about 3.0 in column (2).

For completion we study the external reward, namely, obtaining a private sector job. This is motivated by the rich literature linking regulatory actions to the outside option. Our dataset does not cover career trajectories in the private sector, but we do observe the employee's decision to quit and anecdotal evidence suggests that forced removal is extremely rare in the federal government. Thus, we follow Kalmenovitz, Vij, and Xiao (2025) and view the exit decision as a proxy for obtaining a private sector job. We estimate the following specification in the cross-section of rulemakers:

$$Exit_i = \beta_1 \cdot CompletionRatio_i + \beta_2 \cdot CompletionRatio_i \cdot \mathbb{1}(Female_i) + \lambda, \tag{8}$$

where $Exit_i = 1$ if rulemaker i left their government post and $CompletionRatio$ is the fraction of rulemaker i's rulemaking projects successfully completed throughout their entire government career. Similar to the panel regression, we add the interaction of cohort, agency, location, and occupation fixed effects. Effectively, we compare the exit decision of two rulemakers who started their government career at the same time, working for the same agency at the same city, and hold a similar occupation. For instance, two attorneys hired by the SEC in 2000 and work from its DC headquarters. We are interested in seeing whether exit relates to project completed (β_1) and whether there are gender gaps in the exit-to-rulemaking sensitivity (β_2). Note that the sign of β_1 can be either positive or negative, related to the

broader question whether the outside option motivates regulators to decrease regulatory burden ($\beta_1 < 0$) or increase it ($\beta_1 > 0$).[22]

The results are in column 3, showing a significant positive association between rule completion and exits. In particular, a 100% completion rate increases exit probability by 62 percentage points, compared to 0% completion rate, and a one-standard-deviation (45%) increase in completion rate is associated with 28 percentage point greater likelihood of exit. More importantly, we find no gender differences, as the coefficient β_2 is statistically insignificant. In other words, rulemakers who complete their projects are more likely to resign, whether they are males or females. We further estimate Equation (8) with the final number of years in government service as the outcome. It is inversely related to exits, as employees who choose to exit cut their tenure short. Here, we find that a one-standard-deviation increase in completion rate is associated with a 6-year shorter tenure. Again, we find no differences between males and females in their choice to exit the government.

In sum, the completion of rulemaking projects is associated with greater pay (internal reward) and with more exits from the public sector (external reward). More relevant to our paper, female rulemakers are rewarded similarly (externally) or better (internally) than their male counterparts. It suggests that the return on effort for females is either similar or larger than the return on effort for males. This is consistent with our earlier findings, whereby females put more effort into their actions, while reinforcing the question why female-inclusive teams have lower completion rates than all-male teams.

3.4 Evaluation

In this section, we investigate how evaluation contributes to the observed gender gaps. Through the lens of the model, the question is whether all-male teams are more likely to receive approval from their superiors, conditional on the initial project quality and the effort

[22]These predictions loosely correspond to the regulatory capture (Stigler, 1971; Peltzman, 1976; Shleifer and Vishney, 1993) and regulatory schooling (Bond and Glode, 2014; Bar-Isaac and Shapiro, 2011) theories. Empirical findings point in different directions (e.g., Duchin, Kalmenovitz, and Wang (2024) versus Tabakovic and Wollmann (2018).

exerted by the team. Although we do not observe the evaluation process directly, we again leverage the power of the BERT model. While this analysis and that in Section 3.2 both rely on BERT, the estimations are conducted independently, and the resulting outcomes have distinct interpretations, as we explain below.

3.4.1 Methodology

Our goal is to test whether male teams receive more favorable treatment from agency leadership when their draft rules are reviewed. In an ideal experiment, the agency would evaluate two identical drafts produced by different teams – one male, one female-inclusive – and we could directly compare their approval rates. While such an experiment is not feasible, we approximate it using C-BERT, a causal variant of BERT designed to estimate counterfactual outcomes. Intuitively, we focus on the draft text itself, "swap" the team's gender across otherwise comparable drafts, and measure whether male teams are still more likely to complete their projects.

Our approach follows the application in Hochberg et al. (2023) and Veitch, Sridhar, and Blei (2020), with details provided in Appendix A.3. Here we summarize the main steps. Unlike in Section 3.2, we now restrict attention to projects where the team has produced at least one draft rule, reducing the sample from 37,525 to 20,526 projects. We further exclude drafts with strongly gendered language, keeping only those that could plausibly have been written by either male or female-inclusive teams. Specifically, we use a neural network to estimate the probability that a draft was written by a female-inclusive team, and drop the 2,258 projects in the top and bottom 5% of this distribution. In this subsample, the completion rate conditional on having at least one draft is naturally higher than in the full sample (90.1% vs. 70.8%). We then convert the full text of each draft – not just its title and abstract – into high-dimensional embeddings using BERT to capture the contextual relationships between words. Next, we train two separate neural network models to predict the probability of project completion: one using only all-male projects (the male model) and one

using only female-inclusive projects (the female model).[23] We then pass each female-inclusive draft through the male model, and each all-male draft through the female model. This yields $\hat{p}_r$, the counterfactual probability that the project would be completed if the same draft had been produced by a team of the opposite gender.

Finally, we compute the tilt score for project r, defined as:

$$Tilt_r = Completion_r - \hat{p}_r, \tag{9}$$

where $Completion_r$ is the realized project outcome and $\hat{p}_r$ is the counterfactual probability of completion with the opposite gender. The difference between the two, $Tilt$, measures the extent and direction of the agency's evaluation of a draft relative to the counterfactual benchmark. A positive value ($Tilt > 0$) indicates that the draft received more favorable treatment than predicted for the opposite gender, while a negative value ($Tilt < 0$) implies less favorable treatment. The absolute value $|Tilt|$ reflects the strength of that departure: larger magnitudes correspond to more decisive evaluations.

To illustrate, consider two projects: one with a high counterfactual probability of completion ($\hat{p}_i = 0.7$), and another with a low probability ($\hat{p}_j = 0.2$). For rule i, if it ultimately failed, $Tilt_i = -0.7$; if it succeeded, $Tilt_i = 0.3$. For rule j, if it failed, $Tilt_j = -0.2$; if it succeeded, $Tilt_j = 0.8$.

	$Tilt$	
	$Completion = 1$	$Completion = 0$
Project i ($\hat{p} = 70\%$)	30%	-70%
Project j ($\hat{p} = 20\%$)	80%	-20%

Example. This table illustrates possible values of $Tilt$ for two rules with different counterfactual probability of completion ($\hat{p}$).

[23] The model demonstrates strong performance in predicting outcomes conditional on gender, achieving an accuracy of 86.1%, a precision of 86.2%, a recall of 99.9%, and an F1 score of 92.5%.

As the table shows, both the sign and the magnitude of $Tilt$ provide meaningful information on the evaluation stage. For example, in the bottom-left cell, rule j had a low counterfactual probability (20%) yet was completed, implying a strongly favorable evaluation ($Tilt = 80\%$) that pushed the draft over the finish line.

3.4.2 Main result

With the $Tilt$ measure in hand, we examine whether female-inclusive teams receive less favorable treatment at the final evaluation stage of the rulemaking process. As an initial step, Table A.1 reports mean $Tilt$ scores by team type. The average $Tilt$ score is substantially higher for all-male teams (8.3%) than for female-inclusive teams (3.1%), and the difference is statistically significant. This pattern suggests that, conditional on producing a draft, male teams receive systematically more favorable assessments from agency leadership relative to the counterfactual benchmark, pointing to a gender gap in the evaluation stage of the rulemaking process. To formally assess this gap, we return to our OLS specification in Equation (4):

$$Tilt_{r,a,t} = \alpha + \beta \cdot Female + \lambda + \epsilon, \tag{10}$$

where $Tilt_{r,a,t}$ measures the principal's assessment of the team's draft rule, as defined in Equation (9). As before, our coefficient of interest is β, and we include the same set of fixed effects to control for the rule's agency, cohort, and level of importance. The results, reported in the first four columns of Table 5 reveal significant differences between team types. Female-inclusive teams exhibit $Tilt$ scores that are 4.4–5.3 percentage points lower than those of all-male teams, corresponding to 4.9–5.9% of the mean outcome. Put differently, even after conditioning on the linguistic content of the draft, female-inclusive teams receive systematically less favorable evaluations from agency leadership, making their projects significantly less likely to be approved and advance to final rule status.

To ensure that our findings are not driven by specific modeling choices or correlated team

characteristics, we conduct several robustness checks, reported in Table A.2. First, we replace the binary female-inclusive indicator with the share of females in the rulemaking team. The estimated effect of female representation on $Tilt$ ranges from -2.6 to -1.7 percentage points, indicating that both the extensive and intensive margins of female inclusion are associated with lower $Tilt$ scores. Second, we examine whether team frictions contribute to worse evaluations – for example, if larger teams present a less cohesive case to their superiors and are therefore assessed less favorably. Females are significantly more likely than males to serve on multi-member teams (34% vs. 37% working solo), a pattern that is notable given the over-representation of males in the pool of rulemakers (see Section 1.2). To test this, column 3 of Table A.2 adds a control for team size. While larger teams do tend to have lower $Tilt$ scores, our main measure of female inclusion remains negative and statistically significant. Finally, we distinguish between solo projects (single rulemaker) and team projects (multiple rulemakers). For solo projects, the female-inclusive indicator equals 1 if the sole rulemaker is female; for team projects, it equals 1 if at least one member is female. Results in columns 4–5 show that the gap in $Tilt$ is especially pronounced in team projects (-4.7% of the mean outcome) compared to solo projects (-1.6% of the mean outcome).

Taken together, these results show that female-inclusive teams receive systematically lower tilt scores than all-male teams, even after holding constant the observable characteristics of the draft proposals they produce. This gap persists across specifications and robustness checks, and is especially pronounced in collaborative projects. Because we control for the full linguistic content of each draft, capturing the team's output prior to evaluation, these differences cannot be readily attributed to project selection or team performance. Instead, the evidence points to differential treatment in the final stages of the rulemaking process.

3.4.3 Additional results

In this section we present additional results related to the evaluation channel.

We first ask whether observable team characteristics – such as experience, seniority,

or location – can account for the differences between female-inclusive and all-male teams. If female-inclusive teams systematically differ along these dimensions, then some of the evaluation gap may reflect variation in team capacity or access to institutional resources rather than differential treatment per se. To examine this, we focus on the subset of rule proposals developed by individuals we successfully matched to federal payroll records (Section 1.1). For each team, we compute the average annual pay in the year of the proposal, the share of members located in Washington, DC, and average tenure measured as the difference between the rule's cohort (year of proposal) and the average hiring year of team members. In addition, we account for the composition of the team by rank and occupation, based on their job classifications at the time of proposal.[24] These variables proxy for human capital, institutional proximity, and bureaucratic seniority, all of which may plausibly affect the team's ability to navigate the evaluation process.

We then re-estimate our evaluation regression (Equation (10)), sequentially adding these controls alongside a full set of fixed effects. The results, reported in Table 6, show that the baseline evaluation gap in this matched subsample is 4.1 percentage points (4.5% of the sample mean). Importantly, the gap remains virtually unchanged after controlling for all available team characteristics. This finding reinforces our earlier conclusion: female-inclusive teams receive significantly lower evaluations even when we hold constant a rich set of observable team attributes. The persistence of the gap suggests that unobserved frictions – such as differential scrutiny in the evaluation stage – may play a role in limiting the success of female-inclusive teams at the final stages of rulemaking.

We separately estimate Equation (10) across various subsamples. In Figure 1, we study heterogeneity by project cohort. Each point represents the coefficient from a separate regression on each cohort, where cohort is defined by the year in which the project was initiated. The results show that the estimated gender effect is consistently negative across

[24]We suppress the coefficients on ranks and occupations, since there are 186 unique ranks and 192 unique occupations in the data. The numerical representation of ranks and occupations does not have any natural interpretation, e.g., rank 1 is not necessarily junior to rank 2.

cohorts, with nearly all point estimates falling below zero and many statistically significant. In Figure 2, we study heterogeneity by regulatory topic. We rely on data from Chang et al. (2023), who decompose each rule proposal in the Unified Agenda into 100 topics using a Latent Dirichlet Allocation (LDA) algorithm.[25] For tractability, we retain only the topic with the highest weight in each rule proposal. To focus on common topics, we further restrict the sample to topics that appear in at least 500 proposals. We then estimate Equation (10) separately for each eligible topic and present the estimated coefficients. In many topics, particularly those related to taxation, legal operations, and banking, female inclusion is associated with significantly lower impact on project completion probability. In contrast, few topics show a positive or null effect. These patterns suggest that gender disparities in rulemaking outcomes are not uniform but vary considerably by policy domain, with some areas exhibiting sharper penalties for female participation. Finally, Figure 3 shows variation across federal agencies. While most agencies exhibit negative coefficients, indicating lower evaluation of female-inclusive teams, several estimates are small or even positive. This heterogeneity suggests that institutional context shapes the relationship between gender composition and rulemaking performance.

3.5 Summary of reduced-form findings

In our model, the likelihood of completing a rulemaking project is based on initial project value (x_i); team effort (e_i), itself a function of costs (γ) and incentives (ω); and principal evaluation (δ_i). The reduced-form analysis shows that completion rates are lower for female-inclusive teams. Females are assigned to rules with lower ex-ante probability of completion, and the drafts they present receive worse evaluation from the agency's leadership. However, females appear to be better rewarded for completing their projects, and female-inclusive teams appear to exert more effort, at least in the earlier stages of the rulemaking process. Combined, these

[25] An increasingly popular tool in the financial economics literature, LDA has been used for instance to study asset pricing puzzles (Bao and Datta (2014)), risks in the financial sector (Hanley and Hoberg (2019)), and FOMC speeches (Hansen, McMahon, and Prat (2018)).

findings support the interpretation that the lower completion rates are driven primarily by worse selection and evaluation, and less so by gaps in effort. To be clear, we cannot rule out – in the reduced-form analysis – the possibility that females exert less effort, especially in the latter stages of the rulemaking process. We will continue to investigate these channels in Section 4 using structural estimation.

4 Quantitative implications

In Section 3, we document the gender gap in completion rates and discuss how it relates to selection, costs, and incentives. To gain better insight, we now turn to the structural estimation of our model from Section 2. This estimation will help quantify the key frictions which yield different completion rates between the genders. Looking at Equation (3), several possibilities emerge. *First*, females may be assigned to rule proposals with lower ex-ante perceived benefit (low x_i). *Second*, females may exert less effort (lower e_j). This can be driven by higher cost of effort (higher γ_j) or weaker compensation incentives (lower ω_j), either one reducing the marginal benefit of effort. *Third*, the principal may evaluate the work of female rulemakers more harshly (low δ_j).

4.1 Estimation

We seek to estimate four structural parameters: the initial net benefit of the rule x, the pay-to-effort sensitivity ω, the cost of effort γ, and the principal evaluation δ. The first parameter varies by rule, while the last three parameters vary by gender and are estimated for male and female rulemakers separately. We start with three data moments, describing the probability that rulemaking project i assigned to rulemaker j will be completed based on the information available at this point. The first moment, $p_{i,\bar{j}}$, is the completion probability before the proposal was assigned to any rulemaker. It is identical to the empirical *Completability* in the reduced-form analysis. Here, $\bar{j}$ means that the expectation is taken across all possible

rulemakers. To obtain it, we estimate the BERT model based only on the content available at this early stage (title and abstract), and using the full sample of both males and females. We compute the implied net benefit of the rule as:

$$y_{i,\bar{j}} \equiv x_i + e_{\bar{j}} + \delta_{i,\bar{j}} = \log\left(\frac{p_{i,\bar{j}}}{1 - p_{i,\bar{j}}}\right). \tag{11}$$

The second moment, $p_{i,j}$, is the completion probability after the proposal is assigned to rulemaker j and after effort has been exerted. The estimation is conducted based on the rulemaker's true gender and the BERT model trained on the text of the full rule drafts.[26] Given this probability, we compute the implied net benefit of the rule as:

$$y_{i,j} \equiv x_i + e_j + \delta_{i,j} = \log\left(\frac{p_{i,j}}{1 - p_{i,j}}\right). \tag{12}$$

The third moment is the counterfactual probability where the gender is switched, $p_{i,-j}$, where $-j$ means the opposite gender to rulemaker j's gender. It is identical to the empirical $\hat{p}_r$ in the reduced-form analysis. The estimation is identical to the one for $p_{i,j}$, except that we use the switched gender. We can similarly compute the implied net benefit of the rule as:

$$y_{i,-j} \equiv x_i + e_j + \delta_{i,-j} = \log\left(\frac{p_{i,-j}}{1 - p_{i,-j}}\right). \tag{13}$$

With those moments in hand, we turn to the four structural parameters. To estimate pay-to-effort sensitivity ω, recall that we observe pay and projects (rules finalized), but we do not observe effort. We overcome this challenge based on the model-implied relation between the rulemaker's salary s, completion probability p, and effort e:

$$\omega = \frac{\partial s}{\partial e} = \frac{\partial s}{\partial p}\frac{\partial p}{\partial e} = p(1-p)\frac{\partial s}{\partial p}, \tag{14}$$

where the last term, $\frac{\partial s}{\partial p}$, is the empirical salary-to-completion sensitivity which we discussed

[26] This sample is smaller than the one used to estimate $p_{i,\bar{j}}$, since not all rule proposals reach this stage.

in Section 3.3.3. It is based on estimating Equation (7) in a panel of rulemakers. The results in Table 4, Panel C, show that pay-to-success sensitivity is \$2,480 for males versus \$4,489 for females (β_1 and $\beta_1 + \beta_2$, respectively). Multiplying these numbers by $p(1-p)$ yields the pay-to-effort sensitivity ω for males and females.

To estimate the principal evaluation δ, we normalize the male valuation to zero and assume that the female valuation is a constant parameter and does not vary across rules. Then, we calculate the female δ as:

$$\hat{\delta} = \frac{1}{N} \sum_i \left((y_{i,j} - y_{i,-j}) \mathbb{1}[j \in female] + (y_{i,-j} - y_{i,j}) \mathbb{1}[j \in male] \right), \tag{15}$$

which is the average difference between the net benefit of the rule, based on the true gender ($y_{i,j}$) and the counterfactual one ($y_{i,-j}$). Using this notation, we can express the principal evaluation of rule i as:

$$\delta_{i,j} = \delta_j = \delta \mathbb{1}[j \in female]. \tag{16}$$

Note we drop the subscript i, because the principal evaluation is assumed to be a function of the rulemaker's gender and does not interact with the content of the rule.

To identify the cost of effort (γ), we compare the net benefit of the rule prior to team selection ($y_{i,\bar{j}}$) with the net benefit after effort is exerted ($y_{i,j}$), thereby differencing out the initial net benefit x_i:

$$y_{i,j} - y_{i,\bar{j}} = (e_j + \delta_j) - (e_{\bar{j}} + \delta_{i,\bar{j}}) \tag{17}$$

Rearranging, we derive an empirical measure of the optimal effort:

$$\hat{e}_j = y_{i,j} - y_{i,\bar{j}} - \delta_j + (e_{\bar{j}} + \delta_{i,\bar{j}}) = y_{i,j} - y_{i,\bar{j}} - \delta_j + \alpha_j, \tag{18}$$

where the last term $\alpha_j \equiv e_{\bar{j}} + \delta_{i,\bar{j}}$ is a structural residual. We then choose the values of γ and α that minimize the distance between the model's predicted effort based on Equation (1)

and the empirical counterpart in Equation (18):

$$\min_{\gamma_j,\alpha_j} \sum_{i=1}^{N} (e_j - \hat{e}_j)^2 = \min_{\gamma_j,\alpha_j} \sum_{i=1}^{N} \left(\frac{\omega_j}{\gamma_j} - (y_{i,j} - y_{i,\bar{j}} - \delta_j + \alpha_j) \right)^2. \tag{19}$$

We estimate these parameters for the subsamples of rules written by males and females, separately. Then, with the estimated cost of effort γ, we can predict the implied effort e from Equation (1). Finally, combining γ with the principal evaluation δ, we calculate the perceived net benefit of the rule at the first stage, before it was assigned to any rulemaker, as:

$$x_i = y_{i,j} - e_j - \delta_j. \tag{20}$$

4.2 Baseline results

Table 7 presents the baseline parameter estimates for male and female rulemakers. Several key patterns emerge. First, females work on rule proposals with 7% lower initial net benefits: the average x is 1.416 for females, compared to 1.515 for males. This demonstrates how female rulemakers are more likely to be tasked with lower-quality projects, that is, projects with a lower perceived benefit. This selection mechanism drives down the likelihood of female-inclusive teams to complete their rulemaking projects.

Second, we observe a stark contrast between the two effort parameters. While the pay-to-effort sensitivity ω is 44.8% higher for females (0.620 vs. 0.343), their cost of effort γ is 44.4% higher (2.272 versus 1.262). This creates a genuine trade-off, as females must balance the higher cost of effort against the greater return to effort. On balance, we find that the benefit of effort outweighs its costs, resulting in slightly more effort from females (0.273 vs. 0.272) based on Equation (1). This is consistent with our reduced-form analysis (Section 3.3), where we find that females appear to exhibit greater effort in the earlier stages of the rulemaking process. The empirical proxies capture realized behavior in the early stages of the drafting process, while the structural estimation represents the expectation for total

effort throughout the entire process. Combined, it suggests that females put more effort into their actions, which reinforces the question as to why female-inclusive teams have lower completion rates than all-male teams.

Third, the principal's evaluation of the rulemaker's output δ shows the largest gender gap. For male rulemakers, the normalized $\delta = 0$, while for females, it is negative and sizable ($\delta = -0.166$). This implies that, holding the rule's perceived quality (x) and the team's effort (e) constant, the principal systematically undervalues proposals developed by females. In the context of the model, this subjective penalty amplifies the gender gap in completion probabilities. Note that we do not take a stand on the root causes for the differences in principal evaluation. It is possible, for instance, that a doom loop emerges. Suppose that females are assigned to low-quality rules and thus exhibit lower completion rates. This can contribute to lower evaluation of females by their principals in the future, leading to even worse assignment and then even lower completion rates. It is also possible that evaluation and selection gaps arise endogenously, as a response to diversity-promoting efforts which assign greater reward per unit of effort to underrepresented groups.

Together, the structural estimation highlights how multiple structural frictions – selection, effort, and evaluation – contribute to the observed gender disparities in rulemaking outcomes. They are largely consistent with the empirical counterparts we estimate in the reduced-form analysis, while quantifying the unobserved aspects of the rulemaking process which cannot be otherwise measured.

4.3 Counterfactual policies

In this section we simulate four counterfactual scenarios, each designed to isolate a specific mechanism in the structural model. Concretely, we start with male parameters as the baseline and change them sequentially to the female equivalents. These exercises will help us quantify the contribution of various factors to the observed gender gap in project completion. The results are summarized in Figure 4. The horizontal dash line displays the observed

gender gap of 2.9 percentage points, which is the estimate from the baseline OLS reported in the first column of Table 2. The colored arrows represent estimated contributions from specific channels, where red (blue) indicates a positive (negative) contribution that increases (decreases) the gap in completion rates. These effects accumulate from left to right, and for illustration purposes, the four channels are sorted by magnitude.

We start with the selection channel. To that end, we estimate a scenario where males are matched to rules with initial net benefit x similar to females' rules. The simulation reveals that selection disfavors females and creates a substantial gender gap, which equals 36% of the observed gender gap. It indicates that rule selection is a major source of disparity. Next, we replace the cost parameter (γ) of males with that of females. This simulation reveals that the true differences in cost account for 44% of the observed gender differences. It underscores the role of effort costs in limiting the effectiveness of female rulemakers, consistent with the interpretation that higher effort costs constrain the ability to improve a rule's perceived quality. Next, we replace the principal evaluation δ for males (the baseline zero) with the females' evaluation. This reveals that the true evaluation gap contributed to 61% of the observed gender gap, which is bigger than the cost channel and smaller than the selection channel. It suggests that gender gaps in evaluation is a consequential factor.

The last exercise focuses on the compensation channel. Here, we replace the male pay-to-effort sensitivity ω with that of females. Following the baseline estimation, we can see how the compensation channel helps limit the gender gap, as females have a higher ω than their male counterparts. In particular, it account for 70% of the observed success gap. In other words, the higher ω observed for females in the baseline estimates is a favorable feature of the female compensation structure that partially offsets disadvantages elsewhere. Eliminating this advantage worsens the female-male gap in rule success.

Overall, the decomposition highlights how selection, cost of effort, and principal evaluation are roughly similar contributors to the gender gap in completion rates. Compensation incentives, by contrast, serve as an equalizing force. Thus, while females appear to receive

better rewards for a given outcome, they are also assigned to less favorable projects, pay higher costs per unit of effort, and are evaluated more harshly. This pattern is consistent with a setting where explicit efforts to promote diversity – such as rewarding underrepresented groups – are present, but insufficient to fully offset other frictions.

5 Conclusions

In this paper, we examine how gender diversity among federal rulemaking teams shapes the progression and completion of rulemaking projects. By merging biannual Unified Agenda data with federal payroll records, we construct a panel of 15,000 career bureaucrats working on 38,000 rule proposals. Our reduced-form analysis shows that female-inclusive teams draft more extensively and reach interim milestones more often, yet they are less likely to complete a final rule than all-male teams.

To uncover the mechanisms behind this gap, we develop and estimate a three-stage structural model featuring parameters for initial project value (selection), cost and return to effort, and principal evaluation bias. Structural estimates and reduced-form tests reveal that female rulemakers are assigned to projects with lower perceived net benefit, face higher effort costs, and suffer an evaluative discount from agency leaders, which together outweigh their stronger performance incentives. Counterfactual simulations indicate that equalizing assignments, costs of effort, and evaluations would substantially narrow the completion gap, whereas aligning effort incentives will widen the gap.

These findings suggest that subtle frictions in assignment and evaluation, rather than lack of effort or skill, drive the observed gender disparity in rulemaking outcomes. Implementing blind assignment protocols and standardized evaluation rubrics could substantially narrow the gap in completion rates. Moreover, providing targeted support – such as research assistance or administrative relief – may alleviate the higher effort burdens that female rulemakers seem to incur. Future research could assess the welfare impacts of rules produced by diverse

teams, for example by looking into stakeholder comments and post-implementation outcomes. Extending our structural framework to other dimensions of diversity or to other regulatory settings would help generalize these insights. Finally, investigating how diversity policies affect long-run career trajectories could inform the design of interventions that enhance both equity and regulatory effectiveness.

References

Adams, R. B., and D. Ferreira. 2009. Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics* 94:291–309.

Administrative Conference of the United States. 2015. The Unified Agenda: Proposals for Reform. *Available online* .

Ahern, K. R., and A. K. Dittmar. 2012. The changing of the boards: The impact on firm valuation of mandated female board representation. *The Quarterly Journal of Economics* 127:137–97.

Allen, A., and A. S. Wahid. 2023. Regulating gender diversity: Evidence from california senate bill 826. *Management Science* 70:2023–46.

Azmat, G., and R. Ferrer. 2017. Gender gaps in performance: Evidence from young lawyers. *Journal of Political Economy* 125:1306–55.

Bagues, M., and B. Esteve-Volart. 2010. Can gender parity break the glass ceiling? evidence from a repeated randomized experiment. *The Review of Economic Studies* 77:1303–38.

Bao, Y., and A. Datta. 2014. Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science* 60:1371–91.

Bar-Isaac, H., and J. Shapiro. 2011. Credit ratings accuracy and analyst incentives. *American Economic Review* 101:120–24.

Bertrand, M. 2011. New perspectives on gender. In D. Card and O. Ashenfelter, eds., *Handbook of Labor Economics*, vol. 4B, 1545–92. Elsevier.

Bertrand, M., C. Goldin, and L. F. Katz. 2010. Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics* 2:228–55.

Bond, P., and V. Glode. 2014. The labor market for bankers and regulators. *The Review of Financial Studies* 27:2539–79.

Card, D., F. Colella, and R. Lalive. 2025. Gender preferences in job vacancies and workplace gender diversity. *The Review of Economic Studies* 92:2437–71.

Chang, S., J. Hajda, J. Kalmenovitz, and A. Lopez-Lira. 2023. Follow the pipeline. *Working paper* .

Coles, J. L., M. L. Lemmon, and J. F. Meschke. 2012. Structural models and endogeneity in corporate finance: The link between managerial ownership and corporate performance. *Journal of Financial Economics* 103:149–68.

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

Dal Bó, E., F. Finan, and M. A. Rossi. 2013. Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly Journal of Economics* 128:1169–218.

deHaan, E., S. Kedia, K. Koh, and S. Rajgopal. 2015. The revolving door and the SEC's enforcement outcomes: Initial evidence from civil litigation. *Journal of Accounting and Economics* 60:65–96.

Delfino, A. 2024. Breaking gender barriers: Experimental evidence on men in pink-collar jobs. *American Economic Review* 114:1816–53.

Deschamps, P. 2024. Gender quotas in hiring committees: A boon or a bane for women? *Management Science* 70:7486–505.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–86.

Duchin, R., J. Kalmenovitz, and Y. Wang. 2024. Between boardrooms and the beltway: The career paths of senior regulators. *Working paper* .

Eckbo, B. E., K. Nygaard, and K. S. Thorburn. 2021. Valuation effects of norway's board gender-quota law revisited. *Management Science* 68:4112–34.

Exley, C. L., and J. B. Kessler. 2022. The gender gap in self-promotion. *The Quarterly Journal of Economics* 137:1345–81.

Field, L. C., M. E. Souther, and A. S. Yore. 2020. At the table but can not break through the glass ceiling: Board leadership positions elude diverse directors. *Journal of Financial Economics* 137:787–814.

Gao, J., W. Wang, and Y. Wu. 2024. Human capital portability and careers in finance. *The Review of Financial Studies* 37:2732–78.

Gneezy, U., M. Niederle, and A. Rustichini. 2003. Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics* 118:1049–74.

Gopalan, Y., A. Kalda, and A. Manela. 2021. Hub-and-spoke regulation and bank leverage. *Review of Finance* 25:1499–545.

Gormley, T. A., V. K. Gupta, D. A. Matsa, S. C. Mortal, and L. Yang. 2023. The big three and board gender diversity: The effectiveness of shareholder voice. *Journal of Financial Economics* 149:323–48.

Griffin, D., K. Li, and T. Xu. 2021. Board gender diversity and corporate innovation: International evidence. *Journal of Financial and Quantitative Analysis* 56:123–54.

Hanley, K. W., and G. Hoberg. 2019. Dynamic interpretation of emerging risks in the financial sector. *The Review of Financial Studies* 32:4543–603.

Hansen, S., M. McMahon, and A. Prat. 2018. Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics* 133:801–70.

Hirtle, B., A. Kovner, and M. Plosser. 2020. The impact of supervision on bank performance. *The Journal of Finance* 75:2765–808.

Hochberg, Y. V., A. Kakhbod, P. Li, and K. Sachdeva. 2023. Are patents with female inventors under-cited? evidence from causal text estimation. *Working paper* .

Huang, J., and D. J. Kisgen. 2013. Gender and corporate finance: Are male executives overconfident relative to female executives? *Journal of Financial Economics* 108:822–39.

Kalmenovitz, J. 2021. Incentivizing financial regulators. *The Review of Financial Studies*

34:4745–84.
Kalmenovitz, J., and S. Vij. 2025. Regulatory risk perception and small business lending. *Management Science* .
Kalmenovitz, J., S. Vij, and K. Xiao. 2025. Closing the revolving door. *The Journal of Finance* .
Kerwin, C. M., and S. R. Furlong. 2018. *Rulemaking: How government agencies write law and make policy*. Cq Press.
Leaver, C., O. Ozier, P. Serneels, and A. Zeitlin. 2021. Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from rwandan primary schools. *American Economic Review* 111:1052–88.
Leibbrandt, A., L. C. Wang, and C. Foo. 2018. Gender quotas, competitions, and peer review: Experimental evidence on the backlash against women. *Management Science* 64:3501–16.
Lewellen, K. 2025. Women in charge: Evidence from hospitals. *The Journal of Finance* 80:2199–253.
Lucca, D., A. Seru, and F. Trebbi. 2014. The revolving door and worker flows in banking regulation. *Journal of Monetary Economics* 65:17–32.
Morellec, E., B. Nikolov, and N. Schürhoff. 2012. Corporate governance and capital structure dynamics. *The Journal of Finance* 67:803–48.
Murciano-Goroff, R. 2022. Missing women in tech: The labor market for highly skilled software engineers. *Management Science* 68:3262–81.
Niederle, M., and L. Vesterlund. 2007. Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics* 122:1067–101.
Nikolov, B., and T. M. Whited. 2014. Agency conflicts and cash: Estimates from a dynamic model. *The Journal of Finance* 69:1883–921.
Ors, E., F. Palomino, and E. Peyrache. 2013. Performance gender gap: Does competition matter? *Journal of Labor Economics* 31:443–99.
Page, T. B. 2018. CEO attributes, compensation, and firm value: Evidence from a structural estimation. *Journal of Financial Economics* 128:378–401.
Peltzman, S. 1976. Toward a more general theory of regulation. *The Journal of Law and Economics* 19:211–40.
Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66:688–.
Shleifer, A., and R. Vishney. 1993. Corruption. *The Quarterly Journal of Economics* 108:599–617.
Stigler, G. J. 1971. The theory of economic regulation. *The Bell Journal of Economics and Management Science* 3–21.
Tabakovic, H., and T. G. Wollmann. 2018. From revolving doors to regulatory capture? Evidence from patent examiners. Working Paper, National Bureau of Economic Research.
Tate, G., and L. Yang. 2015. Female leadership and gender equity: Evidence from plant closure. *Journal of Financial Economics* 117:77–97.
Taylor, L. A. 2010. Why are CEOs rarely fired? Evidence from structural estimation. *The Journal of Finance* 65:2051–87.
———. 2013. CEO wage dynamics: Estimates from a learning model. *Journal of Financial Economics* 108:79–98.
VanHelene, A. D., I. Khatri, C. B. Hilton, S. Mishra, E. D. Gamsiz Uzun, and J. L. Warner.

2024. Inferring gender from first names: Comparing the accuracy of genderize, gender api, and the gender r package on authors of diverse nationality. *PLOS digital health* 3:e0000456–.

Veitch, V., D. Sridhar, and D. M. Blei. 2020. Adapting text embeddings for causal inference 124:919–28.

von Meyerinck, F., J. Romer, and M. Schmid. 2025. Ceo turnover and director reputation. *Journal of Financial Economics* 163:103971–.

Wu, Y. 2017. What's behind smooth dividends? evidence from structural estimation. *The Review of Financial Studies* 31:3979–4016.

Figure 1: **Variation in Rulemaking Gender Output Gap by Cohort**

Each point represents the coefficient from estimating Equation (10) by cohort (the year in which the rule was first proposed). The dependent variable is *Tilt*, representing how the principal evaluates the draft rule presented by the rulemaking team. The independent variable is an indicator for any female member in drafting team. The vertical dotted line represents the mean of the beta coefficient across all cohorts. 95% confidence interval with standard errors clustered at rule level represent by gray lines. See Section 3.4.3.

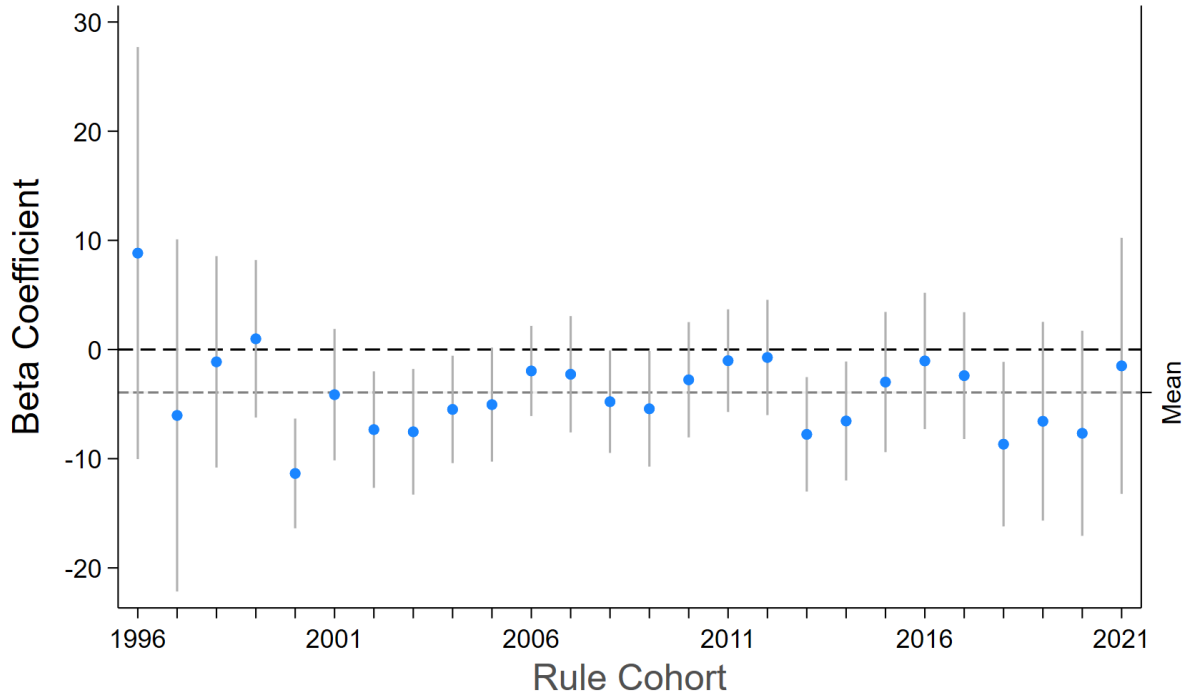


Figure 2: **Variation in Rulemaking Gender Output Gap by Topic**

Each bar represents the coefficient from estimating Equation (10) by topic. We use data from Chang et al. (2023), who decompose each rule proposal into 100 topics using a Latent Dirichlet Allocation (LDA) algorithm. We keep topics which impact at least 250 proposals. The dependent variable is *Tilt*, representing how the principal evaluates the draft rule presented by the rulemaking team. The independent variable is an indicator for any female member in drafting team. The vertical dotted line represents the mean of the beta coefficient across all topics. 95% confidence interval with standard errors clustered at rule level represent by gray lines. See Section 3.4.3.

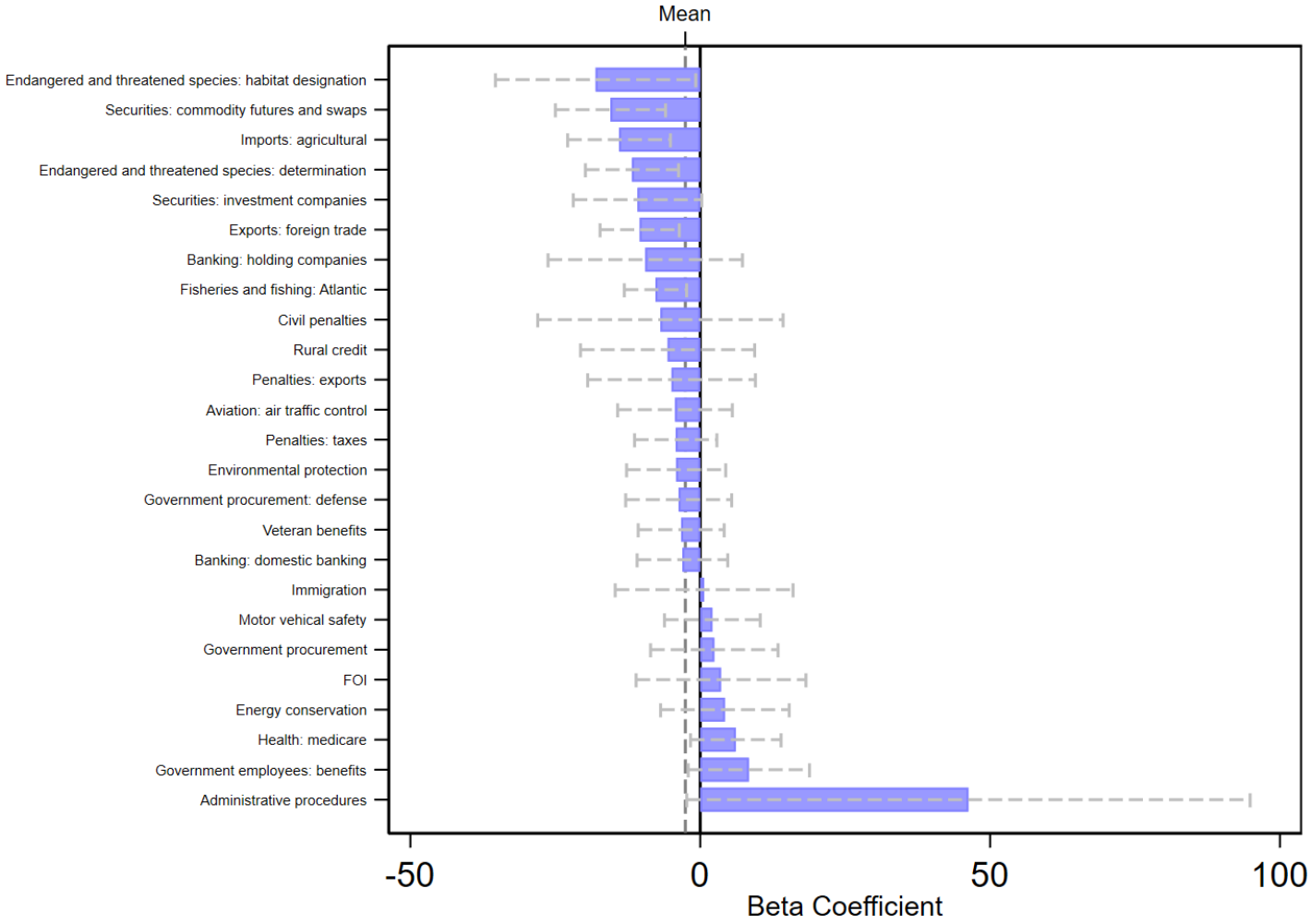


Figure 3: **Variation in Rulemaking Gender Output Gap by Agency**

Each bar represents the coefficient from estimating Equation (10) by parent agency. We club all parent agencies with less than 200 rules into other. The dependent variable is *Tilt*, representing how the principal evaluates the draft rule presented by the rulemaking team. The independent variable is an indicator for any female member in drafting team. The vertical dotted line represents the mean of the beta coefficient across all agencies. 95% confidence interval with standard errors clustered at rule level represent by gray lines. See Section 3.4.3.

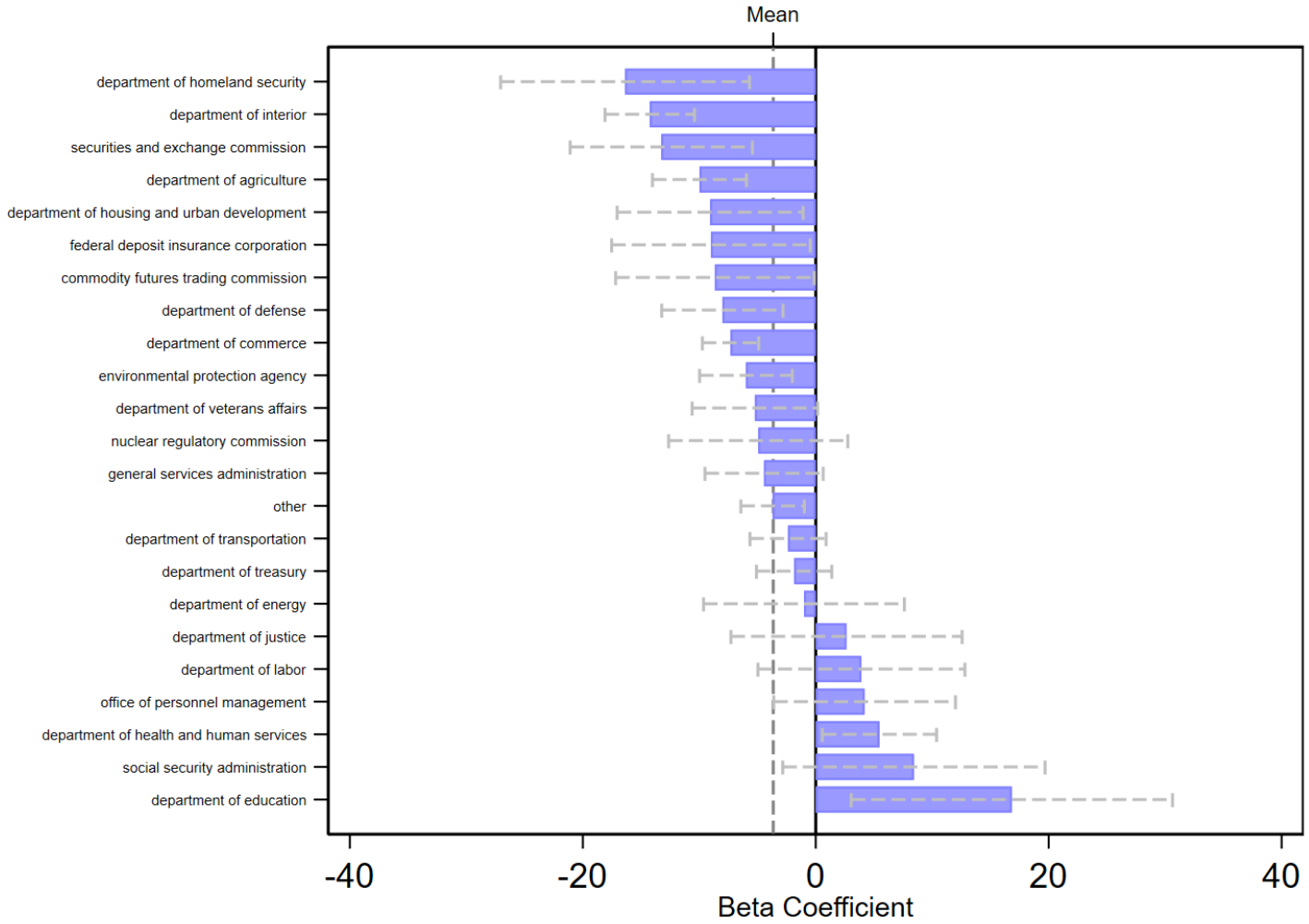


Figure 4: **Decomposition of the Gender Output Gap**

This figure shows the decomposition of the gender differential using the structural model. The horizontal dash line displays the observed gender differential in the completion rate of the rule (Table 2). The colored arrows represent estimated contributions from specific channels, where red (blue) color indicates a channel that increases (decreases) the gender gap. The effects accumulate from left to right; the top (bottom) of a red (blue) bar reflects the cumulative contribution up to that point. The selection channel is estimated by replacing the initial rule quality (x) assigned to males with that of females. The evaluation channel is estimated by replacing the subject evaluation discount (δ) of males with that of females. The cost channel is estimated by replacing the cost parameter (γ) of males with that of females. The incentive channel is estimated by replacing the wage to effort sensitivity (ω) of males with that of females. See Section 4.3.

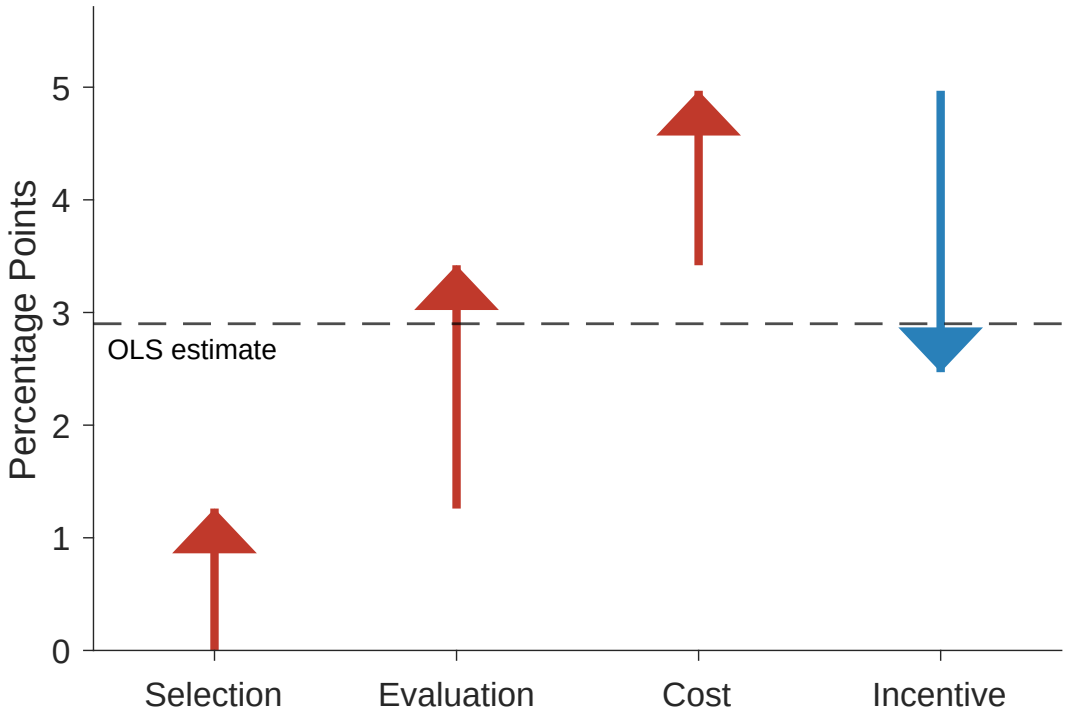


Table 1: **Descriptive statistics**

Panel A. Teams. The sample includes all federal rule proposals since 1996, excluding those still under consideration as of September 2023. We report the number of regulators working on the proposal (*Rulemakers*) and the gender composition; the rule's importance based on the government's scaling system, as well as indicator for major rules; the project's outcome; the time the rule spent in the pipeline (in days); and an indicator if the rule was put on pause for more than 12 months. See Section 1.2.

Statistic	Average	S.D.	Min.	Max.	Obs.
	(1)	(2)	(3)	(4)	(5)
Teams:					
Team Size	1.6	1.1	1.0	22.0	37,739
Any Female (%)	52.9	49.9	0.0	100.0	37,739
Males (%)	53.0	45.1	0.0	100.0	37,739
Females (%)	42.8	44.6	0.0	100.0	37,739
Mean Team Pay (2023 $000s)	176.3	41.0	24.9	345.8	17,094
Mean Team Tenure (in Years)	18.2	9.5	0.0	52.0	17,094
Washington, DC (%)	54.7	49.4	0.0	100.0	17,094
Rule Outcomes:					
Full rule (%)	70.8	45.5	0.0	100.0	37,739
Full repeal (%)	29.2	45.5	0.0	100.0	37,739
Pending time (days)	748.2	971.3	0.0	9,924.0	37,739
Suspended (%)	24.5	43.0	0.0	100.0	37,739
Revisions	0.5	0.4	0.0	1.7	12,804
C-BERT moments:					
Completability (%)	74.7	30.6	2.7	98.2	35,635
Tilt (%)	5.4	35.0	-95.2	57.1	20,526
Importance:					
Major (%)	5.4	22.6	0.0	100.0	35,308
Significant (Tier 1) (%)	27.7	44.8	0.0	100.0	37,739
Substantive (Tier 2) (%)	64.3	47.9	0.0	100.0	37,739
Routine (Tier 3) (%)	8.0	27.1	0.0	100.0	37,739

Panel B. Individuals. For each individual rulemaker, we compute the total number of rule proposals, outcomes, and importance. We also report their gender and team members. Finally, we match 10,246 rulemakers to their payroll records and report their hiring date, number of promotions, and the average basic pay and average annual pay raise. See Section 1.2.

Statistic	Average	S.D.	Min.	Max.	Obs.
	(1)	(2)	(3)	(4)	(5)
Rulemaking work:					
Total rules	3.8	11.1	1.0	377.0	15,492
Major (%)	7.3	22.6	0.0	100.0	14,802
Significant (Tier 1) (%)	34.7	42.4	0.0	100.0	15,492
Substantive (Tier 2) (%)	58.1	43.8	0.0	100.0	15,492
Routine (Tier 3) (%)	7.3	22.7	0.0	100.0	15,492
Rulemaking Outcomes:					
Full rule (%)	66.3	39.8	0.0	100.0	15,492
Full repeal (%)	33.7	39.8	0.0	100.0	15,492
Pending time (days)	1,241.7	1,260.8	0.0	9,772.0	15,492
Suspended (%)	36.1	41.6	0.0	100.0	15,492
Revisions	0.6	0.4	0.0	1.7	8,056
C-BERT moments:					
Completability (%)	70.6	27.7	2.9	98.2	15,089
Tilt (%)	0.3	34.1	-94.9	55.6	10,361
Demographics:					
Females (%)	45.3	49.8	0.0	100.0	14,792
Coauthors	1.5	1.7	0.0	21.0	15,492
Female coauthors (%)	43.1	36.2	0.0	100.0	11,333
Career Trajectory:					
Hiring Date	1,991	14	1,947	2,022	10,246
Pay (2023 $000s)	156.6	47.0	40.1	271.1	10,232
Δ Pay (2023 $000s)	4.8	3.2	0.0	17.9	9,965
Promotions	3.8	7.3	0.0	50.0	10,251
Washington, DC (%)	54.6	44.9	0	100	10,251

Table 2: **Female representation and successful project completion**

Results from estimating Equation (4). *Completion* = 1 if the team successfully completed the project, turning the rule proposal into a full rule. *Female* equals 1 if at least one of the team members is a female, and 0 for all-male teams. Standard errors clustered at the rule level are in parentheses. See Section 3.1.

Outcome:	*Completion*			
	(1)	(2)	(3)	(4)
Female	-2***	-1.2**	-1.2**	-1.5**
	(0.5)	(0.5)	(0.5)	(0.6)
Obs.	37,739	37,739	37,739	37,739
R^2	0.00	0.07	0.10	0.30
Effect (%)	-2.9	-1.7	-1.6	-2.1
Agency FE		Y	Y	Y
Agency×Tier FE			Y	Y
Agency×Tier×Cohort FE				Y

Table 3: **The selection channel**

Results from estimating Equation (5). *Completability* is the proposal completion probability based on the initial text of the title and abstract, using a single model for all rules irrespective of gender. *Female* equals 1 if at least one of the team members is a female, and 0 for all-male teams. Standard errors clustered at the rule level are in parentheses. See Section 3.2.

Outcome:	*Completability*			
	(1)	(2)	(3)	(4)
Female	-2.1***	-1.4***	-1.4***	-.88**
	(0.3)	(0.3)	(0.3)	(0.4)
Obs.	35,635	35,635	35,635	35,635
R^2	0.00	0.12	0.15	0.36
Effect (%)	-2.8	-1.8	-1.8	-1.2
Agency FE		Y	Y	Y
Agency×Tier FE			Y	Y
Agency×Tier×Cohort FE				Y

Table 4: **The effort channel**

Panel A. Textual revisions. Results from estimating Equation (4). *Revisions* is the average change in topic composition between the early and final versions of the rule proposal. *Female* equals 1 if at least one of the team members is a female, and 0 for all-male teams. Standard errors clustered at the rule level are in parentheses. See Section 3.3.1.

Outcome:	*Revisions*			
	(1)	(2)	(3)	(4)
Female	.066***	.041***	.039***	.038***
	(0.0)	(0.0)	(0.0)	(0.0)
Obs.	12,804	12,804	12,804	12,804
R^2	0.01	0.26	0.30	0.54
Effect (%)	13.9	8.8	8.4	8.1
Agency FE		Y	Y	Y
Agency×Tier FE			Y	Y
Agency×Tier×Cohort FE				Y

Panel B. Mid-point milestones. Results from estimating Equation (4). $Final = 1$ ($Interim = 1$) if the official rule was final or interim. $Draft = 1$ if a draft proposal was published for comments. $Suspended = 1$ if the rule was temporarily put on hold. $Cancelled = 1$ if the rule was officially withdrawn from the pipeline. Duration measures the time in days till final proposal outcome. *Female* equals 1 if at least one of the team members is a female, and 0 for all-male teams. Robust standard errors are in parentheses. See Section 3.3.2.

Outcome:	*Final*	*Interim*	*Draft*	*Suspended*	*Canceled*	*Duration*
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-2***	1.2***	3.8***	6.5***	1.8***	272***
	(0.6)	(0.4)	(0.6)	(0.5)	(0.6)	(12.0)
Obs.	37,739	37,739	37,739	37,739	37,739	37,739
R^2	0.31	0.3	0.34	0.39	0.3	0.4
Effect (%)	-2.9	13.3	7.4	26.6	6.5	36.3
Agency×Tier×Cohort FE	Y	Y	Y	Y	Y	Y

Panel C. Incentives. In columns 1-2, we estimate Equation (7) in a panel of rulemakers. In columns 3-4, we estimate Equation (8) in the cross-section of rulemakers. *Pay* (*LogPay*) is the rulemaker's pay (log of pay) at time t, $Exit = 1$ if the rulemaker resigned their government position, *Tenure* is final number of years in government service. $Female = 1$ if the rulemaker is a female. *CompletionRatio* is the fraction of rulemaker i's rulemaking projects successfully completed, up to year $t-1$ (columns 1-2) or throughout their tenure (columns 3-4). Controls represent the number of projects or up to year $t-1$ (columns 1-2) or throughout the rulemaker's tenure (columns 3-4); and the salary in year $t-1$ (columns 1-2) or the initial salary (columns 3-4). We winsorize salary outcomes at a 1% level. Standard errors clustered at the rulemaker level are in parentheses. See Section 3.3.3.

Model:	**Panel**		**Cross Section**	
Outcome:	*Pay*	*Log(Pay)*	*Exit*	*Tenure*
	(1)	(2)	(3)	(4)
CompletionRatio	2,480***	.013**	.62***	-14***
	(619)	(0.01)	(0.08)	(1.10)
CompletionRatio × *Female*	2,009**	.026***	-0.01	1.10
	(906)	(0.01)	(0.12)	(1.50)
Obs.	97,753	97,753	3,033	3,033
R^2	0.98	0.97	0.56	0.77
Controls	Y	Y	Y	Y
Agency×Location×Occupation×Year FE	Y	Y	Y	Y
Person FE	Y	Y		

Table 5: **The evaluation channel**

Results from estimating Equation (10). *Tilt* is the difference between realized project outcome (equals 1 if the draft became a final rule) and the completion probability if the draft rule has been written by the opposite gender, calculated using the C-BERT methodology and the full text of draft rules. *Female* equals 1 if at least one of the team members is a female, and 0 for all-male teams. Standard errors clustered at the rule level are in parentheses. See Section 3.4.1 and Section 3.4.2.

Outcome:	*Tilt*			
	(1)	(2)	(3)	(4)
Female	-5.3***	-5.3***	-5.2***	-4.4***
	(0.5)	(0.5)	(0.5)	(0.6)
Obs.	18,268	18,268	18,268	18,268
R^2	0.01	0.04	0.07	0.39
Effect (%)	-5.9	-5.8	-5.8	-4.9
Agency FE	-	Y	Y	Y
Agency×Tier FE		-	Y	Y
Agency×Tier×Cohort FE	-	-	-	Y

Table 6: **The role of team characteristics**

Results from estimating Equation (10). *Tilt*(%) is the difference between actual proposal completion and completion probability if the presence of any female in the team is reversed, calculated using the C-BERT methodology on a model trained using the full text of proposals. In columns 2-6 we add controls for average team salary, tenure, ratio of team located in Washington, ratio of team members at each different rank, and ratio of team members in each different occupation. We drop proposals where text completely predicts gender of authors (top and bottom 5 percentile). Standard errors clustered at the rule level are in parentheses. See Section 3.4.3.

Outcome:	*Tilt*					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-4.10***	-4.10***	-4.06***	-4.07***	-3.98***	-3.98***
	(0.96)	(0.96)	(0.97)	(0.97)	(1.00)	(1.03)
Mean Team Pay (2023 $000s)		0.00	0.00	0.00	-0.01	-0.01
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Mean Team Tenure (in years)			-0.03	-0.03	-0.02	0.03
			(0.05)	(0.05)	(0.06)	(0.07)
Washington, DC (%)				0.01	0.01	0.02
				(0.03)	(0.03)	(0.03)
Obs.	9,096	9,096	9,096	9,096	9,096	9,096
R^2	0.44	0.44	0.44	0.44	0.46	0.47
Effect (%)	-4.5	-4.5	-4.4	-4.5	-4.3	-4.4
Agency×Tier×Cohort FE	Y	Y	Y	Y	Y	Y
Rank Controls					Y	Y
Occupation Controls						Y

Table 7: **Structural estimation: Parameter values**

This table shows the parameter estimates for males and females, respectively. The details of the estimation are described in Section 4.1.

Parameter	Male	Female
Initial net benefits (x)	1.5152	1.4162
Wage to effort sensitivity (ω)	0.3428	0.6204
Cost of effort (γ)	1.2618	2.2716
Subjective assessment (δ)	0.0000	-0.1658
Residual (α)	0.6648	0.8062

Internet Appendix

A.1 Name matching

In this section we explain how we match names across two datasets: the Unified Agenda (UA) and the federal payroll (see Section 1.1). We represent each name in four formats: (1) first, middle, and last name; (2) first name, middle initial, and last name; (3) first and last name; and (4) first initial and last name only (for incomplete entries). We perform a fuzzy match using cosine similarity on each group, keeping matches with over 70% similarity within the same agency. If multiple matches exist, we retain the one with the highest similarity score. If duplicates remain, we prioritize matches in the order of group 1 through group 4. From this, we construct a preliminary set of matches, which we further filter to retain only high-quality matches. Specifically, we require at least 97% similarity on last names and 95% on first and middle names (where available). To account for reversed first and middle names, we check both name orders. If only initials are available for first or middle names, we require exact initial matches.

We begin with our sample of 37,739 rule proposals. We exclude proposals associated with agencies for which employee names are masked.[1] This leaves a sample of 29,528 proposals from eligible agencies. Among these, we are able to match at least one team member to the payroll for 24,190 rules, representing 82% of the eligible sample. We further restrict the analysis to proposals for which the entire team can be matched, yielding a final sample of 21,359 rule proposals (a 72% match rate), of which 17,100 (58%) include complete payroll information for all team members. Match rates vary across agencies. We achieve over 80% match rates in the Department of Commerce, the Federal Housing Finance Board, and the Commodity Futures Trading Commission. Several other financial regulatory agencies – such as the Securities and Exchange Commission and the Federal Deposit Insurance Corporation – also exhibit match rates exceeding 70%.

A.2 Literature review

Female inclusion and firm performance. Women's representation on corporate boards remains low despite steady gains over time: Field, Souther, and Yore (2020) document that although the percentage of female directors on large U.S. boards has increased, women remain dramatically under-represented in key leadership roles such as Chair and Lead Independent Director, indicating persistent informal barriers to advancement. Institutional investors have begun to push for change: Gormley et al. (2023) show that large index funds ("the Big Three") effectively use their proxy votes to pressure firms to increase female board representation, with the greatest impact at firms in which they hold concentrated stakes.

A growing body of evidence links board gender diversity to stronger governance and firm performance. Adams and Ferreira (2009) show that U.S. boards with more female directors

[1] We are unable to match any employees in the Environmental Protection Agency, Department of Defense, Department of Justice, Internal Revenue Service, Federal Reserve System, Army, Navy, Air Force, and specific bureaus within the Department of the Interior (such as the National Indian Gaming Commission), the Department of the Treasury (such as the Bureau of Customs), and the Department of Energy (such as Energy Efficiency and Renewable Energy), where names are fully or partially masked.

exhibit higher attendance, stronger monitoring, and higher Tobin's Q, suggesting that women bring complementary skills that improve both governance and firm value; von Meyerinck, Romer, and Schmid (2025) find that firms losing a female director suffer larger and more persistent declines in firm value than those losing a male director, implying that female directors contribute unique monitoring and advisory value. At the same time, Griffin, Li, and Xu (2021) provide international evidence that board gender diversity correlates positively with corporate innovation measures such as patenting and R&D intensity, especially in countries with supportive national innovation policies.

Executive-level studies highlight both behavioral differences and broader real-world impacts. Huang and Kisgen (2013) report that male executives display greater overconfidence while firms led by female executives earn higher announcement-period returns. In the manufacturing sector, Tate and Yang (2015) analyze U.S. plant closures and find that plants led by female executives mitigate adverse employment and wage declines for female workers relative to male leadership. In the healthcare industry, Lewellen (2025) examine decision-making and compensation in U.S. hospitals and find that male and female CEOs respond similarly to financial shocks, but female CEOs exhibit lower turnover rates and adopt more conservative pay structures.

Quotas have produced mixed results. Exploiting California's Senate Bill 826 mandate, Allen and Wahid (2023) show that requiring female board quotas increased gender diversity without harming firm financial performance. Norway's 2003 quota reform causally raised firm valuations by about 8 percent, with the strongest effects in state-controlled enterprises Ahern and Dittmar (2012). By contrast, Eckbo, Nygaard, and Thorburn (2021) revisit Norway's board-quota law and find that mandated female appointments led to directors with somewhat lower experience and mixed effects on stock-market valuations. Beyond boards, Leibbrandt, Wang, and Foo (2018) conduct lab and field experiments on grant-allocation competitions and demonstrate that gender quotas boost women's funding success but provoke evaluators to score female proposals more sternly when quotas are known to be in place; similarly, Deschamps (2024) implement a 30% gender quota on candidate-selection panels at a large consultancy and find that while women's hiring rates increase under quotas, committees also evaluate female applicants more harshly in non-quota contexts, revealing a backlash effect.

Gender performance gaps. Azmat and Ferrer (2017) find that male attorneys bill more hours and generate more revenues, and Bertrand, Goldin, and Katz (2010) document diverging earnings trajectories among MBA graduates due in part to women's greater likelihood of prioritizing family. Exley and Kessler (2022) show that women rate their performance lower than equally performing men, resulting in worse hiring and compensation outcomes. This adds to the literature showing that women tend to underestimate their own ability (Niederle and Vesterlund (2007)), and exhibit more cautious self-presentation leading to reduced recognition and opportunities (Murciano-Goroff (2022)). Several studies attribtute performance gaps to competitive environment. For instance, while women tend to perform equally or better than men in cooperative or non-tournament settings (Gneezy, Niederle, and Rustichini (2003)), female students perform worse than males on competitive entrance exams for elite French business schools, despite outperforming men in earlier academic stages (Ors, Palomino, and Peyrache (2013)). This could reflect aversion to competition, stemming from both lower

confidence and greater risk aversion (Croson and Gneezy (2009); Bertrand (2011)).

Gender diversity in the public sector. Gender diversity in government agencies can influence both performance and representation. Regarding performance, in Rwandan primary schools, long-term incentives enhance teacher effort and retention particularly among women (Leaver et al. (2021)). Regarding female representation, Bagues and Esteve-Volart (2010) demonstrate that female-majority examination panels significantly boost the hiring of qualified women. Evidence from Switzerland indicates that public-sector recruiters continue to post more male-oriented vacancies (Card, Colella, and Lalive (2025)). Finally, a field experiment in UK social care finds that highlighting male success stories can attract more men into traditionally "pink-collar" roles (typically associated with women) without deterring female applicants (Delfino (2024)).

A.3 C-BERT methodology

A.3.1 BERT

BERT makes use of Transformers, an attention mechanism that learns contextual relationship between words. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore, it is considered bidirectional, though it would be more accurate to say that it is non-directional. This characteristic allows the model to learn the context of a word based on all its surroundings (left and right of the word). The input is a sequence of tokens, which are first embedded into vectors and then processed in the neural network. The output is a sequence of vectors of size H, in which each vector corresponds to an input token with the same index.

BERT uses two training strategies. The first strategy is **Masked LM** (MLM), a method that can be thought of as filling in the blanks. Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. The prediction requires adding a classification layer on top of the output of the encoder, transforming the output vectors into vocabulary dimension (by multiplying output vectors and embedding matrix), and then calculating the probability of each word in the vocabulary with softmax.

The second strategy is **Next Sentence Prediction**. Here, the model receives pairs of sentences and learns to predict the second sentence in the pair. The data is split into two parts, where half of the inputs have a sequential pair while the remaining half have random pairs. To help the model distinguish between the two sentences in training, the input is processed in the following way before entering the model. First, a [CLS] token is inserted at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence. Then, a sentence embedding indicating Sentence A or Sentence B is added to each token. Sentence embeddings are similar in concept to token embeddings with a vocabulary of 2. Finally, a positional embedding is added to each token to indicate its position in the sequence. To predict if the second sentence is indeed connected to the first, the

following steps are performed. First, the entire input sequence goes through the Transformer model. Then, the output of the [CLS] token is transformed into a 2×1 shaped vector, using a simple classification layer (learned matrices of weights and biases). Next, we calculate the probability of the second sentence being connected to the first with softmax, which converts the raw scores into a probability distribution over the two possible labels: "next sentence" and "random sentence." This enables the model to output interpretable predictions about sentence continuity.[2]

When training the BERT model, Masked LM and Next Sentence Prediction are trained together. The goal is to minimize the combined loss function[3] of the two strategies. In linear regressions, for example, the loss function is the squared distance between the predicted and actual values (Mean Squared Error or MSE). However, BERT training involves classification tasks, that is, predicting masked tokens and sentence relationships. For these, the appropriate loss function is Binary Cross-Entropy (or more generally, Cross-Entropy), which compares the predicted probabilities to the actual class labels and penalizes incorrect predictions accordingly. It then calculates the score that penalizes the probabilities based on the distance from the expected value.

A.3.2 Implementation

First, we fine-tune a pre-trained BERT model (Devlin et al., 2019) to generate high-dimensional representations of text, using 768-dimensional embeddings for each token. BERT is initially pre-trained on a large corpus of unlabeled text to learn contextualized word representations. Fine-tuning adapts the model to the specific linguistic patterns found in regulatory documents. For this purpose, we use the titles and abstracts of all rule proposals announced in the Unified Agenda. This domain-specific fine-tuning improves BERT's ability to perform downstream tasks, such as predicting the probability of rule success.

We then build a custom neural architecture, which we refer to as C-BERT, to examine the relationship between gender and the likelihood that a rule proposal is finalized. The architecture consists of three models trained jointly to estimate this treatment effect. First, we train a single dense (fully connected) layer to map BERT-generated embeddings to the gender propensity of the authoring team. A dense layer connects each neuron to every neuron in the preceding layer, enabling the model to capture complex patterns in the input representation. Simultaneously, we train two separate two-layer perceptrons to predict the completion probability of projects for all-male and female-inclusive teams, respectively. For each network, we compute the binary cross-entropy loss, and the final loss function is defined as a weighted average of the three individual losses. This composite loss guides model training. Training proceeds until the loss function converges. In our implementation, we find that performance plateaus after approximately 20 epochs, indicating that additional training yields minimal improvement in fit.

With the trained mappings in place, we assess the impact of female inclusion on the rulemaking project's completion likelihood. Specifically, we ask: what would be the predicted

[2]Given scores z_1 and z_2, softmax computes probabilities $p_1 = \frac{e^{z_1}}{e^{z_1}+e^{z_2}}$ and $p_2 = \frac{e^{z_2}}{e^{z_1}+e^{z_2}}$. These values lie between 0 and 1 and sum to 1, making them suitable for binary classification tasks.

[3]This function captures how well an algorithm performs on a given data set, by mapping an event onto a real number which represents some cost associated with the event.

probability of completion of the female-inclusive project if it was an all-male project, and vice versa? We frame this using the potential outcomes framework (Rubin, 1974), using gender as the binary treatment variable. Let $T_i \in \{0, 1\}$ indicate presence of a female in the rulemaking team of proposal i, where $T_i = 1$ denotes a female presence and $T_i = 0$ denotes all-male team. Define $Y_i(1)$ as the potential outcome (i.e., completion probability) if the team includes a female, and $Y_i(0)$ as the outcome if the team is all-male. For each proposal, we observe only one of the two potential outcomes, depending on T_i. To estimate the counterfactual, we use the trained neural networks as structural mappings. Specifically, for all-male proposals ($T_i = 0$), we estimate $\hat{Y}_i(1)$ by applying the female-trained network to the BERT embedding x_i. Equivalently, for proposals authored by females ($T_i = 1$), we estimate $\hat{Y}_i(0)$ using the male-trained network on the same embedding x_i. This yields counterfactual predictions $\hat{Y}_i^{(F \to M)}$ and $\hat{Y}_i^{(M \to F)}$, which allow us to estimate the conditional average treatment effect (CATE) for subgroups, or the average treatment effect (ATE) across the full sample:

$$\text{ATE} = \sum_{i=1}^{N} [\mathbb{E}(Y_i(1)|x_i) - \mathbb{E}(Y_i(0)|x_i)]$$

By holding proposal content fixed (via the embedding x_i) and varying only the gender-based mapping, we isolate the impact of gender on project completion.

Table A.1: **Team composition and female representation**

This table disaggregates the team-level sample averages from Table 1, Panel A, by team composition. We report relevant outcomes separately for rules authored by all-female teams and for those with no female rulemakers. The final column presents the difference in means between the two groups.

Statistic	Female-inclusive	All-male	Diff.
	(1)	(2)	(3)
Teams:			
Team Size	1.9	1.3	0.6***
Mean Team Pay (2023 $000s)	171.6	180.1	-8.5***
Mean Team Tenure (in Years)	17.7	18.6	-0.9***
Washington, DC (%)	54.9	54.6	0.3
Rule Outcomes:			
Full rule (%)	69.9	71.6	-2.0***
Full repeal (%)	30.1	28.1	2.0***
Pending time (days)	881.2	598.6	282.6***
Suspended (%)	27.7	20.9	6.7***
Revisions	0.5	0.4	0.1***
C-BERT moments:			
Completability (%)	73.7	75.8	-2.1***
Tilt (%)	3.1	8.3	-5.2***
Importance:			
Major (%)	6.0	4.7	1.2***
Significant (Tier 1) (%)	29.1	26.1	3.0***
Substantive (Tier 2) (%)	63.3	65.4	-2.1***
Routine (Tier 3) (%)	7.60	8.40	-0.9***

Table A.2: **The evaluation channel: Robustness**

Columns 1-2 use the ratio of females in a team, instead of a binary indicator for any female in a rule-making team. Column 3 controls for team size. Column 4 estimates results for proposals composed by single authors. Column 5 estimates results for proposals composed by multiple authors. Standard errors clustered at the rule level are in parentheses. See Section 3.4.3.

Outcome:	*Tilt* (%)				
Sample:	*All Rules*			*Solo*	*Coauthored*
	(1)	(2)	(3)	(4)	(5)
Ratio Female	-2.6***	-1.7**			
	(0.5)	(0.7)			
Female			-2.0***	-1.4*	-4.1***
			(0.7)	(0.8)	(1.5)
Team Size			-4.4***		
			(0.4)		
Obs.	18,268	18,268	18,268	11,227	7,041
R^2	0.00	0.39	0.40	0.46	0.50
Effect (%)	-2.9	-1.9	-2.3	-1.6	-4.7
Agency×Tier×Cohort FE		Y	Y	Y	Y