I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17480

Designing Gender Equity: Evidence from Hiring Practices

Tatiana Mocanu

NOVEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17480

Designing Gender Equity: Evidence from Hiring Practices

Tatiana Mocanu
Columbia University and IZA

NOVEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Designing Gender Equity: Evidence from Hiring Practices*

I combine novel data on job applications and hiring decisions for the universe of public sector jobs in Brazil and a natural experiment that decreased discretion in hiring to analyze how screening determines gender application and hiring gaps. I find that hiring practices have crucial gender equity consequences for selection and sorting, and not all approaches to reduce discretion have the same implications. Limiting discretion in existing tools or adding new impartial tools reduces the gender hiring gap by a third. However, policies that eliminate subjective tools like interviews are ineffective, suggesting employers should carefully weigh bias-information trade-offs.

JEL Classification: M51, D73, J16, J45, J71

Keywords: hiring practices, gender bias, public sector personnel

Corresponding author:
Tatiana Mocanu
Columbia University
116th and Broadway
New York, NY 10027
USA
E-mail: tm3326@columbia.edu

* I am deeply grateful for the guidance and support from David Albouy, Alex Bartik, and Dan Bernhardt. This paper has benefited greatly from feedback and discussions with Vittorio Bassi, Barbara Biasi, Nicola Bianchi, Sandra Black, Eliza Forsythe, Yana Gallen, Andy Garin, Pat Kline, Dmitri Koustas, Gianmarco León-Ciliotta, Day Manoli, Sendhil Mullainathan, Kieu-Trang Nguyen, Matthew Notowidigdo, Dan-Olof Rooth, Matteo Paradisi, Nicola Persico, Brendan Price, Simon Quinn, Natalia Rigol, Olga Stoddard, Pedro Tremacoldi-Rossi, Sergio Urzua, Chris Walters, Russell Weinstein, Guo Xu, Owen Zidar, as well as seminar and conference participants at Columbia, UC Berkeley, Northwestern Kellogg, Cornell, Boston University, Oxford, Bocconi, USC Marshall, Princeton, EIEF, CEMFI, University of Virginia Batten, Stockholm University, UCL, Rutgers, University of Illinois Urbana-Champaign, Sciences Po Summer Workshop: Labor, NBER Summer Institute Gender in the Economy, Discrimination in the 21st Century Conference (University of Chicago), Stanford Institute for Theoretical Economics Gender, Emerging Scholars Conference (University of Wisconsin-Madison), CESifo Area Conference on Labor Economics, Call to Service Conference (Stanford University), SOLE, Annual Congress of the International Institute of Public Finance (Utah State University), North America Summer Meeting of the Econometric Society, CESifo/ifo Junior Workshop on Big Data, ERMAS Applied Monthly Seminar, and the 3rd Monash-Warwick-Zurich Text-as-Data Workshop. I also thank several career public servants in various Brazilian government levels for sharing their experience in participating and conducting hiring processes in the country's public sector.

1 Introduction

Much progress has been made in the last decades toward closing gender gaps in labor market outcomes. Despite that, women continue to face barriers to make inroads in many careers, putting into focus the equity consequences of how firms and organizations recruit and select job applicants. While the existence of discrimination in hiring has been well documented,[1] little is known about the extent to which different screening practices may mitigate or exacerbate gender disparities.

How to best design hiring processes that are less susceptible to bias is a question that remains open. Employers have at their disposal both subjective and objective screening tools when making hiring decisions. On one hand, subjective tools, such as interviews, that give evaluators more discretion may provide additional productivity information, but are more prone to evaluator biases. On the other hand, more objective tools, such as written tests, constrain evaluator behavior but may screen out otherwise qualified workers.

These different properties of screening tools imply that alternative ways to reduce discretion may have distinct consequences for diversity. First, the choice of screening practices may impose a bias-information trade-off. Removal of subjective steps discards information about a candidate that is difficult to learn otherwise. And removing information about minority applicants, even when biased, may not necessarily be effective since evaluators may instead rely more heavily on priors about group characteristics. Second, introducing objective tools, while constraining evaluator behavior, could inadvertently introduce disparate impact.[2] And to make things even more complicated, changes in screening tools can also shift whether women apply for a job in the first place.

Using Brazil's public sector as setting, this paper studies the role of screening methods in determining gender application and hiring gaps. Tackling these questions empirically has been challenging because hiring decisions are effectively a black box. Economists rarely observe how organizations screen candidates and make selection decisions. In addition, since employers' choices of how to screen their employees is endogenous, exogenous variation necessary for causal analysis has limited empirical work. As Oyer and Schaefer (2011) put it: "What manager, after all, would allow an academic economist to experiment with the firm's screening, interviewing or hiring decisions?".

[1] See, for example, Goldin and Rouse (2000), Bertrand and Mullainathan (2004), Kline et al. (2022).

[2] The standard definition for disparate impact since Griggs v. Duke Power Co. (1971) is that, even if discrimination is unintentional, a job requirement that leads to disparities between applicants from different groups can only be legally justified if it connects to job duties. A prominent example in the US is the use of aptitude test in police hiring, on which black applicants perform worse white applicants, but the entrance examination scores are not predictive of performance on the job (McCrary (2007)).

I combine novel data on job applications and hiring decisions for the universe of public sector jobs and a natural experiment that allow me to overcome these challenges. To open the black box of hiring, I develop a two-step, flexible natural language processing algorithm that distills over 35 million official government text files into data. These documents contain detailed records of the universe of Brazilian government screening processes, but have remained unavailable to researchers due to their highly confounding and unstructured nature. My novel query-based based algorithm generates a rich database detailing job applicant performance and evaluator decision making process, including job openings and offers, applicant and evaluator identities, and candidates' individual scores by screening method.

The paper leverages variation generated by the introduction of a requirement to public sector hiring mandating impartial screening. Public employees in Brazil are hired through a formal, merit-based selection system that has been in place for decades. With the passage of the country's new Constitution in 1988, parts of the constitutional text were modified to align with a new social focus. Concerning the selection of public servants, this came in the form of codifying the understanding that hiring practices that could enable differential treatment would be legally challenged, which led to a redesign of hiring processes toward a reduction in discretion.

The impartiality requirement was introduced without specific compliance directions. This feature generated a number of different changes to screening practices, with a treatment assignment process determined by the occupation. This followed the tradition of organizing rules related to public servants around careers, which in turn drove the discussions about how to modify hiring practices. Bureaucrats and legal experts at high organizational levels took into consideration the pre-existing screening tools in an occupation and modified selection processes in one of two general ways: make tools already in use more impartial or change the screening tools. While the first case amounted to fully or partially blinding written exams, the second included cases when written exams were added and non-written exams were removed or replaced. This richness in variation provides a unique empirical setting to study the equity consequences of alternative ways to reduce discretion.

In line with the legal interpretation provided by bureaucrats, 95% of hiring processes within a job title adopted only one change in screening methods. Yet, each treatment involved a variety of occupations across several dimensions. For example, kitchen assistants and civil engineers had screening modified in the same way, just like teachers and police officers were assigned the same treatment. Importantly, the introduction of the impartiality requirement did not attend specific policy goals, like increasing diversity, nor comprised a broader anti-discrimination effort, which I support empirically by leveraging thousands of carefully documented records about the drafting of Brazil's new Constitution.

My empirical analysis first documents the aggregate effects on gender gaps using a triple-difference design. I use an institutional delay of several years in adoption by states, in contrast with the sharp implementation by federal jobs, to compare outcomes of women and men in job processes within the same occupation in federal and states postings. The parallel trends assumption rules out time-varying unobserved variables that correlate with gender gaps and that could change differentially between state and federal jobs. Consistent with that, there were no contemporaneous changes to labor markets affecting treatment relative to control, and no pre-trends in gender gaps.[3]

To leverage the fact that my data allows to observe job seekers' applications and employers' hiring decisions, I decompose the components that determine the share of female hires in a hiring process — selection (demand) and sorting (supply) and estimate each effect separately. While the share of female hires captures whether the workforce is becoming more diverse and is usually the outcome observable to researchers, measuring demand and supply effects individually is important to understand how each margin adjusts to changes in the hiring process design, since they have distinct potential remedies and policy consequences.[4]

I estimate that women's final scores increased by 0.07 standard deviation, accompanied by a decrease of similar magnitude in men's scores. This resulted in the overall gender score gap closing in response to the reduction of hiring discretion. Women's final scores increase was driven by higher marks in written exams, which universally became blind. When still used, non-written stages had gender score gaps unchanged, suggesting evaluators did not respond strategically by favoring men in more discretionary practices. The decrease in the gender final score gap translates into an improved conditional probability of being hired for women and a narrowing of the gender hiring gap of about 44% of the pre-treatment level, even after controlling for job process competitiveness.

Who is hired ultimately depends not only on the employer's hiring decision, but also on who applies to begin with. Men and women may respond differently to perceived discrimination associated with unfair hiring practices or more costly preparation as selection processes are modified. Using the entire pool of applicants, including those rejected, I find that application rates of women relative to men increased by about 1 percentage point. Taken together, both higher employer demand and application rates of women result in the increase of gender diversity among hires just a few years after the new Constitution came into effect.

[3] I further support the credibility of the design underpinning my aggregate estimates through several checks. These include whether establishments tried to front-run their legislative process that required a long delay to implement the impartiality requirement for states (anticipation), or if the treatment assignment to federal jobs led to spillovers to control job processes. The weight of the evidence rejects the plausibility of these violations.

[4] This distinction is for example crucial to measure discrimination, since outcomes constructed from a sample of hired workers (e.g. share of female hires) are conditional on application decisions, which are endogenous (Durlauf and Heckman (2020)).

My initial set of results reveal that even within hiring systems with structured and professionalized selection procedures, discretion in hiring can still prevent women's access to jobs. This is particularly important in the context public bureaucracies, where having a more representative body of civil servants is linked to broader benefits to state capacity (Kingsley (1944), Alsan et al. (2019), Miller and Segal (2019), Xu (2021)), and the establishment of meritocratic recruitment alone, as with Brazil's case, may not achieve gender diversity as a goal.

However, these aggregate estimates showing the overall decrease in gender gaps mask important heterogeneity in the way discretion was reduced in practice. Exploiting another unique feature of my data detailing the specific screening methods for all job processes, in the second part of the paper I combine a theoretical framework with a multiple-treatment empirical strategy to isolate the effect of five alternative screening changes to reduce discretion in the sorting and selection of job seekers.

Before the impartiality requirement, occupations screened candidates using written, non-written, or a combination of both steps. The initial set of hiring tools in an occupation defined treatment assignment, a feature that I exploit for my empirical design and to test the identifying assumptions. Occupations with pre-treatment use of written tests fully blinded the exams; those using interviews or practical, oral exams either switched to a blind written test, or kept the non-written tools but added the less subjective step; those with a pre-existing combination either removed the non-written step or kept it. These different changes in screening methods, which define the treatments, all targeted a reduction in the level of discretion.

To validate the treatment assignment process induced by compliance with the new constitutional text, I conduct a series of tests. The space of possible treatments for an occupation includes at most one out of two changes in screening methods. This "one treatment or another" design where job processes are treated based on plausible exogenous variation generated by their job titles allows me to conduct three tests to support my empirical strategy. First, pre-treatment covariates including degree of feminization in the occupation, and share of female applicants and competitiveness of the job process, collectively explain a negligible portion of treatment assignment.

A second test addresses the possibility that policymakers based the changes in screening for an occupation on treatment effects, so that relevant factors in their information set uncorrelated with the previous covariates could drive selection. But to the extent that these unobservables also matter for gender hiring gaps, they would plausibly lead to pre-treatment differences in the outcome paths of jobs that sorted into alternative treatments. This suggests a visual inspection that combines the usual pre-trends probe between treated and control with the evaluation of systematic differences in the outcome between occupations following alternative treatments.

Gender gaps pre-treatment show parallel trajectories and indistinguishably-estimated differences in magnitudes between the pairs of occupations that could in theory follow alternative treatments. This indicates that, on average, selection of an occupation into a particular treatment shows no relationship with pre-existing differential gender hiring gaps, which in turn would be likely to pick up time-invariant unobservable factors. Finally, I leverage the fact that many occupations determine a treatment to assess systematic differences in estimated effects between each occupation, which further lends credibility to my design.

I build on a canonical model of statistical discrimination by introducing two new features. The first is tool discretion, which regulates evaluators' expression of their own bias. The second is tool bias, which allows for disparate impact independently of evaluator behavior. Written and non-written screening tools differ in the level of discretion, bias, and precision, closely matching the empirical setting. The model pins down gender hiring gaps in each of the five potential treatments that correspond to the different ways to reduce discretion as a function of three forces. These reduced-form predictions will guide the interpretation of the treatment effect estimates, many of which can have a theoretically ambiguous impact on diversity due to trade-offs between information an bias.

I break down the impact of each of the different ways to reduce discretion by first looking at the treatments that kept the tools in use. Starting with occupations that fully blinded job processes, the estimated impact on the gender hiring gap is a decrease of about 33% at the baseline. This treatment captures the causal effect of removing all sources of evaluator bias, and therefore reveals that it was an important force driving gender disparities.

The other treatment of this kind — occupations initially using a mix of written and non-written methods — kept the subjective stage and blinded the written step. Similar to fully blinding, this change in screening does not impose a potential trade-off because overall precision and bias remain unchanged. While the average treatment effect of partially blinding is null, the estimate is a function of the exogenously-determined weights given to blind written stages in job processes. In hiring processes with greater weight to fully blind exams, the gender hiring gap also considerably narrows. Importantly, I also show no evidence of strategic response by evaluators in non-written scores as their importance to the overall hiring decision varies across processes.

The second set of estimates comes from treatments that involved changes in the tools being used to screen job applicants. The first treatment introduces an objective tool to a pre-existing interview. A common objection against requiring standardized tests is that they could disadvantage minorities through a disparate impact channel, if for example women are worse test-takers in a way that is not predictive of worker productivity.[5] While my theoretical predictions show that this is a possibility, as long as these exams impose a smaller disparate impact than interviews, overall bias in screening will be lower. Combined with greater precision from their introduction, which favors minorities because performance signals influence evaluators' (men-favoring) priors more strongly, this condition would predict an increase of women hiring rates. Empirically, I estimate a decrease in the gender hiring gap of about 5.9 percentage points, or 35% of the initial hiring gap from using only an interview.

The next treatment removes subjective stages from the selection process. Similar to the other treatments that contain a mix of high and low-skill, here careers like cleaners, programmers, civil engineers, and nutritionists adopted the same change in screening. If evaluator bias matters and bias is particularly enabled by tools with high discretion, removing them from selection processes has an intuitive appeal. Contrary to this, my estimates reveal that removing subjective stages has no effect on the gender gap. This is consistent with a bias-information trade-off where the loss in screening precision from losing a productivity signal offsets the potential gains from eliminating disparate treatment and impact with the use of the tool. This results in evaluators putting more weight on their prior, making above-average women more difficult to be identified.[6]

To conclude the main set of results with multiple treatments, I also find that the way employers screen has important sorting effects. Changes in screening that improve perceived fairness without making the process more costly — because of additional hiring steps — increase women's application rates relative to men, further accounting for the increase in diversity captured by aggregate estimates. Taken together, I find that changes in screening tools have important consequences for both employer hiring decisions and job seekers' application behavior.

This paper makes several contributions. First, to the personnel economics of state literature (Finan et al. (2017), Besley et al. (2022)), which has studied governments' long-held goal of making civil service more professionalized (Grindle (2012)) in the context of large-scale reforms that curbed political patronage (Aneja and Xu (2022), Moreira and Pérez (2021a), Moreira and Pérez (2021b), Estrada (2019), Ornaghi (2019)). One limitation from prior studies stems from the overreaching essence of these reforms, where often the insulation of hiring from political influence also extends to protection from arbitrary firings, how promotion decisions are made, and general job incentives based on political cycles. This has made it difficult to isolate responses

[5] For example, Baldiga (2014) finds that women are more likely to skip than to guess on SAT questions that penalize a wrong answer, which decreases their test scores. Importantly, the pattern is not explained by gender differences in knowledge or confidence.

[6] The final treatment, where occupations change the screening design completely by replacing an interview with a written stage decreases the gender hiring gap by about 41%, another large estimated effect.

due to hiring design from changes in job attributes. My setting allows me to separately identify the role of discretion in hiring. Additionally, this is the first paper to directly observe and document how public servants are screened, overcoming another limitation of prior studies which could only provide little information into how governments carry out meritocratic selection due to data availability. By further observing the pool of applicants for the universe of government postings, I also show that different screening strategies employed by the public sector affect who applies and ultimately becomes a public servant, with a direct impact to the gender composition of the workforce.[7]

My findings also contribute to the literature on discrimination in hiring. Papers including Goldin and Rouse (2000), Bertrand and Mullainathan (2004), and Kline et al. (2022) have made great headway in documenting the existence of discrimination in hiring. But because of the nature of their evidence, prior work has been unable to identify whether disparities are driven by evaluators or the hiring practices used by employers. Combining the unique ability to observe the implementation of hiring practices with an empirical setting that isolates changes in screening methods, I show that hiring practices are a key source of gender gaps. Screening tools ultimately regulate the expression of bias, and directly targeting how screening is conducted can be an effective way to foster diversity.

This paper further provides insight to a strand of studies analyzing the importance of the degree of subjectivity of assessments in organizations. While much of the literature has focused on performance evaluations for workers once they are hired (e.g. Prendergast and Topel (1993), Prendergast and Topel (1996), MacLeod (2003), Frederiksen et al. (2020), De Janvry et al. (2023)), Autor and Scarborough (2008) and Hoffman et al. (2018) have analyzed the role of tests on firm hiring. This study augments this research by considering a broader set of choices employers have at hand to decrease subjectivity — remove discretion from tools already in use or replace, remove, or add tools with different levels of discretion. This provides important lessons: not all approaches to decrease discretion have the same implications for diversity, particularly when achieved through a removal of performance information about a candidate.[8] I estimate that a reduction in bias at the expense of losing screening precision may be ineffective on average,

[7] This also contributes to a strand of work that focuses on which factors can affect the applicant pool to public sector positions, including with the use of financial and career incentives (Dal Bó et al. (2013), Deserranno (2019), Ashraf et al. (2020)). Relatedly, my results add hiring practices and their role in shaping anticipated discrimination to factors influencing the gender make-up of job applicants (e.g. Flory et al. (2015), Kuhn et al. (2020), Abraham et al. (2020), Card et al. (2021), Delfino (2021), Flory et al. (2021), Kuhn and Shen (2023)).

[8] Moreover, a growing literature demonstrates the potential negative consequences to diversity from partially concealing or removing non-performance information about job applicants, e.g., ethnicity (Behaghel et al. (2015)), age (Neumark (2021)), credit information (Bartik and Nelson (2022)), criminal records and history checks (Holzer et al. (2006), Agan and Starr (2018), Doleac and Hansen (2020)).

with evidence from additional analyses suggesting that this bias-information trade-off may vary depending on occupational skill.

Finally, this work provides a methodological contribution to the growing use of text analysis tools in empirical economics. Researchers have relied mostly on *ad hoc* dictionary methods to parse and interpret information in text form into a predictor of underlying phenomena (e.g., Gentzkow and Shapiro (2010), Baker et al. (2016)). More recent methods are useful in applications with structured layouts to identify text regions (Shen et al. (2021)). In many cases, however, researchers are interested in extracting actual structured data from text, a task that is especially challenging when the text is displayed without regular layout and contains confounding information. The natural language processing algorithm I develop leverages semantic patterns of raw text surrounding numeric data, without requiring structured layouts. This query-based approach offers a text analysis tool to enrich new methods being developed in economics.

The paper is organized as follows. Section 2 provides the institutional details and the main source of variation used in the paper. Section 3 describes the natural language processing algorithm I develop to distill hiring records into data. Section 4 shows the aggregate results of reducing discretion on sorting and selection. Section 5 sets up a model and empirical strategy to identify the effects on gender gaps from different changes in screening methods. Section 6 concludes.

2 Institutional Details and Setting

2.1 Hiring in the Brazilian Public Sector

Brazil has been using a formal, merit-based civil service hiring system for decades, being considered a primary example of a meritocratic and legally professionalized bureaucracy (Grindle (2012)). Since the 1960s, over 80% of public positions have been filled exclusively through this mandatory competitive examination system, known as "*Concurso Público*".[9]

Public sector hiring follows strict rules and is organized in the following way. A number of constitutional requirements apply to all job selection processes at federal, state, and municipal levels, with each public employer implementing and conducting their own hiring. The first rule that all hiring processes must follow is transparency. Every step of the the *concurso* is pub-

[9] See Figure A.1 for a complete history of meritocracy implementation and public servant selection rules. There are strict limits to the number of openings and the types of jobs that can hire outside the *concurso* system. The remaining 20-30% public sector jobs legally exempt from formal civil service exams are temporary jobs, positions of trust, and commissioned posts. These posts are particularly common around politicians and specific bureaucrats, like congressional staff.

lished in a designated daily government gazette (similar to the Federal Register in the US). The second rule is that job openings must be allocated following the final ranking of candidates, which is exclusively determined by their scores in the screening steps.

Public hiring processes begin with a job announcement posting, called *Edital*. Job announcements need to detail the number of job openings, application requirements, and job attributes, and crucially for the research question, which screening methods will be used, and the weight of each exam toward a candidate's final score. Occupational attributes, including wages, are fixed by bureaucrats at high organizational levels and not subject to bargaining or employer discretion. The rules and details laid out in the job announcement are legally binding. In practice, this means that evaluators cannot ignore or modify a screening step.

The hiring process proceeds according to the general steps depicted in Figure 1. First, candidates apply to the job opening, have their applications screened based on announced requirements (e.g. attain the education level required), and have their names published on a subsequent journal issue. At this stage, the entire pool of candidates is publicly visible. The employer conducting the hiring process then publishes individual performance (scores) for each selection stage as the hiring process unfolds, including interviews and tests, identifies the candidates who are ultimately offered jobs, wait-listed, and hired. This concludes each *concurso*.

2.2 Impartiality Requirement

I exploit the introduction of a new requirement that altered one feature of public employee selection, while leaving the overall hiring procedures, rules, and the organization of public careers unchanged. The new requirement imposed that screening would have to be impartial, which in practice resulted in a shock to screening tools that reduced discretion.

In October 1988, Brazil passed a new Constitution expanding state guarantees and responsibilities as the country exited a 25-year period under military dictatorship. Known as "Citizen Constitution", the new legal framework modified several areas in the pre-existing constitutional text to align with the new social focus. This included the legal text organizing the hiring of federal public workers, which began to include a new provision mandating impartial screening.

The previous Constitution stated, regarding public sector hiring, that "*Public sector positions are accessible to all Brazilians* [...] *and hiring must be conducted through formal process* (*concurso*) *using exams or exams and candidate qualifications*" (1967 Constitution of Brazil, Section 7, Article 95). The 1988 Constitution amended this text by including a provision that the selection process should follow the principle of "impartiality" (*impessoalidade*). The new requirement aimed to codify the understanding that practices that could give an unfair advantage to an individual would violate the new Constitution and therefore be legally challenged. There were no specific directions or implementation prescription given, only that all hiring had to comply with the stated requirement.

2.2.1 Implications for the selection of public employees

The impartiality requirement generated a variety of changes to screening steps aimed at decreasing discretion. That is because, due to the lack of specific implementation rules in the Constitution, federal bureaucrats and government legal experts at high institutional levels provided different guidelines to public sector hiring. Keeping with tradition of organizing rules related to public servants around careers, discussions on new hiring practices were made at the occupation level. These centralized recommendations would then be implemented downstream by all federal employers.

Changes in screening methods followed two criteria. The first criterion focused on the legality aspect, which led to two universal changes. First, all occupations using written tests had to impose a blind implementation. Second, occupations could no longer screen candidates using exclusively non-written methods. The conclusion by government officials was that allowing for a candidate to be identified in a written exam could enable unfair treatment, and that relying only on subjective screening steps could lead to hiring decisions without a clear way to validate candidate performance (lack of "receipts").[10]

The second criterion was that, given that modifications were deemed necessary, those were made striking a balance between compliance with the impartiality requirement and maintaining screening tools considered important in a particular job. These considerations have been widely cited in Brazilian courts, which confirms their relevance in how treatment assignment was determined. For instance, in a 1993 legal proceeding (TC 013.813/93-5) involving a hiring process for journalists by the Federal University of Pernambuco, the Federal Court of Accounts (TCU) stated that a practical exam in that specific occupation was desirable, since "*journalists are commonly on camera and therefore observing their identity is necessary*".

[10] Figure A.2 shows an example of the language that began being added to job processes post-1988 about blind exams. In this selection process for federal judges published on September 4, 1989, a rule states that candidates identifying themselves in written exams will be excluded from the hiring process. This language was virtually nonexistent in announcements before the introduction of the impartiality requirement. In line with that, the website Jusbrasil.com, which compiles over 100 million of litigation cases in Brazil, lists no civil proceedings involving applicants removed from job processes because they identified themselves in the exam until 1988, while after the new Constitution there are hundreds of cases related to non-compliance with blind exams.

2.2.2 Treatment timing and assignment

Preexisting screening methods in an occupation therefore drove which screening changes were adopted to comply with the impartiality requirement and reduce subjectivity, providing exogenous variation in screening tools. Empirically, 95% of hiring processes within an occupation in the federal sector modified screening in the same way, underscoring the top-down nature of treatment assignment, which was not determined by employers or evaluators conducting hiring. Each change in screening methods involved a variety of occupations across several dimensions. For example, kitchen assistants and civil engineers had screening modified in the same way, just like teachers and police officers were screened the same way.[11]

Important for my identification purposes, the introduction of the impartiality requirement did not attend specific policy goals. For example, if policymakers intended to increase the participation of minorities in government, the provision could have been added with the intent to reform hiring. If that translated into an active implementation, employers or occupations deemed more problematic (e.g. lower women representation, more intensive use of discretionary screening tools) may also have been treated differently. In contrast with these considerations, evidence from over 14,000 pages of the Constituent Assembly records that I investigate show a clear pattern of only procedural or technical discussions around the implementation of the new provision.[12]

The provisions in the new Constitution applied to public sector hiring at all government levels. However, because public servant selection is conducted by states and municipalities independently of the central authority, states had to individually pass the appropriate legal framework to comply with the new impartiality requirement. It took several years for the sharp shift in federal employer behavior with respect to hiring to trickle down to state agencies and governments. This means that, while state jobs are suitable as the control group around the Constitution introduction, this institutional feature also limits the duration for which I can credibly study post-1988 outcomes. Since the first changes in screening methods in state job announcements start to appear in 1991, I limit the baseline analysis in the paper until 1990.[13]

[11] See Table A.1 for several examples of occupations in each treatment.

[12] Constituent Assembly discussion records are retrievable, in Portuguese, here `https://www25.senado.leg.br/web/atividade/anais/constituintes#1988` and here `https://www.senado.leg.br/publicacoes/anais/constituinte/sistema.pdf`.

[13] Due to the lack of systematic roll-out across states, many of which officially passed the requirement much later, but already showed informal changes to screening gradually over time starting in the early 1990s, alternative empirical strategies exploiting time variation in adoption across federal and states are not well-suited.

2.2.3 Broader institutional context around the new Constitution

While the impartiality requirement effectively represented a shock to public sector hiring, its introduction occurred with Brazil's new Constitution, which bundles broader social and institutional changes, including the exit from a period under military dictatorship. This potentially introduces contemporaneous changes that could confound the analysis. Importantly, my empirical design is only subject to identification threats that shift unobserved attributes of federal jobs relative to state jobs. This could be the case if, for example, the Constitution established female-friendly job amenities for federal positions only. In contrast, factors like the expansion of services provided by the federal government or modifications in the tax code would not be consequential.

Alleviating this concern, there were no relative changes to labor markets between federal or state governments or even between public and private sectors more broadly. Institutional changes introduced around the same time like the shortening of the workweek or extension of maternity leave were rights extended to all formal jobs in the economy.

3 Opening the Hiring Black Box

3.1 Raw Text Sources

The raw data used in this paper come from over 35 million of official government journal pages from Brazil (known as "*Diário Oficial*") since 1980. These gazettes are similar to the Federal Register in the US and publish the universe of public notices spanning public procurement processes, executive orders, and information on public servants, including hiring. Every government branch maintains its own decentralized repository with daily scanned issues of official journals. Many journals require a page-by-page access point.[14]

3.2 A New Approach to Transform Unstructured Text Into Data

3.2.1 Practical challenges

While these documents provide a rich source of hiring records, two main challenges prevent the direct retrieval of data on selection processes from Brazil's government gazettes.

First, there is no systematic way to search, identify, or link job hiring processes in the government journals. Hiring processes are mixed-in with extraneous and highly confounding

[14] Table A.2 shows a complete list of the separate government entities used to retrieve the gazettes, as well as when issues first become available online.

information, and postings from the same process published over time do not have an identifier that enables linkage. Information confoundness is a key issue which makes off-the-shelf text analysis tools based on similarity or proportionality ineffective.[15]

The second challenge is that, even if job processes could be identified and linked, text content in the government journals is highly irregular and unstructured. This is because there is no pre-determined layout that public agencies must follow when publishing notices, which results in an unwieldy variety of "shapes" to be handled by conventional scraping targeting ex-ante known structure patterns (e.g. a table with certain layout or free text organized in two columns).[16] As a consequence, an alternative approach to extract information not based on layout-pattern recognition is necessary.

3.2.2 Two-step natural language processing algorithm: Summary

To address all of these challenges, I develop a two-step natural language processing algorithm that allows me to first define the relevant text portions from highly confounding text, attribute a posting to a unique job hiring process, link its different postings, and finally transform unstructured text into data. The algorithm generalizes a search query with learning and can be applied to a wide variety of empirical settings that follow the same general extraction data problem and goal of this paper.

Because the algorithm's conceptualization and progression involve an extensive number of rules and technical considerations, I detail its implementation in the Appendix B. To fix ideas here, I use Figure 2 which summarizes the implementation mechanics and data output of the algorithm using one job process as example.

Figure 2 shows screen captures of the raw government gazette pages containing the **(a)** announcement, **(b)** list of applicants, and **(c)** final results for a job process in the federal government to select public prosecutors. For visualization purposes, the announcement and list of applicant posts trimmed and additional posts containing exam scores before the final results are not shown. Text snippets bounded in red are extraneous posts unrelated to the target hiring process, while text bounded in green represents the correct contents. Note that the target text on the first panel is organized in paragraphs within numbered sections, on the second in

[15] Standard text analysis algorithms that are increasingly popular in economics (like term frequency-inverse document frequency and cosine similarity) are poor tools for connecting different text corpora based off *exact* text vectors. Even the sophisticated lexical fingerprinting tools used to detect plagiarism would still rely on the resemblance between text documents that might not be informative for linking purposes. These algorithms require calibration that is context-specific, demanding supervision in a large number of cases, drastically decreasing gains to automation and resulting in a large number of type I and II errors.

[16] See Figure A.3 for some examples. I document over 200 different text layouts, with multiple variations within the same broad layout type.

continuous-like text, where applicant names are divided by a delimiter (;), and in tables in the last panel.

There are two important ideas behind the architecture of the algorithm's first step: posts in the government journals are temporally ordered — a job announcement must be published before its hiring decision. This means that when searching for a complete job process, the algorithm first identifies an announcement, then proceeds to journals published after the announcement date as it searches for the other posts.

Second, because these are official records subject to legal enforcement, each post must have enough information to enable a human reader to distinguish one job process from another. How much information a computer needs to perform entity linking as efficiently as humans is a complicated and actively-discussed problem, but I consider that there must exist a sufficient set of attributes comprised by a message keyword (job), sender (establishment), and keyword feature (which may be needed to resolve ambiguity). In Figure 2, the message keyword takes the correspondence *public prosecutor* and the sender is the *Office of the Federal District Attorney General*. Because that establishment did not have any other job hiring processes for the same job title ongoing until the final result is detected, establishment and job are sufficient to link the different posts of this hiring process.

Having determined that the three posts shown in Figure 2 belong to the same hiring process, the next step is to extract the data. For simplicity, I illustrate this step focusing on the shaded green area in panel **(c)**, which includes the final scores for all job applicants who made to the end of selection. In this step, the algorithm exploits the semantic relationship within a text snippet (instead of across) leveraging the fact that, if I am interested in retrieving numeric strings (the scores), each number can be seen as a target of a message. Thus, every score must have a sufficient number of attributes: a sender (the evaluator), a receiver (job applicant), and a feature (exam type). This approach is flexible because it does not impose a physical relationship between these textual elements that will be carried to the data output; rather, it assumes that for a human to properly assign meaning to a number, these attributes must be in the neighborhood of that number.

3.2.3 Data validation and estimating sample

I organize the algorithm's pipeline that implements step 2 following the sequence in Figure A.4. Implementation performance can be thought as an inference problem: all text snippets extracted in the first step contain the true data, but also false positives (incorrect signals). The goal is to separate the two. To ensure the final dataset contain only correct information, the algorithm first maximizes stringency and therefore the number of false negatives — it throws away important information that should be part of the output. Then, to recover discarded information, I increasingly relax the search parameter's stringency.

In a world without computation constraints, this iterative process would continue until gains in data extracted between two iterations is negligible. At a given point in the pipeline, my data dimensionality spans anywhere from 8 to 60 billion characters, implying more than 2 trillion matching computations needed every time the NLP algorithm runs. To make extraction feasible, I continue to minimize false negatives in the data output until the final data has an accuracy above 90% and the cost of decreasing false negatives is greater than some threshold increase in the number of observations in the final data (e.g., a manual adjustment to the algorithm that takes 1 hour produces 5 additional job processes that were otherwise discarded).

While there are no official sources to validate the data extracted with my approach around the impartiality requirement introduction, later periods have external aggregate statistics that I can use to run evaluation exercises. Figure A.6 plots annual official records provided by Brazil's federal government starting in 1995 containing the number of selection processes, number of hires, share of female hires, and share of high-skill jobs. Next to each statistic, I construct the equivalent measure based on data generated by the algorithm. Though this is out-of-sample in the time-series, the raw source and overall quality and layout of documents are the same as in the 1980s and 1990s, therefore likely to provide a reliable benchmark. The correlation between the generated data and the official records is on average 99% for each series, with point error incidence lower than 2%.

For the analysis in this paper, the sample includes the federal government and the states with official gazette issues available online for the period. These states were Amazonas in the country's north region, Pernambuco in the northeast, Distrito Federal, Mato Grosso, and Mato Grosso do Sul in the central region, São Paulo — the largest and richest state — in the southeast, and Rio Grande do Sul in the south. I use all job processes with complete information on job requirements, screening steps, as well as candidate scores, final ranks, and job offers, if any.[17] I focus the analysis on the 1986-1990 period since states began jointly passing new state Constitutions with similar guidelines to the federal rules in the early 1990s. In the case of states, however, the enforcement of impartiality rules was much less organized, with some state employers changing hiring methods in the 1990s and others in the same state delaying implementation, which for example invalidates a staggered implementation design.

Figure A.5 shows the gender distribution of applicants in the resulting sample by occupation and skill level, and Table 1 provides summary statistics on education requirements,

[17] There are no reasons to expect systematic factors explaining why some states lack online archives of government gazettes available in the 1980s or why states in the central region is over-represented. Mato Grosso and Mato Grosso do Sul share the same digital archive provider and software, and the Distrito Federal (akin to Washington DC in the US) is the seat of the federal government.

screening steps, and job applicants for control and treated groups, before and after the reform. There is considerable gender sorting across occupations and characteristics of posted jobs also vary over time. These composition effects motivate the inclusion of job fixed effects in all my specifications.

4 Overall Effects of Decreasing Discretion in Hiring

This section shows the aggregate results on gender sorting and selection from a reduction in discretion in hiring. The unique ability to observe how public servants are screened, evaluation and outcomes from screening, and the applicant pool, enables the analysis to proceed in the following sequence. First, I verify federal jobs' compliance with the new requirement, also confirming that state jobs did not modify their screening around treatment. Then, by disentangling supply and demand I show that greater impartiality in hiring improved women's final scores relative to those of men, driven by higher marks in objective exams. This translated into a higher probability of hiring. Finally, I leverage the hundreds of occupations, across all skill levels, in the data to check how these results interact with different job characteristics of public servants.

4.1 Did the Impartiality Requirement Change Screening Tools?

I begin by investigating whether and how federal government hiring processes complied with the impartiality requirement. Specifically, I compare how an occupation hiring in the federal sector modifies screening in a variety of ways consistent with decreasing subjectivity and discretion, relative to the same occupation hiring in states.

Figure 3 compares average changes in screening methods for state jobs to changes in federal jobs. The statistics plotted are averages across within-occupation averages, which means that they are free of compositional changes in the supply of different jobs. Comparing similar occupations between treated and control groups is important because their relative make-up may be unbalanced (e.g. healthcare services are usually provisioned at the local and state levels) and fluctuate over time due to different shocks to state and federal governments, including political cycles and budget decisions.

As expected given the institutional context of the new Constitution, state jobs show no response in the use of written and non-written screening tools, which is also reflected in the same number of screening rounds implemented. This is also consistent with a lack of strategic response by state employers, underscoring the rigidity of how screening steps are defined in public sector processes. In contrast, federal jobs adjust sharply by decreasing subjectivity, both by reducing the proportion of non-written tests and increasing reliance on written tools. Because these variables capture a compliance margin without blinding effects, these results show that federal job processes also actively modified screening steps. This led to almost double the number of hiring stages on average, driven by the inclusion of blind written tests to the existing process.

4.2 Empirical Strategy

My baseline empirical strategy implements a triple-difference specification comparing changes in outcomes in the same occupation for federal and state jobs. Specifically, I run:

$$\begin{aligned} y_{ijt} = {} & \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j \right) \\ & + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i \right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i \right) \\ & + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i \right) + \delta_{j(o)} + \theta_t + u_{ijt} \end{aligned} \tag{1}$$

which tracks outcomes for applicant i in job process j in a given year t. If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$.[18] The dummy Female_i corresponds to whether the job applicant is a woman, and the specification controls two-way for occupation and year fixed effects. The triple-difference estimate β measures how gender gaps in the federal sector changed after the decrease in discretion compared to changes in gender gaps in the same occupation in state jobs. Relative differences between the same occupation in federal or state employers, including posted wages or prestige, are differenced out by $\delta_{j(o)}$. Year fixed effects θ_t also control for aggregate time-varying factors like inflation shocks and spending cycles.[19]

In all specifications, standard errors are clustered at the job process level. Because scores and offers are made within a selection process, this level of clustering allows for correlated factors across job applicants: a job process with greater candidate demand, a more stringent set of evaluators, exam contents, and so forth. I assign gender to job applicants using Brazil's Census Bureau *Gender of Names* database, which contains nearly 200,000 unique first names and their corresponding gender. The matching precision between job applicant names and gender from this dictionary is above 98%.

[18] Since the new Constitution came into effect in October 1988, I consider the post-treatment period as the beginning of 1989. In line with that, I start observing changes in hiring processes announcements in 1989.

[19] Specification (1) contains the additional δ_1 and δ_5 parameters since the unit of observation (job applicant i) is not nested within the treatment level (federal government). This generates variation to both parameters beyond the fixed effect δ_2. The interpretation of β remains the usual triple-difference treatment effect.

Assessing the plausibility of the identification strategy. The main identification assumption is that, in the absence of the institutional change, gender gaps in federal and state job processes would exhibit parallel trends. Figure 4 shows event-study style estimates of model (1) separately for treatment and control groups. The panel on the left plots gender application gaps and the one on the right gender score gaps. This visualization helps to conduct the usual pre-trend assessment, and also to inspect the behavior of gender gaps separately for federal and state jobs around the introduction of the new Constitution.

Pre-treatment paths for gender hiring gaps are parallel between treated and control groups, indicating an absence of pre-trends. Following treatment, the outcome in the federal government shows a sharp change as soon as new job processes under the new requirements start to roll out, while the gender hiring gap in state jobs remains with almost an identical magnitude relative to pre-1989. State employers show no sign of "front-running" the implementation of impartial screening by their legislative bodies, attempting to hire more men before the requirement came into effect outside my estimation window.

A potential concern in difference-in-differences designs is the violation of the Stable Unit Treatment Value Assumption (SUTVA) assumption. In my setting, the shift in how screening had to be conducted in the federal sector could have changed application rates to state jobs in ways inconsistent with this assumption. One example would include the possibility that women start applying more often to federal jobs in detriment of state jobs or that men directed job search to states, avoiding a decrease in discretion in the treated job processes. Panel (a) of Figure 4 shows that gender application gaps in state hiring processes remain stable over time, overlapping very closely with the outcome in federal jobs pre-treatment. This indicates, in addition to a lack of pre-trends again, no evidence of spillovers to control job processes.

4.3 Disentangling Hiring Supply and Demand

Observational data available to researchers often allow one to only observe hires as the outcome of a selection process. However, changes in screening can affect both employers' hiring decisions, as well as job seekers' application decisions. In order to separate between these demand and supply effects, one necessarily requires knowledge on the applicant pool, which is a unique feature of my data of public sector labor markets in Brazil. To see this more clearly, consider the following decomposition:

$$\underbrace{\mathbb{P}(\text{Female}|Hired = 1)}_{\text{Share of female hires}} = \underbrace{\mathbb{P}(Hired|\text{Female} = 1)}_{\text{Demand}} \times \underbrace{\mathbb{P}(\text{Female} = 1)}_{\text{Supply}} \times \underbrace{\frac{1}{\mathbb{P}\left(Hired = 1\right)}}_{\text{Competitiveness}} \tag{2}$$

which separates the choice of women to participate in the hiring process from the conditional probability of them being hired. Once adjusted by the competitiveness of the job process — the probability that an applicant of any gender is hired — they jointly determine the share of female hires. Systematic gender differences in the probability of who applies and who gets selected therefore determine gender gaps in hires.

While the share of female hires captures whether the workforce is becoming more diverse, separately estimating selection (demand) and sorting (supply) effects is important to understand how each margin adjusts to changes in the hiring process design, since they have distinct potential remedies and policy consequences. For example, suppose that the share of female hires at an establishment is low. This implies (i) a potential propensity of the employer to discriminate (under a set of assumptions) against women, (ii) lower quality or fewer qualified women applying for the job, or both. As a consequence, using $\mathbb{P}(\text{Female}|Hired = 1)$ amounts to conditioning on an endogenous variable (the available candidate pool), which can lead to incorrect inference about the existence of employer bias. This could result in the inaccurate enforcement of anti-discrimination laws. It also provides employers with little information about the sources of the observed lack of female representation and how to best address it.

On the demand side, biased screening can over-select men by making them seemingly better than equally qualified women. But the use of biased screening can also have a gendered effect on supply, leading to suboptimal hiring if more qualified minority candidates refrain from applying. That is because women may react to even perceived unfair treatment (Small and Pager (2020)).[20] That further interacts with gender differences in behavior involving competitive environments, which includes job selection processes.[21]

Exploiting the ability to separate these two important channels, I organize the results below on the overall effects of greater impartiality in hiring in the following sequence. First, I show how the new requirement affected gender score gaps, which leverage detailed records for applicant performance in each screening stage of a job process. Then, I measure changes in gender hiring gaps, i.e. the labor demand component $\mathbb{P}(Hired|\text{Female} = 1)$. Finally, I estimate gender gaps in supply responses ($\mathbb{P}(\text{Female} = 1)$), ultimately assessing the importance of each channel to affect employee diversity.

[20] Women are more likely to place greater weight than men on fair treatment, and the perception of fair treatment is more strongly linked to women's than to men's willingness to apply at a previously rejecting firm (Brands and Fernandez-Mateo (2017)). There is also evidence that women respond to nudges that they are "welcomed", as with ads stating a preference for diversity (Flory et al. (2021)).

[21] Extant literature shows that women are less likely to apply for promotions and less likely to enter tournaments due to a lower willingness to compete or self-stereotyping (Niederle and Vesterlund (2007)), Hospido et al. (2019), Bosquet et al. (2019), Coffman et al. (2023)), and tend to sort into female environments to avoid competing against men (Gneezy et al. (2003)).

4.3.1 Effects on gender score gaps

Table 2 begins by comparing final scores in the hiring process received by female and male candidates. Scores are standardized within each job process, so that they are comparable across all processes. Final scores of women increase by 0.07 standard deviation after the new Constitution is implemented, and the final scores of men decrease by slightly more. Combined, these effects imply a 0.14 standard deviation reduction in the gender score gap, which closed the pre-existing gap.

Next, I break down scores by screening tool type. Columns (4) through (6) show that written exam scores, which necessarily became blind for federal jobs, improve significantly for women's relative to men. The overall effect on the gender gap in written scores is about 0.13 standard deviation. Columns (7) through (9) repeat the analysis for scores in non-written exams, which, unlike written tests, were not directly altered in response to the impartiality requirement. Results show no changes in gender gaps for oral, interview, or practical exam scores, confirming that evaluators did not seem to adjust non-written scores strategically in response to treatment.[22]

4.3.2 Effects on hiring decisions

Next, I investigate how the closing of the gender score gap translates into hiring odds. Candidates' final scores determine their ranking in the job process, with the highest-scoring applicants being hired.

While the scores of the average woman applicant increased relative to men, this may not necessarily imply a narrowing of the gender hiring gap. For that to happen, women who were just not hired must be able to take over the position of the marginally hired men. If overall increases in scoring came from the highest or lowest performing women their hiring status would likely not change. Table 3 confirms that the higher female performance translated into improved hiring odds for female applicants. Columns (1) and (2) show that the probability of being hired for women increase while for men it decreases. Women become 0.3 percentage point more likely to be hired after the reform and men's hiring rates decrease by about the same magnitude.

Combined, the point-estimates imply a reduction of the gender hiring gap of 0.7 percentage point (column (3)), equivalent to 44% of the initial hiring gap. The fact that the gender score gap closed but the hiring gap only narrowed, albeit considerably, points to increased perfor-

[22] Note that the separate estimates on gender score gaps by screening tool type restrict the underlying sample to those hiring processes that had the tool before and after treatment. Thus, it excludes job processes that for example switched from a non-written to a written test, or that had a combination of both methods pre-treatment but dropped the interview. The analysis in Section 5 considers each specific change to screening methods.

mance effects somewhat dispersed across women's distribution. Raising scores for women at the left tail of the distribution is not enough to make them competitive, just like higher scores for those at the top who would already be selected do not convert to higher hiring odds. This suggests that the benefits of reduced discretion were not concentrated to applicants of only a certain quality.

4.3.3 Effects on application behavior

Next, column (4) in Table 3 estimates the supply response of female applicants, captured by the sorting term $\mathbb{P}(\text{Female} = 1)$. Application rates of women grew by 1 percentage point, for a baseline application rate slightly tilted toward female applicants in federal jobs. This shows that women application rates were relatively more responsive to greater impartiality in screening.

To gain further insight into the general forces behind supply movements, in Table A.3, I run job process-level versions of model (1), first with a focus on the number of job applicants. Estimates show some evidence of an overall decrease in the number of applications, although a statistically insignificant effect. This may be driven by an overall increase in the cost of participating in the average selection process, since there are more screening rounds after treatment. I formally investigate this possibility in Section 5, where I distinguish sorting responses to treatments that clearly reduced discretion without modifying the number and type of tools being used from the ones that did so. Results are consistent with women reacting to perceived fairer screening by increasing application rates.[23]

4.4 Additional Results and Robustness Checks

4.4.1 Job applicant quality

The prior analysis does not directly control for potential changes over time in the composition of job applicants, since it relies on a repeated cross-section from the perspective of candidates. Because of that, the quality of applicants could vary around the treatment. If better-quality women began to apply for federal jobs at higher rates than men, at least some of the decrease in gender gaps could be attributable to changes in quality.

While this potential effect could still be due to a decrease in impartiality — perhaps better prepared women react more strongly to fairer screening since their outside option may be more

[23] Another potential driver is the job process competitiveness, which tends to correlate with the size of the candidate pool, and is determined by budgetary and personnel management constraints on the supply of job openings. However, controlling for process competitiveness or announced job openings does not change these results.

valuable — it is an informative exercise to assess how much changes in quality matter for the results. To test for that, I include a job applicant fixed effect to the previous set of estimates, effectively controlling for changes in the distribution of any individual time-invariant characteristic. This strategy is enabled by two characteristics of my data. First, applicants' names and identifying information allow me to link them across *concursos* at any government establishment. Second, due to the highly competitive nature of selection processes, many individuals apply more than once.

Table A.4 replicates the results in Table 2 for final scores adding applicant fixed effects. The estimate for the increase in women's scores using the sample of repeat-applicants is larger than in the previous specification that allows for compositional changes, while the effects for men are more negative. In addition to confirming the baseline results, this specification further suggests that the average quality of the female applicant pool decreased, which is consistent with the response anticipated when the artificially high hiring threshold to select women pre-treatment is lowered. In summary, the decrease in the gender score and hiring gaps were not driven by more qualified women applying.[24]

4.4.2 Alternative empirical specifications

Table A.5 reports a series of alternative specifications of model (1). To assess the robustness of the baseline specification, I modify how standard errors are clustered, the set of fixed effects, controls, and how job applicants' results in the selection processes are measured. The estimated magnitude of the decrease in gender gaps is stable across all tests, which include clustering at employer, government, or occupation levels, and controlling for occupation skill, geographic region of jobs, interacted fixed effects, or applicant pool size.

4.4.3 Occupational differences in skills and female representation

The magnitude of the estimated treatment effects may depend on characteristics related to the occupation. Women may face more stereotypes for higher-skill jobs at the top of the career ladder, certain jobs main provide more easily observable productivity measures during screening, diminishing reliance on statistical discrimination, and thinner markets in highly-specialized high-skill jobs make taste-based discrimination more costly.

Table A.6 examines how the effect of increasing screening impartiality varies across the occupational skill level. Compared to male candidates in high-skilled occupations (college

[24] This interpretation imposes the assumption that time-varying unobservables correlated with quality did not change differentially by gender over time. That is, female repeat-applicants did not become better relative to men repeat-applicants.

degree or more), female scores increase by 0.2 standard deviation, and their hiring rates by 1.1 percentage points. This represents a magnitude 50% greater than the average effect estimated across all skill levels. Increasing impartiality for low-skill jobs — that require at most a high school diploma — does not have an effect on scores or hiring rates of women.[25]

Table A.7 assesses how my baseline estimates differ by the feminization rate of the occupation, assigning a job title to one of the three groups: female-dominated, neutral, or male-dominated. Comparing columns (A1) and (A3) reveals that the final score gap narrows by approximately 0.3 standard deviations for both female- and male-dominated occupations. Investigating the effects separately by gender in panels (B) and (C) reveals that women's scores increase more in male-dominated occupations, while men's performance falls the most in feminized jobs.[26]

The overall decrease in discretion closed the gender gap in hiring by almost half and increased female representation in Brazil's public sector. This was driven by both an effect on both employers' hiring decisions, as well as job seekers' application decisions with a higher share of women induced to apply. In the next section, I investigate which of the changes in screening adopted to reduce discretion in hiring account for these results.

5 Implications of Different Strategies to Reduce Discretion

The analysis so far shows that application and hiring rates of women increased in response to the decrease in discretion in job processes for federal employees. These aggregate estimates, however, mask heterogeneity in how screening methods were modified to reduce discretion. Distinguishing between these different strategies is important, since alternative ways to reduce discretion can have distinct consequences for diversity. The choice of screening tools may impose a bias-information trade-off: removal of subjective steps may reduce bias but also discards

[25] One possibility accounting for this result is that the specific way to reduce discretion varied along the skill distribution. Another, that the same change produces different effects depending on the skill level. For example, screening tools that give evaluators more discretion may be more valuable in higher-skill settings as private signals observed by evaluators may be more important to determine worker quality relative to lower-skill settings. I investigate these points more formally in the next section.

[26] I measure the degree of feminization as the gender segregation of the occupation in the public and private (when applicable) sectors, the total share of women applying to job openings of that occupation, and when the occupation requires a specific college degree (e.g., structural engineering), the national gender make-up in the major. These factors broadly align. Occupations with less than 40% female participation are defined as male-dominated, 40-60% neutral, and above 60% female-dominated. The overall response here is consistent with the previous results across all occupations, where the downward adjustment in men's scores is quantitatively larger than the increase for women's, $\left(|\widehat{\beta}^F\,[\text{Female}=0]\,| - |\widehat{\beta}^M\,[\text{Female}=0]\,|\right) - \left(|\widehat{\beta}^M\,[\text{Female}=1]\,| - |\widehat{\beta}^F\,[\text{Female}=1]\,|\right) \approx 0.04$. These results can be seen as the "benefit" men derived in female-dominated occupations being larger than the penalty women faced in male-dominated sectors. Similar results carry for gender hiring gaps.

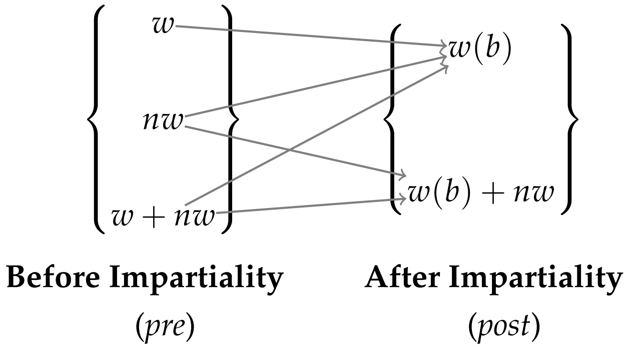
information; while adding objective tools constrain evaluator behavior and bias expression, but can itself introduce disparate impact.

To estimate the equity implications of different screening strategies, I take advantage of two main features of my setting. First, the unique ability to observe the screening methods used in each hiring process, rarely disclosed by employers — the hiring black box. Second, pre-existing differences in how occupations screened before generates exogenous variation in screening changes as a response to the reform, which capture different ways to reduce discretion. Some occupations made the implementation of tools already in use more impartial by partially or fully blinding screening. Others changed the combination of tools being used: either by removing or replacing exams with high discretion (like interviews), or introducing a tool with no discretion (blind written tests).

5.1 Treatments Generated by Discretion Reduction

Screening methods used in job selection processes belong to one of two categories: *written* (w) and *non-written* (nw). *Written* tools include open-ended and sometimes multiple-choice exams. *Non-written* exams comprise oral and practical examinations, and interviews. Screening may use either written or non-written exams, or a combination of both. This results in varying degrees of discretion and subjectivity across occupations, which is the level of assignment of screening methods in Brazil's public sector.

Before treatment, an occupation uses one of the following three possible screening tools: $pre \in \{w, nw, w + nw\}$. With the impartiality requirement, the exclusive use of nw is discontinued and written exams become blind. This implies that there are two possible combinations of exams post-treatment: $post \in \{w(b), w(b) + nw\}$. I define a treatment d as a mapping from *pre* to *post*, that is, a change in screening methods. This approach is consistent with the institutional setting, since the pre-existing mix of screening tools in an occupation determines which post-treatment types of exams will be used. This one-to-one mapping between an occupation and the set of screening tools used reflects the fact that 95% of job processes within an occupation follow only one treatment d.



Specifically, occupations using w before 1989 can only blind fully blind the tool already in use, so that $d = w \longrightarrow w(b)$. Following the same reasoning, those occupations with nw pre-treatment would need to introduce a blind test, generating one of two potential treatments: $nw \longrightarrow w(b)$ or $nw \longrightarrow w(b) + nw$. Finally, occupations with historical use of $w + nw$ can be assigned $w + nw \longrightarrow w(b) + nw$ or $w + nw \longrightarrow w(b)$.[27]

In total, the impartiality requirement generates 5 possible treatments, with each occupation being assigned one type of change of screening tools. Though all of these treatments intended to reduce overall discretion in job processes, the implication of alternative treatments to gender gaps at face value is unclear. That is because different design choices modify bias and precision during selection. For example, how to evaluate whether we should expect $nw \longrightarrow w(b)$ to increase female representation more than $nw \longrightarrow w(b) + nw$? The first treatment trades screening and bias from one tool for another: even if interviews allow for more evaluator bias to be expressed, written tests could impose a greater disparate impact. The second alternative treatment increases overall screening precision by adding a blind test, but could also result in a countervailing force due to added tool bias. To sharply answer questions of this kind, one needs a framework that delivers predictions for gender equity that are empirically relevant.

5.2 Theoretical Framework

While targeting an overall decrease in discretion, different modifications to hiring methods have unclear consequences for gender hiring gaps. That is because most potential treatments generated by the impartiality requirement may introduce a trade-off between bias and infor-

[27] Figure A.6 traces out the complete potential treatment space and gradually eliminates those ruled-out by my empirical design or the underlying setting of public sector hiring. Cases shaded in gray in the space grid are ruled out in a sharp difference-in-differences design (perfect compliance). Of the six remaining transition cases, $w \longrightarrow w(b) + nw$ accounts for less than 1% of transitions in the data.

mation. This subsection sets up a simple theoretical framework that generates empirical predictions on how each change in screening tools impacts hiring rates of men and women.

5.2.1 Model summary

I build on the canonical model of statistical discrimination with fixed wages (Phelps (1972), Autor and Scarborough (2008)), consistent with the setting in Brazil's public sector. My model introduces two new features: tool discretion, which regulates evaluators' expression of their own bias, and tool bias, which allows for disparate impact independently of evaluator behavior.

The first feature of my model is tool discretion, which captures the fact that different tools may allow for different degrees of evaluator bias expression. Evaluators have the task of selecting employees with a mix of screening tools delegated to them by the employer. I allow managers to have a systematic bias for a certain demographic group. The term could be interpreted as taste-based discrimination, or any other type of systematic bias, such as rewarding social skills that do not predict productivity, but favor a certain group. However, the expression of evaluator bias is regulated by how much discretion a specific screening tool enables. Interviews allow for high levels of discretion due to their subjective nature, while the results from written tests are more easily observable to the employer, making bias expression more costly. In contrast, statistical discrimination expression in the model is independent from the degree of discretion of screening practices used. Managers base their prior of a candidate's productivity on her group membership. When candidate identity is concealed, they instead resort to the population mean. Evaluator bias and statistical discrimination combine to represent disparate treatment.

The second addition I make is to allow screening tools themselves to be biased. Independently of evaluator preferences, certain screening tools may disadvantage a particular group. For example, if written tests reward risky behavior by penalizing wrong answers without measuring productivity, women may be disadvantaged and the screening practice would lead to disparate impact. The role of tool bias is equivalent to adding systematic noise to the productivity signal provided to evaluators, favoring less productive applicants of the favored group.

The model pins down gender hiring gaps in each of the five potential treatments that correspond to the different ways to reduce discretion as a function of evaluator bias, tool bias and precision. These reduced-form predictions will guide the interpretation of the treatment effect estimates later in the section.

5.2.2 Environment

An employer (the principal) delegates the screening of a pool of job applicants to an evaluator. The candidate pool comprises individuals from two demographic groups, $x = \{m, f\}$, corresponding to a minority and majority group, female and male, respectively. The employer bases the hiring decision on some indicators of productivity $\theta = \{s, \eta\}$, observable only by hiring evaluators, which noisily measure a candidate's true productivity level, y. The productivity of job candidates is distributed as: $Y \sim N\left(\mu_0\left(x\right), 1/h_0\right)$, where the mean $\mu_0(x)$ is allowed to depend on group membership, and h_0 is assumed to be independent of x. I assume women to be the minority group, in the sense that women's average productivity is perceived to be lower than men's, $\mu_0(f) < \mu_0(m)$.[28]

The employer's objective is to hire a proportion K of workers that maximize expected productivity.[29] But the evaluator objective is imperfectly aligned with that of the firm. The evaluator cares both about productivity and her bias toward a group ($\pi(x)$), which must be jointly maximized when hiring job applicants by

$$u\left(y, \pi\left(x\right)\right) = y + (1 - c_\theta)\pi(x) \equiv y + d_\theta \pi(x)$$

where c_θ is a cost or penalty function defined over $c \in [0, 1]$. This component captures the cost that the evaluator faces by expressing bias, i.e., reporting to the employer a value of a candidate's measured performance that differs from the signal provided by the screening tool. I assume this cost decreases in the subjectivity degree of the screening signal according to $d_\theta \equiv (1 - c_\theta)$. Scoring a candidate's written test differently than the publicly-observable signal poses a much higher threat of detection than underscoring someone after an interview because the person did not appear to be friendly or an "appropriate fit". The degree of discretion enabled by a screening tool plays a central role in the model, as it loads on the bias term and determines its relative weight in the evaluator's utility.

[28] Alternatively, one can model signal precision to depend on group membership (for example, Aigner and Cain (1977), Lundberg and Startz (1983), Cornell and Welch (1996), Bartik and Nelson (2022)). Similar to Autor and Scarborough (2008), I assume it to be independent of group membership to focus the analysis on the new features that I introduce in the model.

[29] The constant aggregate hiring rate K is assumed to be below 50%, as is the hiring rate of each demographic group. This assumption is motivated by the fixed number of positions in the job announcements for public servants, as well as the average probability of being hired being around 4% in the data.

5.2.3 Hiring decision

Workers are hired based on a noisy indicator of productivity generated by screening tool θ:

$$\theta^* = y + \nu_\theta(x) + \varepsilon_\theta, \quad \varepsilon_\theta \sim N(0, 1/h_\theta)$$

where the productivity signal from tool θ is allowed to be mean-biased by including the bias term $\nu_\theta(x)$. This mean-shifter can be interpreted as the disparate impact of the screening tool, which favors men when $\nu_\theta(m) > \nu_\theta(f)$.[30] After observing θ^*, the evaluator updates her assessment of expected productivity of candidates as

$$\mu(x, \theta^*) = \theta \frac{h_\theta}{h_0 + h_\theta} + \mu_0(x) \frac{h_0}{h_0 + h_\theta} + \nu_\theta(x).$$

The expression above represents a weighted average of perceived productivity of group x and the signal provided by the hiring tool, with weights determined by the relative precision of the signal with respect to productivity dispersion. A direct implication from the updated group mean $\mu(x, \theta)$ is that when a screening method is less informative, the evaluator relies more heavily on the group prior.

The hiring decision that maximizes the evaluator's objective function satisfies the rule Hire $= \mathbb{1}\{\mu(x, \theta^*) + d_\theta \pi(x) > k_\theta\}$, where k_θ is the threshold that yields a hiring rate of K. The solution to this problem gives the hiring rate $1 - \Phi\left(z^*_\theta(x)\right)$ with hiring threshold $z^*_\theta(x)$ for each group. Due to the linear form of the signal expression above, the hiring threshold for group x decreases when the group mean productivity is higher, when tool bias favors the group, or when evaluators are biased toward x (regulated by the discretion in the tool being used).

5.2.4 Model predictions

To match the model predictions to our empirical setting, consider the screening tool choices available to employers before and after the impartiality requirement. Before the Constitution, employers could use *i)* a written test, which generates a signal s; *ii)* a non-written test with signal η; or *iii)* a combination of both written and non-written tests.[31] After the requirement, employers in the federal government are constrained to screen candidates using only a blind written test ($s(b)^*$) or a combination of non-written and blind written tools.

[30] I consider that the different screening tools provide signals of productivity determined by one factor, which is to say they measure the same skill. A two-factor model would reformulate productivity as $y = y_1 + y_2$, where, y_1 could represent soft skills and y_2 hard skills (similar to Frankel (2021)).

[31] Within the model, I do not distinguish whether employers use one or multiple tests of the same type. A richer formulation that would incorporate the supply of candidates could take into account the number of exams and therefore the length of the screening process as an application deterrent.

This means that making the screening tools already in use more impartial changes the signals used by evaluators from s^* to $s(b)^*$ (full blinding) or from s^*, η^* to $s(b)^*, \eta^*$ (partial blinding). In the cases where the existing mix of screening tools is changed, productivity signals are modified in three potential ways. From an a non-written η^* to blind test $s(b)^*$; from η^* to $(s(b)^*, \eta^*)$; or from (s^*, η^*) to $s(b)^*$.

The full model solution and predictions to gender hiring gaps for each one of the five changes in screening methods is in Appendix C. To summarize the most important insights, I analyze Figure 5, which contains three panels. Panel **(a)** shows the model-implied gender hiring gap when screening only uses w (gray) or alternatively only uses $w(b)$ (green). The gender gap is plotted as a function of evaluator bias toward a gender. Unless evaluators are biased toward women, the complete removal of discretion reduces the gender gap since it completely eliminates disparate treatment.

Panel **(b)** compares gender gaps when screening only contains a subjective step to when a job process adds a blind written stage. This plot varies on the horizontal axis the relative tool bias of nw with respect to $w(b)$. Unless the disparate impact of blind written tests is much larger than that of interviews, the increase in screening precision in $w(b) + nw$ leads to smaller gender hiring gaps than nw. Lastly, in the case where interviews are removed in panel **(c)**, gender hiring gaps narrow only if the reduction in total bias from removing nw offsets the loss in total screening precision.

5.3 Empirical Strategy for Multiple Treatments

To estimate the causal impact of each change in screening methods individually, I implement the following triple-difference model:

$$\begin{aligned} y_{ijt} = {} & \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j \right) \\ & + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i \right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i \right) \\ & + \beta_d \left(\text{Federal}_j \times \text{Post}_{t(j)} \times \text{Female}_i \right) + \delta_{o(j)} + \theta_t + u_{ijt} \end{aligned} \tag{3}$$

which compares outcomes (y_{ijt}) for female candidates relative to men (Female_i) participating in job processes for the same occupation ($\delta_{o(j)}$) that had screening practices changed following treatment d. Standard errors remain clustered at the job process level, as with the previous specifications. Overall, this is a similar model to (1), except that it is implemented separately for each treatment group.

Guided by the theoretical framework in 5.2, each triple-difference coefficient has different interpretations, depending on the conditioning on d. To fix ideas for the exposition of results below, consider the case $d = w \longrightarrow w(b)$. This treatment was experienced by a variety of jobs like school officers and translators whose hiring in the federal government reduced discretion through only blinding the existing written tests. Because the mix of screening tools in use was kept unchanged, overall tool bias and screening precision are not affected by treatment. Under the identifying assumptions, the only effect measured by $\widehat{\beta}_{w \longrightarrow w(b)}$ is the response in the gender gap when all sources of evaluator bias are completely removed.

Assessing the plausibility of the identification strategy. I discuss the identifying assumptions and provide evidence in their support next. Similar to the binary treatment design in (1), multiple treatments require the assumption of parallel trends between a treatment and its equivalent control group (the same occupation in state jobs).

One additional consideration in this context is the role of selection into different treatments by occupations with the same pre-existing screening mix. Because treatment assignment decisions were made by high-level bureaucrats at a more aggregated level (occupation) than my unit of analysis (job process), employers and evaluators have limited ability to deviate from top-down guidelines. As a consequence, over 95% of occupations followed one treatment type in job processes across different employers. This results in each treatment being generated by many occupations, but each occupation being assigned to one treatment. Crucially, a wide variety of jobs along several dimensions make up each treatment. These institutional features therefore provide plausible exogenous variation to changes in screening methods.

To investigate this more formally, it is useful to view the institutional portion of treatment assignment as generating a set of potential instruments, where occupations determine changes in screening tools. This is an advantage of my setting since the presence of many occupations in each treatment can be used to leverage overidentification-style tests as in Angrist et al. (2017).

As a first step, consider a set of pre-treatment covariates including occupational skill level and degree of feminization, as well as share of female applicants and competitiveness of the job process. These factors can either directly determine selection decisions into treatment or correlate with variables I do not observe that determine assignment on their own. Table A.8 shows almost no systematic relationship between which treatment is assigned to occupations with the same initial mix of screening tools and these observables.

An additional test addresses the possibility that additional relevant factors in the information set of policymakers may not correlate with the covariates. But to the extent that these unobservables would also matter for gender hiring gaps, they would plausibly lead to pre-treatment differences in the outcome paths of jobs that sorted into alternative treatments. For instance, civil engineers and teachers were hired using $w + nw$ before the impartiality requirement and switched to $w(b)$ and $w(b) + nw$, respectively. If attitudes toward women were different or changing between the two occupations so that the choice of each treatment was based on treatment effects, one would expect pre-treatment outcomes to differ between the two careers.

Pre-trends in outcomes in Figure 6 therefore provide a check for systematic differences across occupations with the same initial set of screening methods. The figure shows dynamic effects plotted from an event-study style version of model (3). Each panel divides occupations according to *pre*, so that they compare treated effects in occupations within subsets **(a)** w, **(b)** nw, and **(c)** $w + nw$. Across all five treatments, gender hiring gaps between state and federal job processes were not statistically different from zero prior to the new Constitution.

This provides information above and beyond the standard pre-trend inspection test. Gender hiring gaps both followed parallel trajectories, as well as had indistinguishably estimated magnitudes between the pairs of occupations, relative to control job processes. This indicates that, on average, selection of an occupation into a particular treatment shows no relationship with pre-existing differential gender hiring gaps, which in turn would be likely to pick up unobservable pre-treatment factors.[32]

Combining the evidence presented and how screening decisions are made in Brazil's public sector, I can test directly for violations of selection into treatment as in Angrist et al. (2017). The Sargan (1958) test, where the endogenous regressor (equivalent to the policy variable $\text{Post}_{j(t)} \times \text{Federal}_j$ in a difference-in-differences specification of (3), ran separately by gender and conditioning for the set of pre-treatment tools, *pre*) is instrumented with occupational codes, imposes an intuitive restriction. If all occupations in d adopted that treatment as-good-as randomly with respect to gender disparities in outcomes, one should not expect statistically significant discrepancies of estimated treatment effects between them. Table A.9 confirms this by overwhelmingly failing to reject the overidentifying restrictions for gender gaps in scores and hiring odds as outcomes.

5.4 Main Results

I now proceed to estimate model (3) in five separate regressions for gender final score gaps and gender hiring gaps. Table A.10 shows treatment effects of final scores. For conciseness, I center my discussion in Figure 7, which conducts the same analysis using the gender hiring gap as outcome. Since job offers are solely based on final scores and job openings, any improvement

[32] Differences in pre-treatment outcomes between units can be used formally to assess a plausible channel of selection. In Ghanem et al. (2022), who study a binary treatment difference-in-differences setting, large pre-existing differences in the outcome between treated and control make the assumption of parallel trends less likely to hold when there is selection e.g. based on gains. In my setting, the comparison being made is between units receiving different treatments instead of between treated and control units.

in women's hiring rates relative to men's implies a decrease in the average gender final score gap.

Figure 7 analyzes in three groups the five treatment types generated by the impartiality requirement. Each grouping has the same baseline (pre-treatment) screening tool mix *pre* — w, nw, or $w + nw$ — for which I then estimate treatment effects. Note that, at least observationally, the hiring gap is much larger in job processes relying solely on non-written stages.

Reducing discretion of tools already in use. I first analyze the two treatments that eliminated discretion of tools already in use, without modifying screening stages. With the set of screening tools fixed, overall tool bias (disparate impact) and precision remain unchanged. The two treatments under this category blinded the hiring process either fully ($w \longrightarrow w(b)$) or partially, depending on the relative weights between w and nw ($w + nw \longrightarrow w(b) + nw$).

Starting with occupations that fully blinded job processes in the federal government, the estimated impact on the gender hiring gap is a decrease of 0.5 percentage point, or about 33% of the baseline gap. This treatment captures the causal effect of removing all sources of evaluator bias, and therefore reveals that they were an important force driving gender disparities.

Now consider the treatment that starts with the screening mix $w + nw$, blinding the written stage. Due to the presence of non-written steps, the estimated effect may not be similar to the one in the previous treatment. If the written stage has a small weight toward the final score or if evaluators change their scoring in non-written exams as a spillover from blinding, the gender gap may not necessarily narrow. Figure 7 confirms this possibility, indicating that the average treatment effect of partially blinding in hiring processes is null.

With the pre-determined weights at hand, I can test for the two hypotheses in Table A.11, which breaks down the aggregate null estimate by analyzing how gender hiring gaps responded depending on the weight of blind written tests toward the final score. The firs conjecture — treatment effects on hiring gaps should be stronger for hiring processes with greater weight in $w(b)$ — is confirmed. Columns (2), (4), and (8) show that when enough weight is exogenously placed on the written test (at least 50% of the final score), blinding has a positive and significant effect on women's final evaluation scores and hiring probability. Second, the incentive for evaluators to "compensate" men's scores in interviews would be larger the greater the weight placed on written tests, where they can no longer identify candidates. Results in columns (6) and (7) point to no evidence of strategic response by evaluators in non-written tests, which aligns with the previous findings in Section 4 of broadly unchanged performance in non-written stages of women relative to men following treatment.

Adding an objective tool. While a subset of occupations decreased discretion without modifying screening methods, others changed the tools being used. The first treatment of this kind I analyze shifted screening following $nw \longrightarrow w(b) + nw$, since relying exclusively on a highly subjective stage would not comply with the impartiality requirement.

This treatment provides lessons to a common approach in other contexts where a standardized, objective tool is introduced to reduce discretion in screening. My theoretical predictions show that adding $w(b)$ increases total screening precision without introducing evaluator bias, since the blind tool gives no discretion to evaluators. Higher screening precision in turn favors minorities since it increases the weight placed on performance signals, as opposed to the (men-favoring) priors. As a consequence, another productivity signal makes the optimal hiring threshold less dependent on the unconditional group mean, reducing the level of statistical discrimination.

On the other hand, a common objection against requiring standardized or written tests is that they could disadvantage women or minorities through a disparate impact channel, if for example women are worse test-takers in a way that is not predictive of worker productivity. But based on my theoretical predictions, as long as they are less biased as a tool relative to interviews, overall bias in screening will still be lower, increasing women's hiring rates. This reinforcing mechanism between bias reduction and better precision leads to large treatment effects in practice. The estimated decrease in the gender hiring gap is about 5.9 percentage points, or 35% of the initial hiring gap from using nw.[33]

Removing a subjective tool. The next treatment I analyze also has a pre-treatment mix using both tools, $w + nw$, but instead of only blinding the written exam, also removes the subjective stage. A mix of high and low-skill occupations — cleaners, programmers, civil engineers, and nutritionists — make up this treatment. Similarly, occupations with the same initial set of tools that kept nw include judges, police officers, and cooks again consistent with discussions on how to modify screening being related to idiosyncratic factors of that job. The motivation for this approach in a broader sense is intuitive: if evaluator bias matters and bias is particularly enabled by tools with high discretion, their use may be an important factor for gender disparities.

My estimates in Figure 7 reveal that removing subjective stages has no effect on the gender gap. This is consistent with a bias-information trade-off where the loss in screening precision from removing a productivity signal offsets the potential gains from eliminating disparate treatment and impact with the use of the tool. Less overall precision in screening

[33] Note that in the context of public sector hiring in Brazil, an additional force contributing to a smaller gender gap is a mechanical reduction in the pre-determined weight given to nw, which decreases from 100% to as little as 20% following treatment.

leads evaluators put more weight on their prior, making identifying above-average women more difficult. This result underscores the importance of employers weighing the screening precision loss and net gains from bias reduction when removing a tool.

Replacing a subjective tool. Lastly, I study the treatment where occupations changed the screening design completely — replacing nw with $w(b)$. In this case, occupations traded screening precision and tool bias from non-written exams for those of written tests, with ambiguous predicted effects on the hiring gap. Because of blinding, there is also the complete removal of all sources of evaluators bias. The gender hiring gap decreases by almost 7 percentage points, starting from a baseline gap of almost 17 percentage points. When benchmarked against the initial gap level, the estimated magnitude implies a decrease in the gap of 41%.

Given that removing evaluator bias in nw from the treatment $w + nw \longrightarrow w(b)$ is insufficient to increase female hiring rates, the results for the replacement of subjective tests suggest that either screening precision in written tests is higher than in interviews — since an increase in overall precision is minority-favoring — or that w imposes a smaller disparate impact on women than nw.

Gender effects on sorting. To conclude the main set of results with multiple treatments, Table 4 shows how different changes in screening affect the supply side: job seekers' application decisions. Consistent with the idea from Section 4.3 that perceived discrimination or unfair treatment during hiring may discourage minorities from applying in the first place, columns (1), (4), and (5) all show that blinding a pre-existing written test increases the participation of women in the applicant pool (from 2% to 4% on average).

Column (2) shows that the switch from a non-written exam to a written stage does not affect women's application rates differently than men's. With a completely different screening method, women may perceive the job process as being fairer, but may be uncertain about potentially allocating more time to prepare for the test. In line with a cost of application explanation, column (3) shows that there is also no differential response by gender when employers introduced an additional screening requirement. These results further align with the fact that the reform did not have diversity goals or was promoted as "favoring women", so that applicant responses depends on the specific change to screening methods, with women responding to changes in screening that increase perceived fairness, as long as the hiring process is not seen as more costly. In summary, these results suggests that the way employers screen has important sorting effects and diversity implications.

5.5 Additional Results

In Tables A.12 and A.13, I verify whether estimated effects from different treatments may instead be picking up specific characteristics across occupations. Even though I showed before that each treatment contains a variety job titles and that collectively a number of variables explain a negligible portion of treatment assignment, it could still be possible that a disproportionate presence of say, high skill jobs in one treatment could impact my previous results. The additional set of estimates in both tables, where various occupation characteristics are included as controls show that estimated effects remain unchanged.

While the previous tables show that nothing specific about occupations accounts for the estimated treatment effects, next I investigate potential heterogeneity in treatment effects within each treatment. This is also useful to validate the overidentification test in Section 5.3, as heterogenous effects can confound inference about selection (Angrist et al. (2017)). Table A.14 splits each treatment type into low and high-skill positions and again estimates treatment effects on the gender hiring gap. Overall, results are similar independent of skill level, except in the treatment $w + nw \longrightarrow w(b)$, where for low-skill jobs the estimated effect on the gap is now marginally significant and positive. This is the change in screening where on average the bias-information trade-off was dominated by the loss of screening precision from removing nw. The fact that this result flips for low-skill occupations is consistent with screening precision of interviews being lower for low-skill jobs.

6 Conclusion

Hiring decisions shape firm outcomes and determine individuals' access to labor market opportunities. To ensure fair recruiting processes and increase employee gender diversity, organizations face several challenging questions. For example, does replacing interviews with objective tools lead to disparate impact for female applicants? Should screening practices with high discretion be removed, even if they could provide important information about productivity? Causally linking the design of hiring practices to gender equity requires overcoming the lack of data on the decision making process and generating appropriate variation in the choice and implementation of screening methods.

This paper overcomes these challenges and opens the black-box of hiring decisions by developing a natural language processing algorithm that distills high-dimensional, unstructured text records into a uniquely-detailed dataset on the universe of hiring processes in Brazil's public sector. The data contain complete information on candidate performance, including job offers and individual scores, screening methods, and committee hiring members' evaluations of each job applicant in all hiring stages. Combined with the introduction of a requirement that led to a reduction in discretion through different changes in screening methods, my analysis provides important lessons for both private sector firms and professionalized bureaucracies.

Overall, employers can address gender gaps by targeting screening tools used in hiring, without the need to directly change decision makers' fundamental preferences or behavior. However, not all approaches to reduce discretion have the same implications. Screening changes that limit discretion in existing hiring practices or add new impartial screening tools reduce the gender hiring gap by about one third. They also either increase female application rates relative to men, or leave them unchanged when the overall length and cost of participation in a selection process are expanded. But policies that eliminate subjective tools result in a screening precision loss that can outweigh gains from bias removal, having no effects on gender gaps.

Remaining questions to be addressed in future research include the implications of different screening practices for the quality of female and male job seekers. With on-the-job productivity measures that are relevant for the provision of public services, additional work will be able to investigate the potential existence of a equity-efficiency trade-off and its consequences for productivity in the public sector.

References

Abraham, Lisa, Alison Stein, and J Hallermaier (2020) "Words matter: Experimental evidence from job applications," *Working Paper*.

Agan, Amanda and Sonja Starr (2018) "Ban the box, criminal records, and racial discrimination: A field experiment," *Quarterly Journal of Economics*, Vol. 133, No. 1, pp. 191–235.

Aigner, Dennis J and Glen G Cain (1977) "Statistical theories of discrimination in labor markets," *ILR Review*, Vol. 30, No. 2, pp. 175–187.

Alsan, Marcella, Owen Garrick, and Grant Graziani (2019) "Does diversity matter for health? Experimental evidence from Oakland," *American Economic Review*, Vol. 109, No. 12, pp. 4071–4111.

Aneja, Abhay and Guo Xu (2022) "Strengthening state capacity: Postal reform and innovation during the Gilded Age,"Technical report, National Bureau of Economic Research.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters (2017) "Leveraging Lotteries for School Value-Added: Testing and Estimation," *The Quarterly Journal of Economics*, Vol. 132, No. 2, pp. 871–919.

Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott S Lee (2020) "Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services," *American Economic Review*, Vol. 110, No. 5, pp. 1355–94.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum (2020) "The Evolution of Work in the United States," *American Economic Journal: Applied Economics*, Vol. 12, No. 2, pp. 1–34.

Autor, David H and David Scarborough (2008) "Does job testing harm minority workers? Evidence from retail establishments," *Quarterly Journal of Economics*, Vol. 123, No. 1, pp. 219–277.

Baker, Scott R, Nicholas Bloom, and Steven J Davis (2016) "Measuring economic policy uncertainty," *Quarterly Journal of Economics*, Vol. 131, No. 4, pp. 1593–1636.

Baldiga, Katherine (2014) "Gender differences in willingness to guess," *Management Science*, Vol. 60, No. 2, pp. 434–448.

Bartik, Alexander and Scott Nelson (2022) "Deleting a signal: Evidence from pre-employment credit checks," *Review of Economics and Statistics, Forthcoming*.

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon (2015) "Unintended effects of anonymous resumes," *American Economic Journal: Applied Economics*, Vol. 7, No. 3, pp. 1–27.

Bertrand, Marianne and Sendhil Mullainathan (2004) "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, Vol. 94, No. 4, pp. 991–1013.

Besley, Timothy, Robin Burgess, Adnan Khan, and Guo Xu (2022) "Bureaucracy and development," *Annual Review of Economics*, Vol. 14, pp. 397–424.

Bosquet, Clément, Pierre-Philippe Combes, and Cecilia García-Peñalosa (2019) "Gender and promotions: evidence from academic economists in France," *Scandinavian Journal of Economics*, Vol. 121, No. 3, pp. 1020–1053.

Brands, Raina A and Isabel Fernandez-Mateo (2017) "Leaning out: How negative recruitment experiences shape womens decisions to compete for executive roles," *Administrative Science Quarterly*, Vol. 62, No. 3, pp. 405–442.

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu (2020) "The structure of economic news," *NBER Working Paper*.

Card, David, Fabrizio Colella, and Rafael Lalive (2021) "Gender preferences in job vacancies and workplace gender diversity,"Technical report, National Bureau of Economic Research.

Coffman, Katherine B, Manuela Collis, and Leena Kulkarni (2023) "Whether to Apply?" *Management Science, Forthcoming*.

Cornell, Bradford and Ivo Welch (1996) "Culture, information, and screening discrimination," *Journal of Political Economy*, Vol. 104, No. 3, pp. 542–571.

Dal Bó, Ernesto, Frederico Finan, and Martín A Rossi (2013) "Strengthening state capabilities: The role of financial incentives in the call to public service," *Quarterly Journal of Economics*, Vol. 128, No. 3, pp. 1169–1218.

De Janvry, Alain, Guojun He, Elisabeth Sadoulet, Shaoda Wang, and Qiong Zhang (2023) "Subjective performance evaluation, influence activities, and bureaucratic work behavior: Evidence from China," *American Economic Review*, Vol. 113, No. 3, pp. 766–799.

Delfino, Alexia (2021) "Breaking gender barriers: Experimental evidence on men in pink-collar jobs."

Deserranno, Erika (2019) “Financial incentives as signals: experimental evidence from the recruitment of village promoters in Uganda,” *American Economic Journal: Applied Economics*, Vol. 11, No. 1, pp. 277–317.

Doleac, Jennifer L and Benjamin Hansen (2020) “The unintended consequences of ban the box: Statistical discrimination and employment outcomes when criminal histories are hidden,” *Journal of Labor Economics*, Vol. 38, No. 2, pp. 321–374.

Durlauf, Steven N. and James J. Heckman (2020) “An Empirical Analysis of Racial Differences in Police Use of Force: A Comment,” *Journal of Political Economy*, Vol. 128, No. 10, pp. 3998–4002.

Estrada, Ricardo (2019) “Rules versus discretion in public service: Teacher hiring in Mexico,” *Journal of Labor Economics*, Vol. 37, No. 2, pp. 545–579.

Finan, Frederico, Benjamin A Olken, and Rohini Pande (2017) “The personnel economics of the developing state,” *Handbook of Economic Field Experiments*, Vol. 2, pp. 467–514.

Flory, Jeffrey A, Andreas Leibbrandt, and John A List (2015) “Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions,” *The Review of Economic Studies*, Vol. 82, No. 1, pp. 122–155.

Flory, Jeffrey A, Andreas Leibbrandt, Christina Rott, and Olga Stoddard (2021) “Increasing workplace diversity: Evidence from a recruiting experiment at a Fortune 500 company,” *Journal of Human Resources*, Vol. 56, No. 1, pp. 73–92.

Frankel, Alex (2021) “Selecting Applicants,” *Econometrica*, Vol. 89, No. 2, pp. 615–645.

Frederiksen, Anders, Lisa B Kahn, and Fabian Lange (2020) “Supervisors and performance management systems,” *Journal of Political Economy*, Vol. 128, No. 6, pp. 2123–2187.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019) “Text as Data,” *Journal of Economic Literature*, Vol. 57, No. 3, pp. 535–74.

Gentzkow, Matthew and Jesse M Shapiro (2010) “What drives media slant? Evidence from US daily newspapers,” *Econometrica*, Vol. 78, No. 1, pp. 35–71.

Ghanem, Dalia, Pedro HC Sant’Anna, and Kaspar Wüthrich (2022) “Selection and parallel trends,” *arXiv preprint arXiv:2203.09001*.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini (2003) “Performance in competitive environments: Gender differences,” *Quarterly Journal of Economics*, Vol. 118, No. 3, pp. 1049–1074.

Goldin, Claudia and Cecilia Rouse (2000) "Orchestrating impartiality: The impact of "blind" auditions on female musicians," *American Economic Review*, Vol. 90, No. 4, pp. 715–741.

Grindle, Merilee S (2012) *Jobs for the Boys*: Harvard University Press.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li (2018) "Discretion in hiring," *Quarterly Journal of Economics*, Vol. 133, No. 2, pp. 765–800.

Holzer, Harry J, Steven Raphael, and Michael A Stoll (2006) "Perceived criminality, criminal background checks, and the racial hiring practices of employers," *Journal of Law and Economics*, Vol. 49, No. 2, pp. 451–480.

Hospido, Laura, Luc Laeven, and Ana Lamo (2019) "The gender promotion gap: evidence from central banking," *Review of Economics and Statistics*, pp. 1–45.

Kingsley, J Donald (1944) "Representative bureaucracy: An interpretation of the British civil service," *(No Title)*.

Kline, Patrick, Evan K Rose, and Christopher R Walters (2022) "Systemic discrimination among large US employers," *Quarterly Journal of Economics*, Vol. 137, No. 4, pp. 1963–2036.

Kuhn, Peter and Kailing Shen (2023) "What Happens When Employers Can No Longer Discriminate in Job Ads?" *American Economic Review*, Vol. 113, No. 4, pp. 1013–1048.

Kuhn, Peter, Kailing Shen, and Shuo Zhang (2020) "Gender-targeted job ads in the recruitment process: Facts from a Chinese job board," *Journal of Development Economics*, Vol. 147, p. 102531.

Lundberg, Shelly J and Richard Startz (1983) "Private discrimination and social intervention in competitive labor market," *American Economic Review*, Vol. 73, No. 3, pp. 340–347.

MacLeod, W Bentley (2003) "Optimal contracting with subjective evaluation," *American Economic Review*, Vol. 93, No. 1, pp. 216–240.

McCrary, Justin (2007) "The effect of court-ordered hiring quotas on the composition and quality of police," *American Economic Review*, Vol. 97, No. 1, pp. 318–353.

Miller, Amalia R and Carmit Segal (2019) "Do female officers improve law enforcement quality? Effects on crime reporting and domestic violence," *Review of Economic Studies*, Vol. 86, No. 5, pp. 2220–2247.

Moreira, Diana and Santiago Pérez (2021a) "Civil Service Reform and Organizational Practices: Evidence from the Pendleton Act," *NBER Working Paper*.

——— (2021b) "Who Benefits from Meritocracy?" *Working Paper*.

Neumark, David (2021) "Age discrimination in hiring: Evidence from age-blind vs. non-age-blind hiring procedures," *Journal of Human Resources*, pp. 0420–10831R1.

Niederle, Muriel and Lise Vesterlund (2007) "Do women shy away from competition? Do men compete too much?" *Quarterly Journal of Economics*, Vol. 122, No. 3, pp. 1067–1101.

Ornaghi, Arianna (2019) "Civil service reforms: Evidence from US police departments," *Working Paper*.

Oyer, P and S Schaefer (2011) "Personnel Economics: Hiring and Incentives. Volume 4, Part B, Chapter 20 of Handbook of Labor Economics."

Phelps, Edmund S (1972) "The statistical theory of racism and sexism," *American Economic Review*, Vol. 62, No. 4, pp. 659–661.

Prendergast, Canice and Robert Topel (1993) "Discretion and bias in performance evaluation," *European Economic Review*, Vol. 37, No. 2-3, pp. 355–365.

Prendergast, Canice and Robert H Topel (1996) "Favoritism in organizations," *Journal of Political Economy*, Vol. 104, No. 5, pp. 958–978.

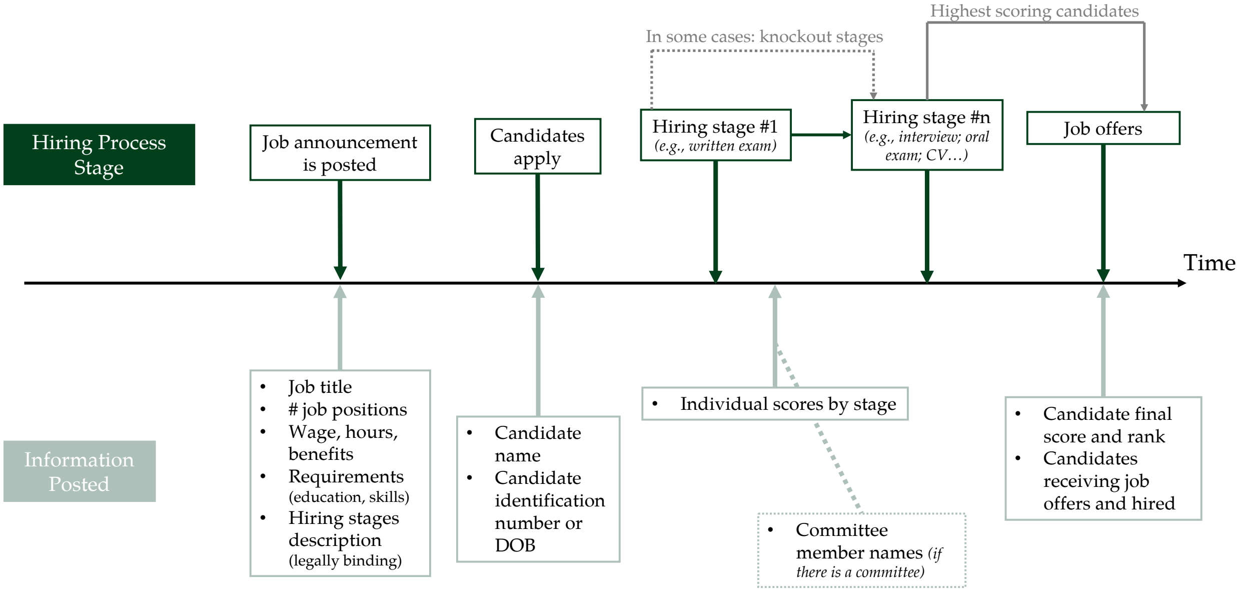
Sargan, J. D. (1958) "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, Vol. 26, No. 3, pp. 393–415.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li (2021) "LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis," *arXiv preprint arXiv:2103.15348*.

Small, Mario L and Devah Pager (2020) "Sociological perspectives on racial discrimination," *Journal of Economic Perspectives*, Vol. 34, No. 2, pp. 49–67.

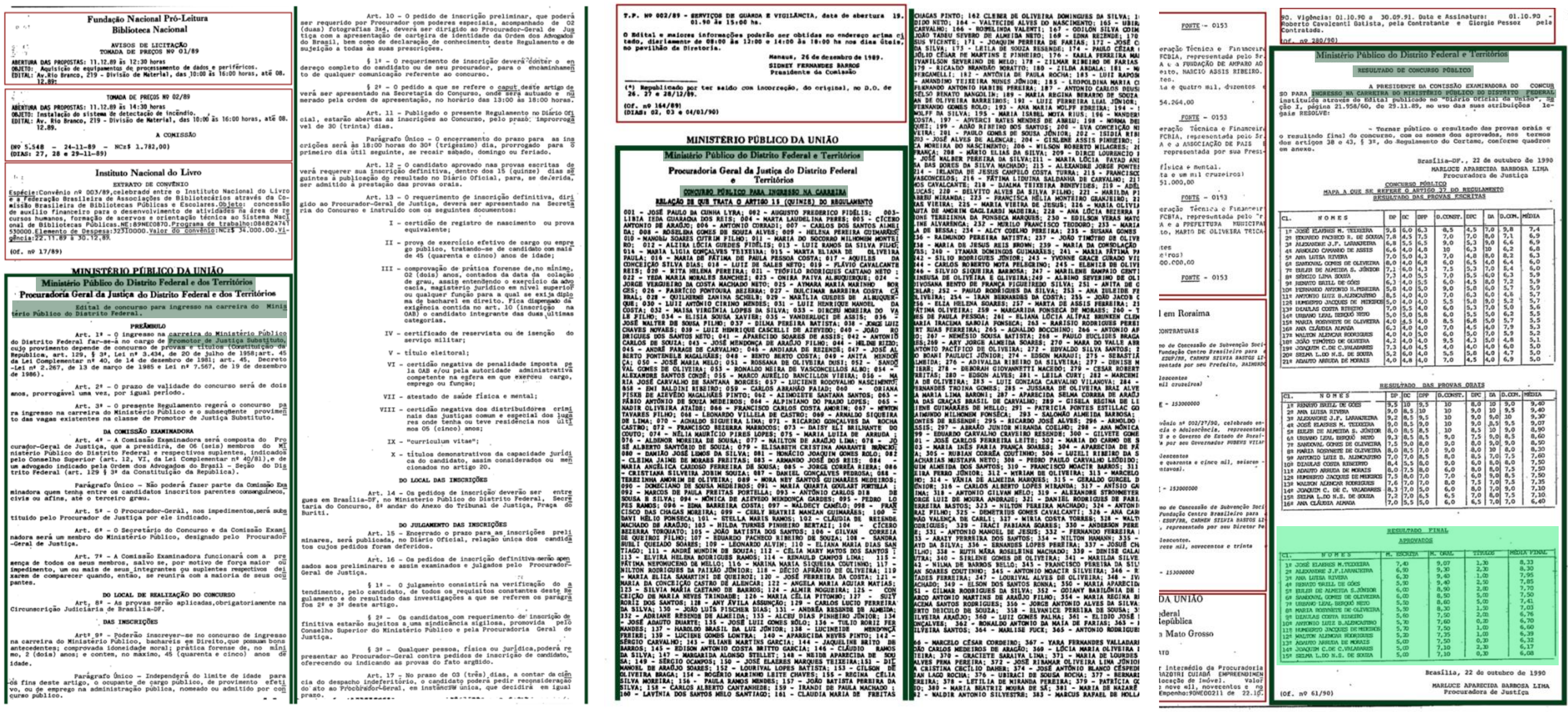
Xu, Guo (2021) "Bureaucratic Representation and State Responsiveness during Times of Crisis: The 1918 Pandemic in India," *Review of Economics and Statistics*, pp. 1–29.

Figures and Tables


Highest scoring candidates
In some cases: knockout stages
Hiring Process Stage
Job announcement is posted
Candidates apply
Hiring stage #1 (e.g., written exam)
Hiring stage #n (e.g., interview; oral exam; CV…)
Job offers
Time
Information Posted
Job title
# job positions
Wage, hours, benefits
Requirements (education, skills)
Hiring stages description (legally binding)
Candidate name
Candidate identification number or DOB
Individual scores by stage
Committee member names (if there is a committee)
Candidate final score and rank
Candidates receiving job offers and hired

Notes: This shows the stylized structure of a hiring process in the Brazilian public sector posted in raw government publications (official gazettes). Information at the top (dark green) describes the screening dynamics from the moment a job is announced until job offers are sent out. The lower part of the figure (light green) shows variables I construct based on observable information in the text of official government documents. The procedure for data extraction is described in Section 3.

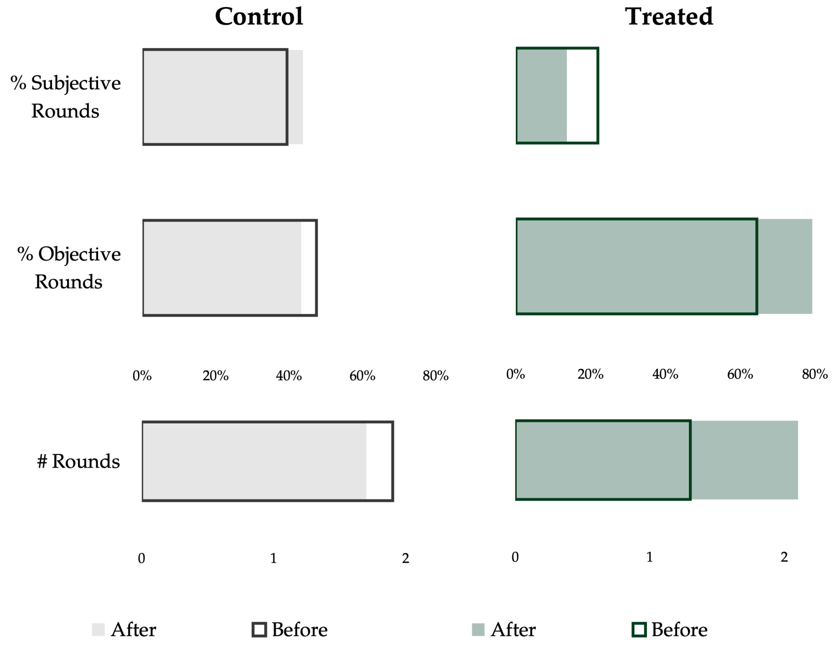
FIGURE 1: Stylized Structure of Public Sector Hiring Processes



(a) Announcement **(b)** Applicants **(c)** Final Results

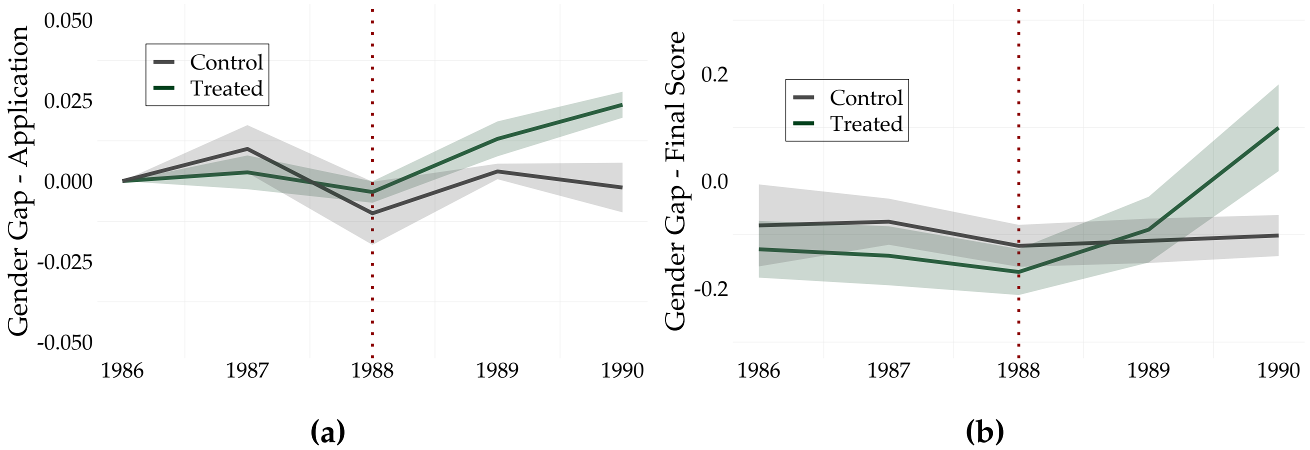
Notes: This figure shows screen captures of the raw government gazette pages containing the **(a)** announcement, **(b)** list of applicants, and **(c)** final results for a job process in the federal government to select public prosecutors. Pictures are trimmed and additional posts containing exam scores before the final results are not shown for visualization purposes. Text snippets bounded in red are extraneous posts unrelated to the target hiring process, while text bounded in green represents the correct contents. Shaded areas represent target semantic relationships within text snippets, corresponding to the attributes sender, receiver, feature, and numeric elements.

FIGURE 2: Illustration of Linkage and Data Extraction of Job Processes



Notes: This figure shows the average share of job processes with subjective rounds, objective rounds, and the number of screening rounds used by federal (treated) and state (control) governments. Black boxed boundaries represent pre-treatment averages, while shaded areas plot the same statistic after treatment. All statistics control for occupation fixed effects.

FIGURE 3: Compliance with the Impartiality Requirement



Notes: The figure on the left plots residualized application gaps estimated from the regression

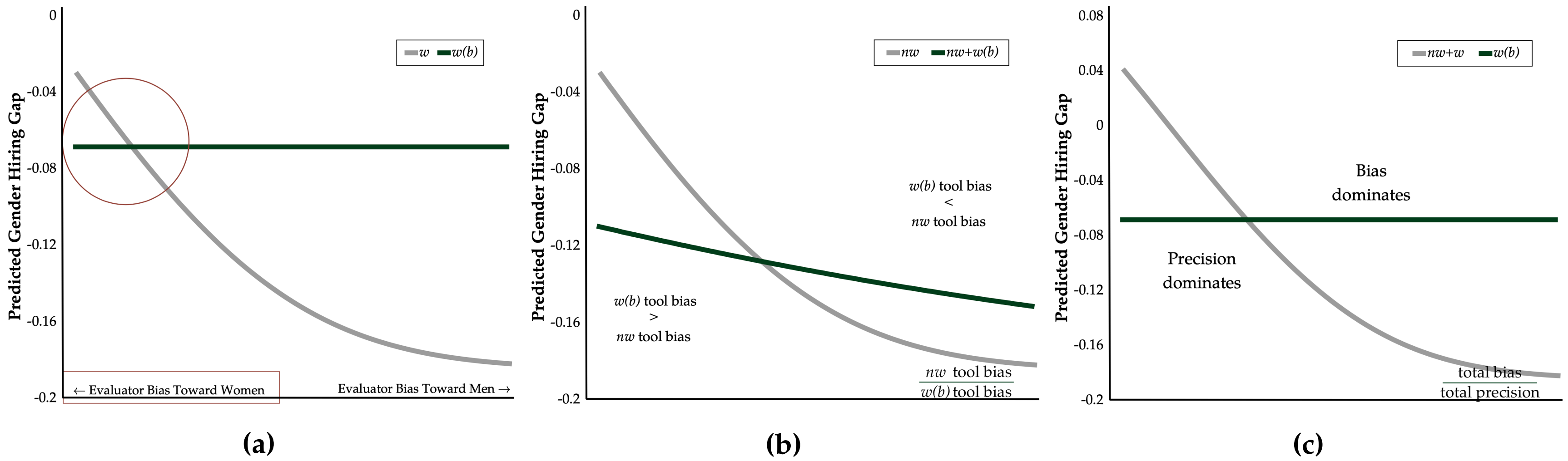
$$\mathbb{P}\left(\text{Female} = 1\right) = \delta_{j(o)} + \theta_t + u_{ijt}$$

separately for control (state governments) and treated (federal government) groups. i denotes a job applicant, j denotes a job process, and o denotes the occupation or job title. 1986 is the omitted year. The figure on the right plots $\widehat{\gamma_0}$ estimates of the regression

$$\text{Final Score}_{ijt} = \delta_{j(o)} + \gamma_0 \text{Female}_i + u_{ijt}$$

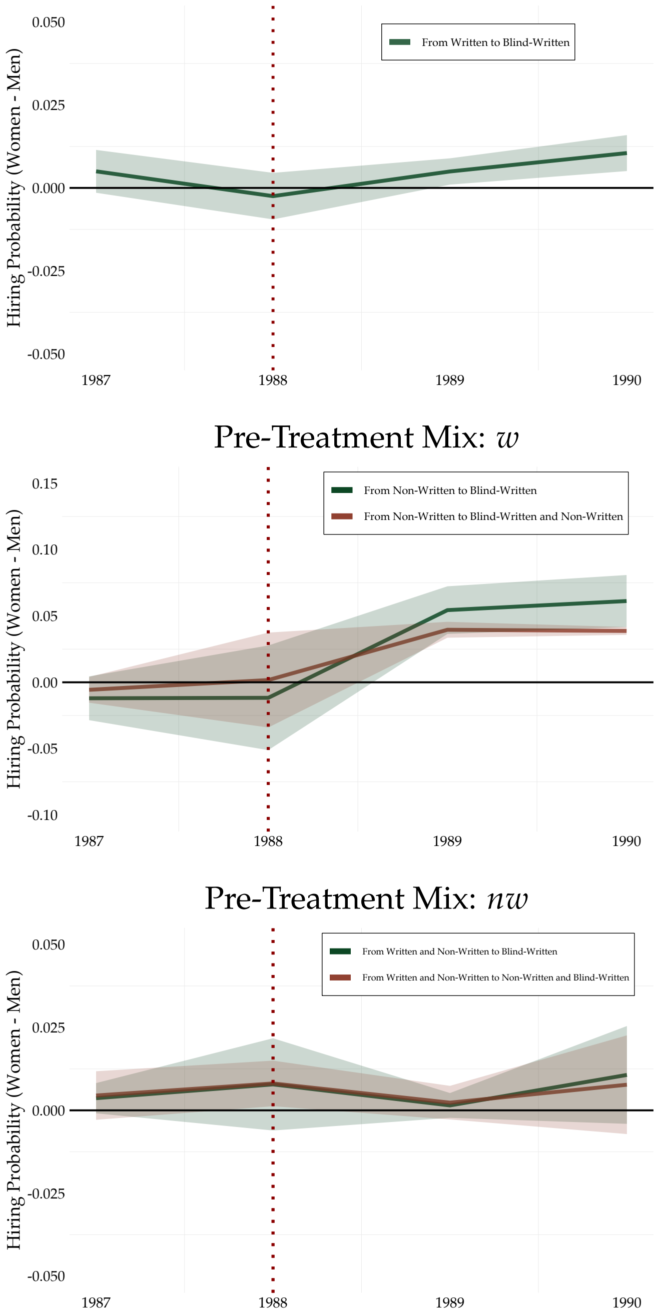
for control (state governments) and treated (federal government) groups separately and for each year. Standard errors are clustered at the job process level across all specifications. Shaded areas are 95% confidence intervals.

FIGURE 4: Dynamic Estimates of Gender Application and Score Gaps



Notes: The figure illustrates predictions for the gender hiring gap using the closed-form solutions in Appendix C. Panel **(a)** shows the model-implied gender hiring gap when screening only uses written w (gray) or alternatively only uses nonwritten $w(b)$ (green) tools. The gender gap is plotted as a function of evaluator bias toward a gender. Unless evaluators are biased toward women, the complete removal of discretion reduces the gender gap since it completely eliminates disparate treatment. Panel **(b)** compares gender gaps when screening only contains a subjective step to when a job process adds a blind written stage. This plot varies on the horizontal axis the relative tool bias of nw with respect to $w(b)$. Unless the disparate impact of blind written tests is much larger than that of interviews, the increase in screening precision in $w(b) + nw$ leads to smaller gender hiring gaps than nw. Panel **(c)** shows the case when nw is removed, showing that gender hiring gaps narrow only if the reduction in total bias from removing nw offsets the loss in total screening precision.

FIGURE 5: Predicted Effects on the Gender Hiring Gap from Different Changes in Screening



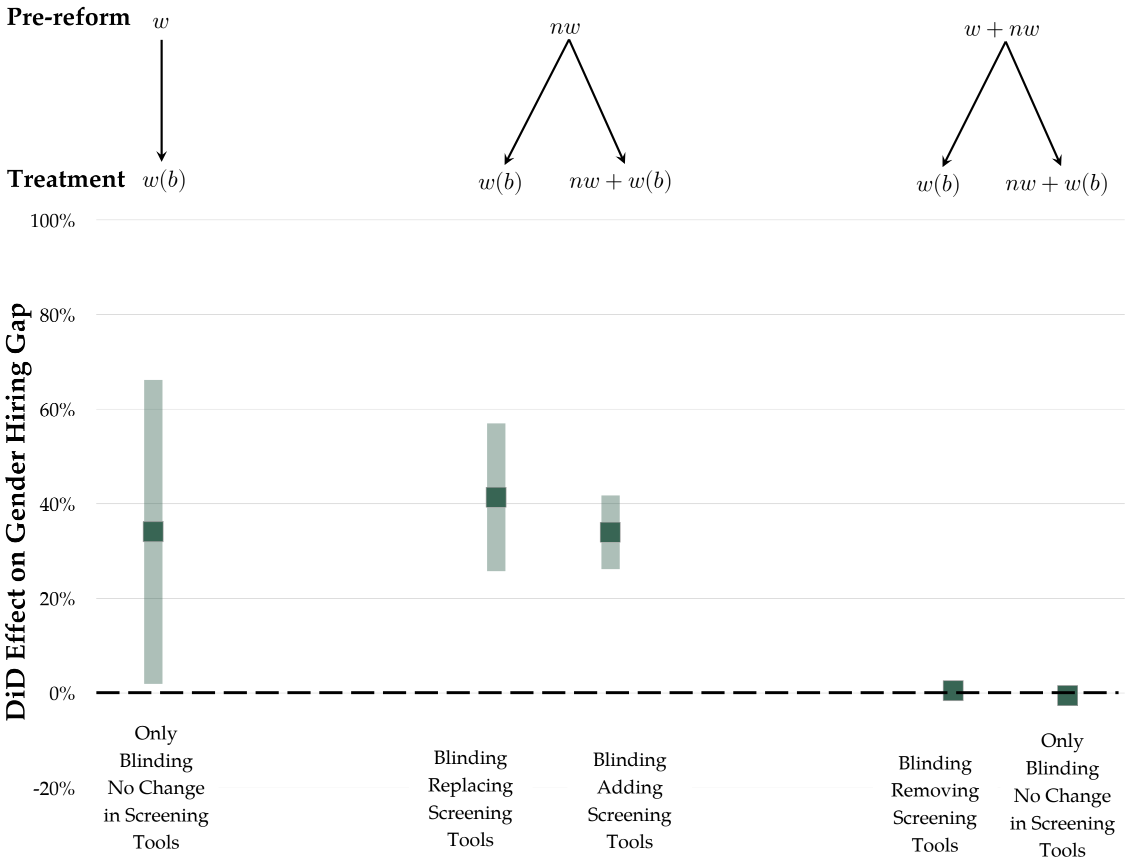
Pre-Treatment Mix: w

Pre-Treatment Mix: nw

Pre-Treatment Mix: $w + nw$

Notes: This figure compares gender hiring gaps in occupations that had the same pre-treatment screening methods and that followed different treatments. The first panel shows job processes transitioning from $w \longrightarrow w(b)$. The second panel shows gender gaps for occupations that followed $nw \longrightarrow w(b)$ or $nw \longrightarrow w(b) + nw$. The third panel compares outcome paths for the case $w + nw \longrightarrow w(b) + nw$ to $w + nw \longleftrightarrow w(b)$. Shaded areas are calculated with standard errors clustered at the job process level.

FIGURE 6: Gender Hiring Gap Paths in Each Treatment Type



Notes: This figure plots treatment effects estimated from the triple-difference model:

$$y_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j \right)$$
$$+ \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i \right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i \right)$$
$$+ \beta_d \left(\text{Federal}_j \times \text{Post}_{t(j)} \times \text{Female}_i \right) + \delta_{o(j)} + \theta_t + u_{ijt}$$

which compares outcomes (y_{ijt}) for female candidates relative to men (Female_i) participating in job processes for the same occupation ($\delta_{o(j)}$) that had screening practices changed following treatment d. Treatments are displayed at the top with their corresponding estimates in the bars displayed below. Bars are 95% confidence intervals obtained with standard errors clustered at the job process level.

FIGURE 7: Treatment Effects of Changes in Screening Tools: Gender Hiring Gap

TABLE 1: Selected Characteristics of Job Processes

	Control		Treated	
	Before Reform	After Reform	Before Reform	After Reform
Education Requirement				
<High School	47%	22%	13%	7%
High School	26%	17%	27%	8%
College or more	27%	61%	60%	85%
Screening Steps				
Average # Rounds	1.9	1.7	1.3	2.1
% Rounds Objective	47.4%	43.3%	64.4%	79.2%
% Rounds Subjective	39.4%	43.8%	21.9%	13.6%
% Rounds Resume	13.2%	12.9%	13.7%	7.2%
Job Applicants				
Avg # Applicants	34.7	33.6	34.4	18.3
Total # Applicants	34,871	41,736	18,726	12,600
# Job Processes	2,422	838	1,005	2,289

Notes: These are descriptive statistics of job processes published during 1986-1990. Treated job processes are those in Brazil's federal sector. The control group comprises processes in the states of Amazonas in the country's north region, Pernambuco in the northeast, Distrito Federal, Mato Grosso, and Mato Grosso do Sul in the central region, São Paulo — the largest and richest state — in the southeast, and Rio Grande do Sul in the south. Sample statistics for *Screening Steps* are net of occupation fixed effects to compare temporal changes within different hiring processes for the same job title.

TABLE 2: Estimates of Screening Impartiality on Candidate Scores

	Final Score			Written Score			Non-Written Score		
	Women	Men		Women	Men		Women	Men	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\text{Federal}_j \times \text{Post}_{j(t)}$	0.067**	−0.075*		0.024	−0.109*		−0.010	0.020	
	(0.030)	(0.037)		(0.044)	(0.059)		(0.071)	(0.099)	
$\text{Federal}_j \times \text{Post}_{j(t)}$			0.141***			0.134*			−0.031
$\times\ \text{Female}_i$			(0.048)			(0.074)			(0.122)
Pre-treatment mean	−0.04	0.07	−0.12	−0.02	0.03	−0.05	−0.02	0.13	−0.15
Occupation FE	X	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X	X
Obs.	54,892	32,067	86,959	34,511	15,546	50,057	29,444	10,764	40,208

Notes: The table shows difference-in-differences estimates of outcomes for applicant i in job process j in a given year t. The first specification is $y_{ijt} = \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_{j(o)} + \theta_t + u_{ijt}$ only with female candidates in columns (1), (4), and (7), only with male candidates in columns (2), (5), and (8). The second specification for columns (3), (6), and (9) is $y_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right) + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$. The outcome y represents either a candidate's final score, written score (written exams), or non-written score (interview, oral, practical exams). If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE 3: Estimates of Screening Impartiality on Hiring and Application Rates

	$\mathbb{P}(Hired\vert\text{Female} = 1)$	$\mathbb{P}(Hired\vert\text{Male} = 1)$	$\mathbb{P}(Hired\vert\text{Applied} = 1)$	$\mathbb{P}(\text{Female} = 1)$
	(1)	(2)	(3)	(4)
$\text{Federal}_j \times \text{Post}_{j(t)}$	0.003**	−0.004***		0.010**
	(0.001)	(0.001)		(0.005)
$\text{Federal}_j \times \text{Post}_{j(t)}$			0.007***	
$\times$ Female_i			(0.002)	
Pre-treatment mean	0.09	0.11	−0.015	0.52
Occupation FE	X	X	X	X
Year FE	X	X	X	X
Obs.	54,892	32,067	86,959	86,959

Notes: The first column shows a regression coefficient capturing the probability that a given female job applicant receives a job offer: $\mathbb{P}(Hired = 1)_{ijt} = \delta_2\text{Federal}_j + \delta_3\left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_{j(o)} + \theta_t + u_{ijt}$ which is ran only on female individuals, column (2) runs the same regression with male applicants, column (3) runs $\mathbb{P}(Hired = 1)_{ijt} = \delta_1\text{Female}_i + \delta_2\text{Federal}_j + \delta_3\left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4\left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5\left(\text{Federal}_j \times \text{Female}_i\right) + \beta\left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$ which measures the probability female job applicants receive an offer relative to male job applicants — the gender hiring gap. Finally, column (4) regresses the specification $\mathbb{P}(\text{Female} = 1)_{ijt} = \delta_2\text{Federal}_j + \delta_3\left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_{j(o)} + \theta_t + u_{ijt}$. If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE 4: Treatment Effects of Changes in Screening Tools: Share of Female Applicants

	% Female Applicants				
	$w \to w(b)$	$nw \to w(b)$	$nw \to w(b) + nw$	$w + nw \to w(b)$	$w + nw \to w(b) + nw$
	(1)	(2)	(3)	(4)	(5)
$\text{Federal}_j \times \text{Post}_{j(t)}$	0.023***	−0.004	−0.022	0.054***	0.043***
	(0.008)	(0.013)	(0.020)	(0.009)	(0.012)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Obs.	1,145	900	1,822	4,252	3,106

Notes: This table plots treatment effects for each treatment type d using the regression:

$$\text{\% Female Applicants}_{jt} = \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j \right) + \delta_{o(j)} + \theta_t + u_{ujt}$$

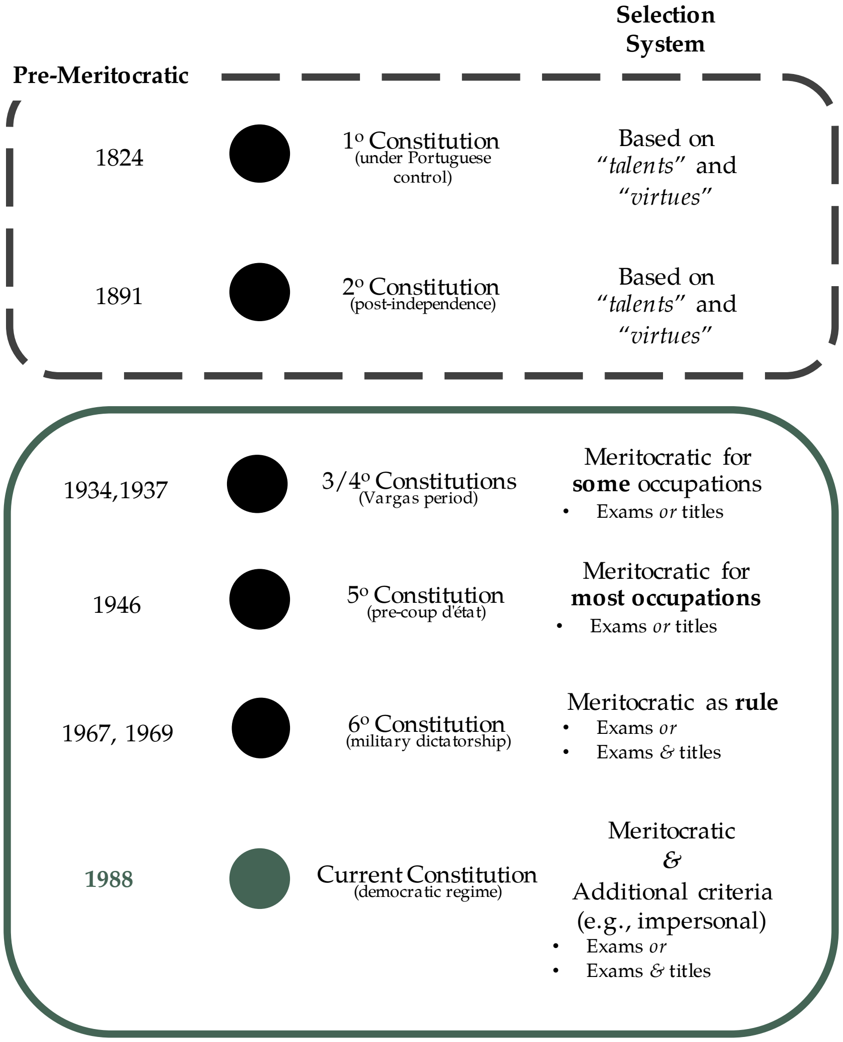
where $\text{Post}_{j(t)}$ is a dummy equal to 1 if a job process occurred $t \geq 1989$ and $\text{Federal}_j = 1$ if it was for a federal job. Treatment type d represents job process transitioning from written exam to blind-written exam ($w \longrightarrow w(b)$), a job process comprising a non-written exam switching for a blind-written ($nw \longrightarrow w(b)$), or only adding the blind-written test ($nw \longrightarrow w(b) + nw$), and a hiring process using a mix of written and non-written tools potentially dropping the non-written and blinding the written ($w + nw \longrightarrow w(b)$) or just blinding the written ($w + nw \longrightarrow w(b) + nw$). All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

Designing Gender Equity: Evidence from Hiring Practices

INTERNET APPENDIX

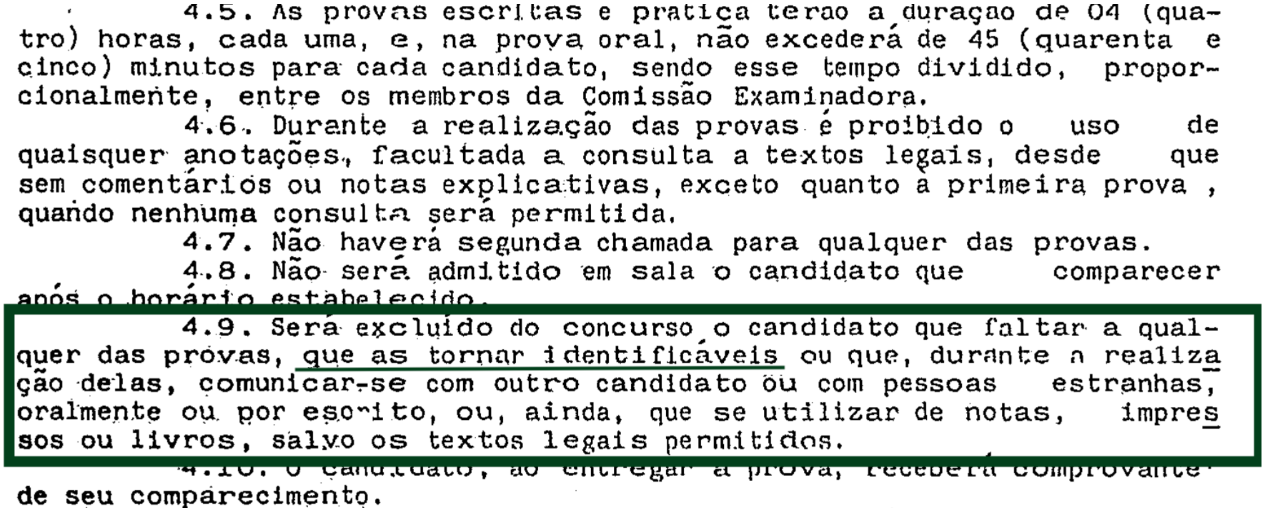
FOR ONLINE PUBLICATION

Appendix A Figures and Tables



Notes: This figure shows the history of all changes to the selection process of public servants in Brazil, beginning from when the country was still under Portuguese domain, and spanning democratic and military control periods. Brazilian legal experts and historians consider the 1934 Constitution (amended in 1939) to establish meritocratic public servant selection — one of the first countries in Latin America. This early stage, however, provisioned the use of examinations or titles (resume) for some occupations. The 1946 Constitution expanded the selection criteria for most government jobs, until in 1967 and 1969 under military regime, the selection of every public servant through the legal device known as *Concurso* had to include at least one type of examination, ruling out the sole use of resumes. Despite the language, the definition of examination at that moment was fairly broad, so that interviews would be character or personality "exams", for example. In the end of 1988, Brazil passed a new Constitution which kept all public servant selection criteria from the previous Constitution but required public sector job processes to be conducted impartially. I exploit the introduction of this requirement as the main source of variation for part of the empirical analysis in the paper.

FIGURE A.1: History of Changes in the Selection of Public Servants in Brazil



4.5. As provas escritas e pratica terao a duraçao de 04 (quatro) horas, cada uma, e, na prova oral, não excederá de 45 (quarenta e cinco) minutos para cada candidato, sendo esse tempo dividido, proporcionalmente, entre os membros da Comissão Examinadora.

4.6. Durante a realização das provas é proibido o uso de quaisquer anotações, facultada a consulta a textos legais, desde que sem comentários ou notas explicativas, exceto quanto à primeira prova, quando nenhuma consulta será permitida.

4.7. Não haverá segunda chamada para qualquer das provas.

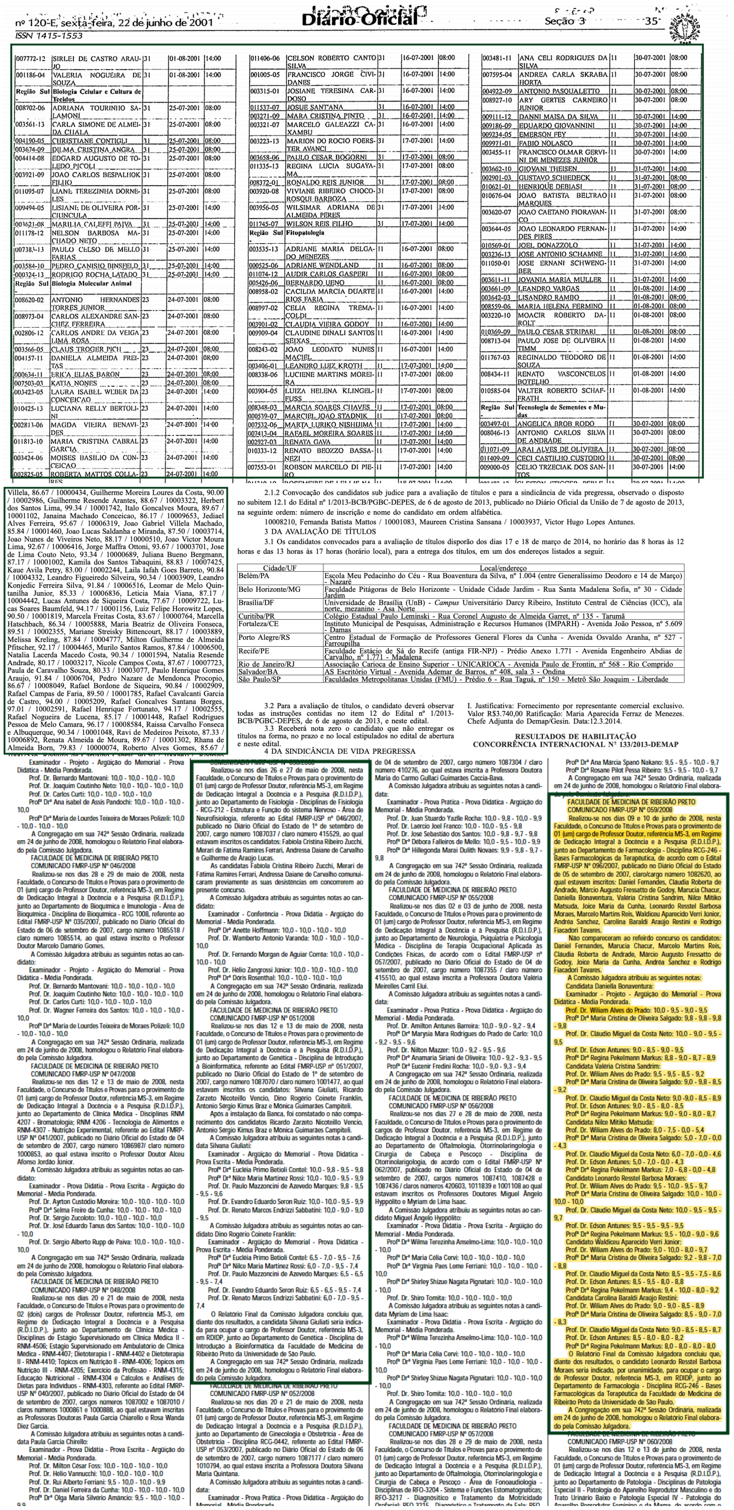
4.8. Não será admitido em sala o candidato que comparecer após o horário estabelecido.

4.9. Será excluído do concurso o candidato que faltar a qualquer das provas, que as tornar identificáveis ou que, durante a realização delas, comunicar-se com outro candidato ou com pessoas estranhas, oralmente ou por escrito, ou, ainda, que se utilizar de notas, impressos ou livros, salvo os textos legais permitidos.

4.10. O candidato, ao entregar a prova, receberá comprovante de seu comparecimento.

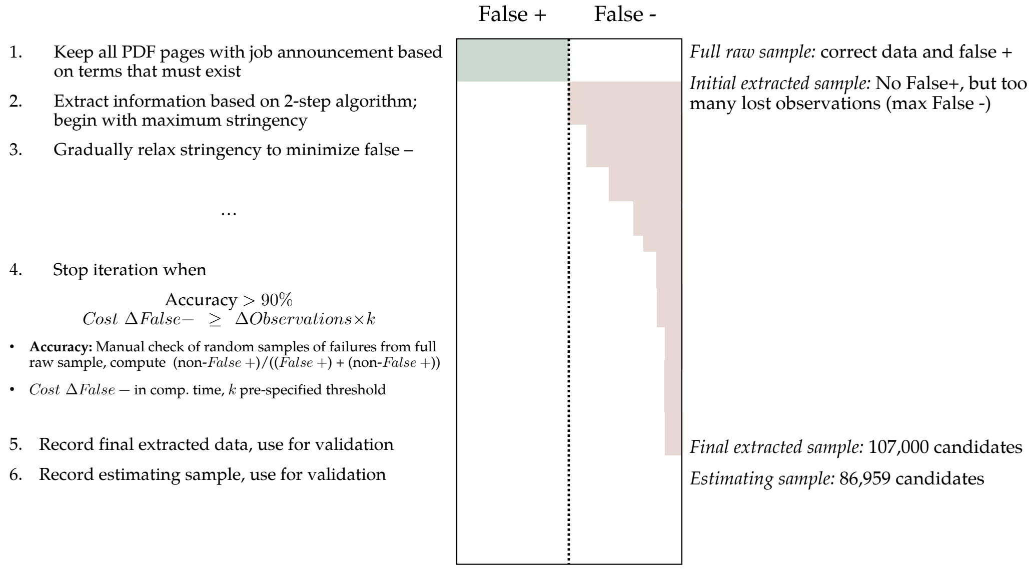
Notes: Selection Process Rules for Hiring Federal Judges (Sep 4, 1989). Reads as: *"Candidates identifying themselves in any exam will be excluded from the hiring process."*

FIGURE A.2: Enforcing Blind Exams After Reform



Notes: This is a sample of some layouts of job selection process postings in Brazilian government gazettes (similar to the Federal Register). Information on candidate scores, hiring committee members, exam types, selection process rules, results and other variables described in Figure 1 are contained in over 200 different layout types and can be found in the example inside green boxes. The underlying full data size spans over 35 million text documents published over almost 50 years.

FIGURE A.3: Examples of Raw Text Containing Hiring Information



1. Keep all PDF pages with job announcement based on terms that must exist
2. Extract information based on 2-step algorithm; begin with maximum stringency
3. Gradually relax stringency to minimize false –

…

4. Stop iteration when

$$\text{Accuracy} > 90\%$$
$$Cost\ \Delta False- \ \geq \ \Delta Observations \times k$$

- **Accuracy:** Manual check of random samples of failures from full raw sample, compute (non-$False+$)/(($False+$) + (non-$False+$))
- $Cost\ \Delta False-$ in comp. time, k pre-specified threshold

5. Record final extracted data, use for validation
6. Record estimating sample, use for validation

Full raw sample: correct data and false +

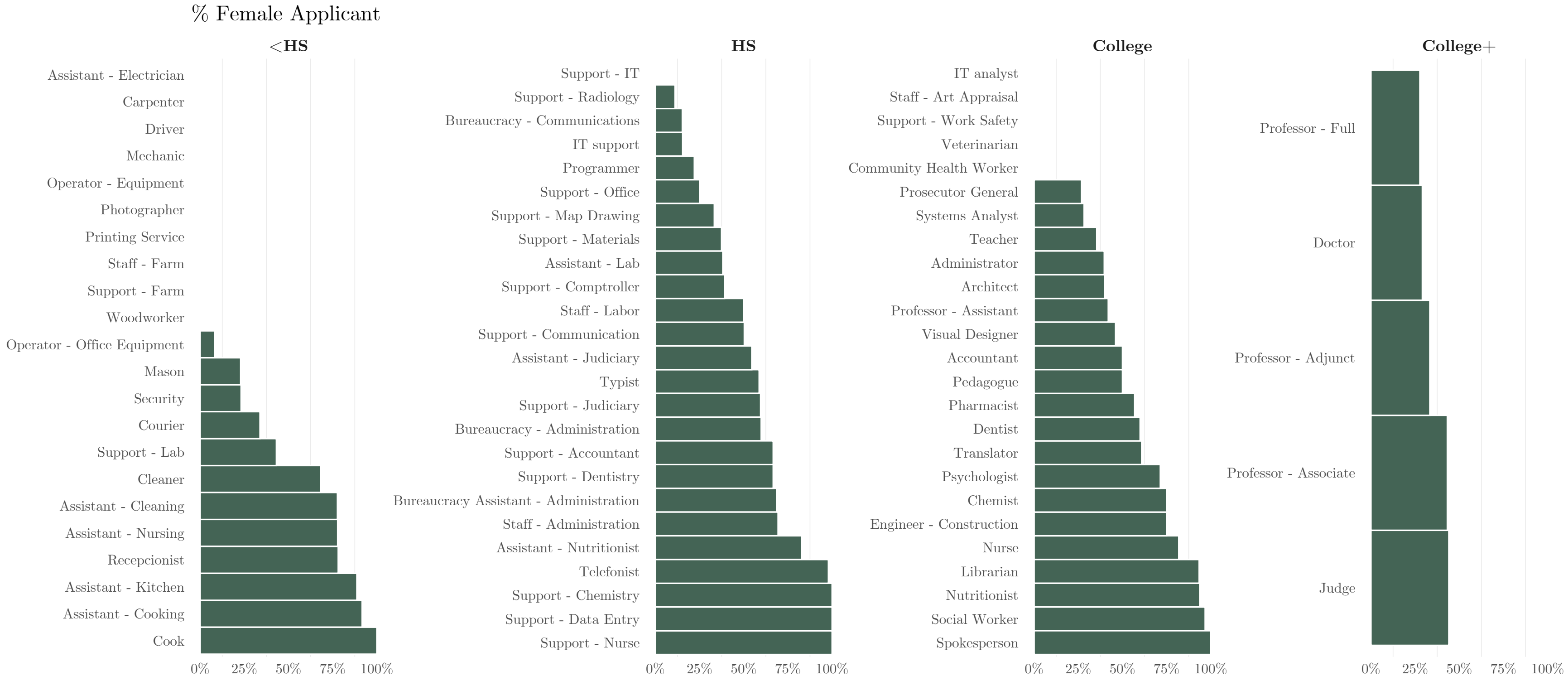
Initial extracted sample: No False+, but too many lost observations (max False -)

Final extracted sample: 107,000 candidates

Estimating sample: 86,959 candidates

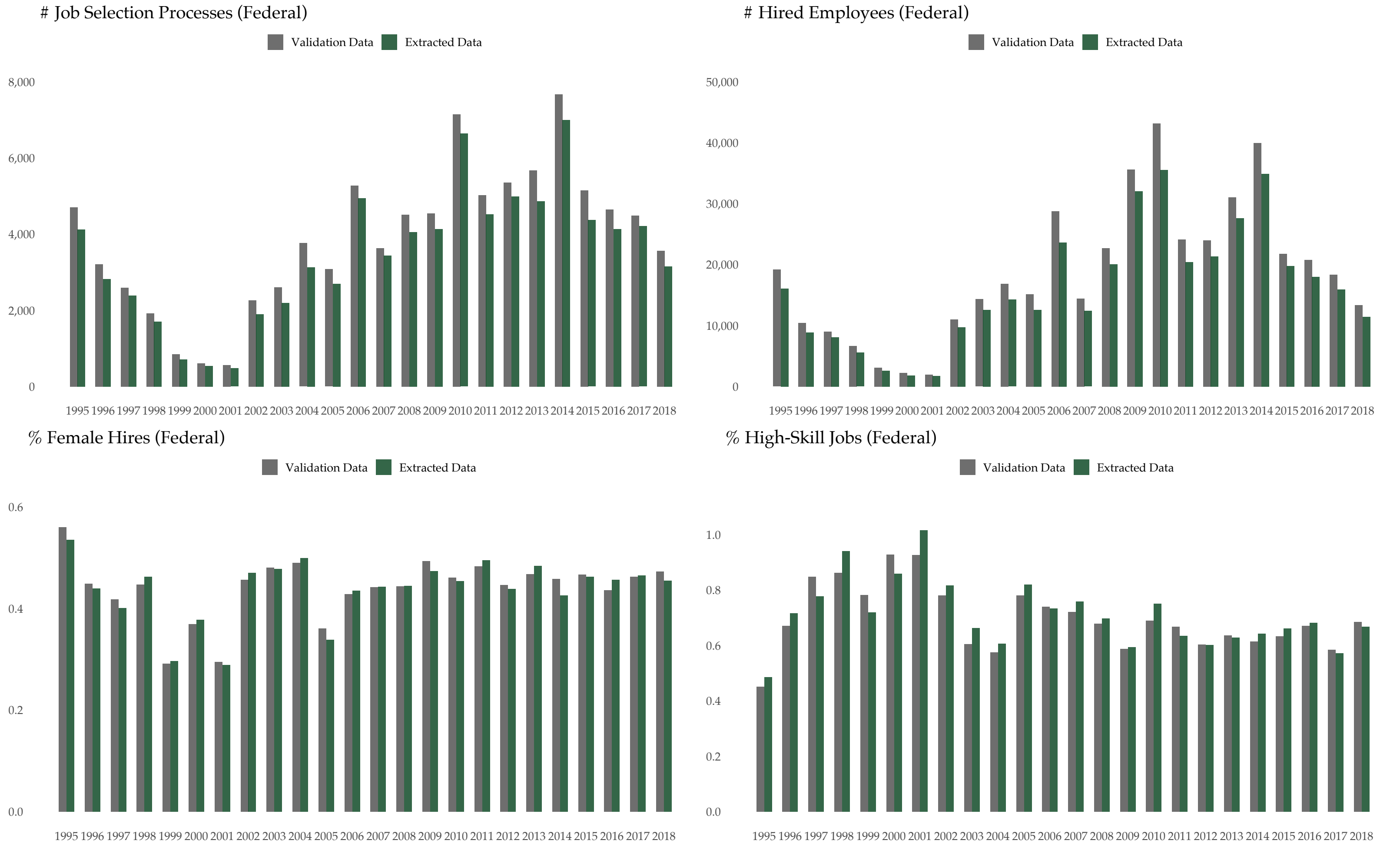
Notes: This figure shows the implementation pipeline of the second step of the natural language processing algorithm developed in the paper to transform unstructured text into ready-to-use data. Each implementation step (on the left) is associated with a level of generated false positives and negatives (center) and underlying sample size (right).

FIGURE A.4: Algorithm Pipeline



Notes: This figure shows the gender share distribution of job applicants to various occupations and skill levels in Brazil's public sector from 1986 to 1990. Occupation titles in the data follow employer-specific career titles given each organizational structure. These titles are translated from Portuguese and then manually assigned occupation categories based on job or title description so that homogeneous occupation groups can be created. The occupation level displayed in the figure is intermediary — equivalent the Census Bureau Standard Occupational Classification (SOC) 4-digit code in most cases and slightly more granular in others. Skill levels are directly informed in job announcements, where only candidates attaining that educational level can apply for the job process. In the rare cases where different job titles are bridged by the same occupation name and they have distinct educational requirements, I consider the in the job title most closely reflecting the underlying occupational name or that is required more frequently. Occupations with blank bars had zero female applicants (e.g., carpenter, driver, mechanics) and some occupations had only women applying (e.g., data entry (support), spokesperson).

FIGURE A.5: Distribution of Female Applicants by Occupation and Skill Level



Notes: This figure compares aggregate statistics informed by Brazil's federal government on its public sector to calculated sample moments using data extracted from official government gazettes. For each statistic — annual number of job selection processes, number of hired employees, share of female hires, and share of high-skill jobs posted — the correlation between actual data and the constructed information using the NLP algorithm is above 99%, with error bands never outside 2%.

FIGURE A.6: NLP Algorithm Validation: Federal Jobs Sample Moments

$z = 0$

		Written	**Non-written**	**Written & Non-written**	**Written Blind**	**Written Blind & Non-written**
	Written	Always Written				
	Non-written		Always Non-written			
$z = 1$	**Written & Non-written**			Always Written & Non-written		
	Written Blind	$w \rightarrow w(b)$ **[20%]**	$nw \rightarrow w(b)$ **[6.7%]**	$w + nw \rightarrow w(b)$ **[6.7%]**	Always Written Blind	
	Written Blind & Non-written	$w \rightarrow w(b) + nw$	$nw \rightarrow w(b) + nw$ **[20%]**	$w + nw \rightarrow w(b) + nw$ **[46.7%]**		Always Written Blind & Non-written

Notes: This figure illustrates all possible potential treatments (strata) generated by the introduction of the impartiality requirement in hiring for federal jobs in Brazil. Areas shaded in gray are ruled out by standard difference-in-differences assumptions and 5 out of the 6 allowed treatments are consistent with the variation induced by compliance: a job process transitioning from written exam to blind-written exam ($w \longrightarrow w(b)$), a job process comprising a non-written exam switching to a blind-written ($nw \longrightarrow w(b)$), or only adding the blind-written test ($nw \longrightarrow w(b) + nw$), and a hiring process using a mix of written and non-written tools potentially dropping the non-written and blinding the written ($w + nw \longrightarrow w(b)$) or just blinding the written ($w + nw \longrightarrow w(b) + nw$). The potential treatment $w \longrightarrow w(b) + nw$ is accounts for less than 1% of transitions in the data. Written exams are shorthand for written or multiple-choice tests, and non-written indicate interviews, practical exams, or oral exams. Numbers in $[\cdot]$ give the frequency of each treatment type in the estimating sample.

FIGURE A.6: Potential Treatment Space Generated by the Impartiality Requirement

TABLE A.1: Examples of Occupations in Each Treatment Type

Treatment	Occupation
$w \longrightarrow w(b)$	translator, librarian, low-level bureaucrat, school officer, office staff, community healthcare worker...
$nw \longrightarrow w(b)$	accountant, psychologist, telephonist, social worker, truck driver...
$nw \longrightarrow w(b) + nw$	nurse, pharmacist, receptionist, typist, visual designer, chemist...
$w + nw \longrightarrow w(b)$	kitchen assistant, nutritionist, cleaner, security agent, courier, civil engineer, programmer...
$w + nw \longrightarrow w(b) + nw$	teacher, judge, mason, cook, professor, high-level bureaucrat, veterinarian, police officer...

Notes: These are examples of occupation titles in each treatment type generated by the introduction of impartiality in hiring in Brazil's public sector.

Table A.2: Raw Text Data Availability: Government Official Gazettes

Entity	Online Archives Available Since	Government Level
Brazil	**1808**	**Federal**
Rondônia	2011	State
Acre	2010	State
Amazonas	**1956**	**State**
Roraima	1998	State
Pará	2016	State
Amapá	1988	State
Tocantins	2005	State
Maranhão	2001	State
Piauí	2004	State
Ceará	1999	State
Rio Grande do Norte	—	State
Paraíba	2003	State
Pernambuco	**1936**	**State**
Alagoas	2010	State
Sergipe	2012	State
Bahia	2007	State
Minas Gerais	2011	State
Espírito Santo	2006	State
Rio de Janeiro	2005	State
São Paulo	**1891**	**State**
Paraná	2004	State
Santa Catarina	2011	State
Rio Grande do Sul	**1968**	**State**
Mato Grosso do Sul	**1979**	**State**
Mato Grosso	1967	State
Goiás	2008	State
Distrito Federal	**1967**	**State**

Notes. This table shows the primary sources of job hiring processes in various levels in Brazil's public sector. Each administrative level displayed publishes its own official gazette in a separate online repository. The middle column lists dates when online archives of each journal became available.

TABLE A.3: Estimates of Screening Impartiality on Job Process Outcomes

	Log # Candidates		
	All	Women	Men
	(1)	(2)	(3)
$\text{Federal}_j \times \text{Post}_{j(t)}$	-0.245	-0.212	-0.316
	(0.322)	(0.382)	(0.251)
Occupation FE	X	X	X
Year FE	X	X	X
Obs.	86,959	86,959	86,959

Notes: The table shows difference-in-differences estimates of the log number of job applicants using the regression:

$$\log\left(\#\ \text{Candidates}_{jt}\right) = \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_{o(j)} + \theta_t + u_{ujt}$$

where $\text{Post}_{j(t)}$ is a dummy equal to 1 if a job process occurred $t \geq 1989$ and $\text{Federal}_j = 1$ if it was for a federal job. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE A.4: Estimates of Screening Impartiality on Candidate Scores, Candidate FE

	Final Score			
	Women		Men	
	(1)	(2)	(3)	(4)
$\text{Federal}_j \times \text{Post}_{j(t)}$	0.067**	0.176***	−0.075*	−0.233***
	(0.030)	(0.018)	(0.037)	(0.092)
Applicant FE		X		X
Occupation FE	X	X	X	X
Year FE	X	X	X	X
Obs.	54,892	10,324	32,067	7,825

Notes: The table shows difference-in-differences estimates of outcomes for applicant i in job process j in a given year t. The first specification is $y_{ijt} = \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j \right) + \delta_{j(o)} + \theta_t + u_{ijt}$ only with female candidates in column (1) and only with male candidates in column (3). The second specification includes an additional job applicant fixed effect, in columns (2) and (4). The outcome y represents either a candidate's final score, written score (written exams), or non-written score (interview, oral, practical exams). If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE A.5: Effects of Hiring Impartiality on Gender Gaps: Alternative Specifications and Robustness

	Final Score							Relative Rank	Relative Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\text{Federal}_j \times \text{Post}_{j(t)}$	0.141**	0.141***	0.141**	0.127***	0.127***	0.139***	0.140***	−0.025***	0.038**
$\times \text{Female}_i$	(0.064)	(0.017)	(0.064)	(0.050)	(0.050)	(0.055)	(0.050)	(0.011)	(0.018)
Clustering									
Employer	*X*								
Government level		*X*							
Occupation			*X*						
Job process				*X*	*X*	*X*	*X*	*X*	*X*
Fixed effects									
Occupation code	*X*	*X*	*X*					*X*	*X*
Occupation skill				*X*					
Census region					*X*				
Occupation code×Year						*X*			
Year	*X*	*X*	*X*	*X*	*X*		*X*	*X*	*X*
Controls									
Applicant pool size							*X*		

Notes: The table reports alternative specifications of the baseline triple-difference model in the paper: $y_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right) + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$. Columns (1) through (3) maintain this specification but alternatively cluster standard errors by employer codes, government (at the treatment level), or occupation codes, instead of the baseline clustering by job processes. Column (4) replaces the occupation code fixed effect $\delta_{j(o)}$ with a set of occupational skill dummies (less than high school high school, less than college, college), column (5) uses Brazilian census regions dummies (North, Northeast, South, Southeast, West), and column (7) uses the interaction $\delta_{j(o)} \times \theta_t$ as fixed effect. Column (7) runs the baseline model but includes the applicant size pool in each job process as control. Column (8) also runs the baseline specification, but uses a job applicant's relative rank in the selection process instead of the standardized final score. A relative rank is the candidate's position in proportion to the total number of ranked applicants, such that the best-ranked candidate is measured as $\frac{1}{\#\ \text{candidates}}$. Column (9) uses a relative final score as outcome, measured as the candidate's score divided by the highest score in that job process, such that the best-ranked applicant is normalized to 1.

TABLE A.6: Estimates of Effects on Scores and Hiring Probability, Skill Level

	Final Score			$\mathbb{P}(\mathit{Hired}\vert\text{Applied}=1)$		
	<High School	High School	College or Advanced Degree	<High School	High School	College or Advanced Degree
	(1)	(2)	(3)	(4)	(5)	(6)
Federal$_j$ × Post$_{j(t)}$ × Female$_i$	−0.016 (0.084)	0.160 (0.111)	0.204*** (0.080)	0.002 (0.003)	−0.001 (0.002)	0.011* (0.005)
Occupation FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Obs.	35,475	22,071	29,413	35,475	22,071	29,413

Notes: This table reports triple-difference estimates of the model $y_{ijt} = \delta_1\text{Female}_i + \delta_2\text{Federal}_j + \delta_3\left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4\left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5\left(\text{Federal}_j \times \text{Female}_i\right) + \beta\left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$, with final scores as the outcome in columns (1) through (3) and a dummy equal to one if the candidate is hired for columns (4) to (6). Within each outcome, columns represent regressions in subsamples based on the education required by a job process from applicants. If a job process occurred $t \geq 1989$, Post$_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, Federal$_j = 1$. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE A.7: Estimates on Scores and Hiring Probability, Occupation Feminization

	Final Score			$\mathbb{P}(Hired \| \text{Applied} = 1)$		
Occupation Gender Identity	Female	Neutral	Male	Female	Neutral	Male
A. All Job Applicants	(A.1)	(A.2)	(A.3)	(A.4)	(A.5)	(A.6)
$\text{Federal}_j \times \text{Post}_{j(t)}$ $\times \text{Female}_i$	0.305*** (0.105)	0.149*** (0.043)	0.267** (0.117)	0.012* (0.006)	0.005* (0.002)	0.002** (0.001)
B. Female Job Applicants	(B.1)	(B.2)	(B.3)	(B.4)	(B.5)	(B.6)
$\text{Federal}_j \times \text{Post}_{j(t)}$	0.066*** (0.019)	0.061** (0.026)	0.151* (0.091)	0.003** (0.001)	0.0007 (0.001)	−0.0005 (0.001)
C. Male Job Applicants	(C.1)	(C.2)	(C.3)	(C.4)	(C.5)	(C.6)
$\text{Federal}_j \times \text{Post}_{j(t)}$	−0.238*** (0.089)	−0.087** (0.038)	−0.115*** (0.035)	−0.008* (0.005)	−0.004** (0.002)	−0.002*** (0.001)
Occupation FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Obs.	48,681	16,848	21,430	48,681	16,848	21,430

Notes: This table displays regression coefficients of the triple-differences specification $y_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right) + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$ in Panel A, and of the difference-in-differences model $y_{ijt} = \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_{j(o)} + \theta_t + u_{ijt}$ only with female applicants in Panel B and male applicants in Panel C. The outcome represents either a female candidate's final score relative to a male candidate in the first three columns, or the probability that a female candidate receives a job offer relative to a male candidate in the last three columns. If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$. Columns (1) and (3) report estimates for a subsample of female-dominated occupations, defined as the proportion of women in that occupation $> 60\%$. Columns (2) and (5) run the regression for a subsample of occupations that are neutral or gender balanced, if the proportion of women in that occupation is between 40% and 60%. Columns (3) and (6) run the regression for a subsample of male-dominated occupations, defined as the share of women $< 40\%$. Standard errors are clustered at the job process level. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE A.8: Correlation Between Occupation Characteristics and Selection into Treatment

	$\mathbb{P}\left(nw \longrightarrow w(b) + nw \mid pre \in nw\right)$	$\mathbb{P}\left(w + nw \longrightarrow w(b) + nw \mid pre \in w + nw\right)$
	(1)	(2)
High-skill	0.183	0.276*
	(0.155)	(0.152)
Female-dominated	0.061	0.068
	(0.151)	(0.150)
% female applicant	0.718	0.771
	(0.611)	(0.628)
Candidates/openings	−0.002	−0.001
	(0.004)	(0.003)
R-squared	0.06	0.09
Occupations	52	47

Notes: This table shows regressions at the occupation-level determining the relationship between the probability that an occupation follows one of two possible treatments and its characteristics. The first column regresses occupations that only used an interview before the impartiality requirement (nw) on the probability of switching to $w(b) + nw$ instead of $w(b)$. Column (2) regresses the probability that occupations using written and non-written exams before the time of treatment ($w + nw$) receive treatment $w(b) + nw$ as opposed to $w(b)$.

TABLE A.9: Selection Tests for Multiple Treatments

	Final Score				$\mathbb{P}\left(Hired = 1\right)$			
Pre-treatment mix:	nw		$w + nw$		nw		$w + nw$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Overidentification test								
p-value	0.86	0.36	0.91	0.75	0.14	0.17	0.00	0.61
χ^2 statistic	3.9	8.7	10.7	15.4	12.2	11.5	122.2	18.66
Female applicant	X		X		X		X	
Male applicant		X		X		X		X
Occupation FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X

Notes: This table reports the results of overidentification tests for the multiple treatments generated by the impartiality requirement in hiring. The reported Sargan (1958) statistics compare whether just-identified estimates using alternative job titles as instruments for a treatment are equal under the null hypothesis. Columns (1), (2), (5), and (6) include treatments $nw \longrightarrow \{w(b), w(b) + nw\}$, while the remaining columns include treatments $w + nw \longrightarrow \{w(b), w(b) + nw\}$. The tests are generated with the model

$$y_{ijt} = \kappa_2 \text{Treatment}_{j(d)} + \kappa_3 \left(\text{Post}_{j(t)} \times \text{Treatment}_{j(d)}\right) + \delta_{o(j)} + \theta_t + u_{ijt}$$

where $\text{Treatment}_{j(d)}$ replaces the $Federal_j$ dummy, mapping to each possible treatment conditional on the pre-existing mix, and equal to 0 for state jobs. Regressions are ran separately for men and women. The outcomes are a job applicant final score and probability of being hired.

TABLE A.10: Treatment Effects on Gender Gaps of Changes in Screening Tools

A. Gender Final Score Gap Gap					
	$w \longrightarrow w(b)$	$nw \longrightarrow w(b)$	$nw \longrightarrow nw + w(b)$	$w + nw \longrightarrow w(b)$	$w + nw \longrightarrow nw + w(b)$
	(A1)	(A2)	(A3)	(A4)	(A5)
Federal$_j$ × Post$_{j(t)}$	0.040***	0.195***	0.141***	0.026*	0.002
× Female$_i$	(0.016)	(0.044)	(0.016)	(0.014)	(0.025)
B. Gender Hiring Gap					
	$w \longrightarrow w(b)$	$nw \longrightarrow w(b)$	$nw \longrightarrow nw + w(b)$	$w + nw \longrightarrow w(b)$	$w + nw \longrightarrow nw + w(b)$
	(B1)	(B2)	(B3)	(B4)	(B5)
Federal$_j$ × Post$_{j(t)}$	0.005**	0.070***	0.058***	0.005	−0.005
× Female$_i$	(0.002)	(0.013)	(0.007)	(0.003)	(0.007)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Obs.	12,066	9,343	18,570	44,964	32,898

Notes: This table shows treatment effects estimated from the triple-difference model:

$$y_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right)$$

$$+ \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right)$$

$$+ \beta_d \left(\text{Federal}_j \times \text{Post}_{t(j)} \times \text{Female}_i\right) + \delta_{o(j)} + \theta_t + u_{ijt}$$

which compares outcomes (y_{ijt}) for female candidates relative to men (Female_i) participating in job processes for the same occupation ($\delta_{o(j)}$) that had screening practices changed following treatment d. Treatments are displayed at the top with their corresponding estimates in the bars displayed below. Bars are 95% confidence intervals obtained with standard errors clustered at the job process level.

TABLE A.11: Decomposing Treatment Effects by Weight of Blind Stages

	Treatment Group (d): $w + nw \longrightarrow w(b) + nw$								
	Final Score					Non-Blind Score		$\mathbb{P}(Hired \mid \text{Applied} = 1)$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\text{Federal}_j \times \text{Post}_{j(t)}$ $\times \text{Female}_i$	0.009 (0.014)	0.052*** (0.022)	-0.011 (0.042)	0.059*** (0.017)	-0.039 (0.040)	0.011 (0.008)	-0.006 (0.009)	0.014*** (0.005)	0.009 (0.026)
Blind Score				0.630*** (0.049)	0.446*** (0.029)				
Job Process Blind Weight		$> 50\%$	$< 50\%$	$> 50\%$	$< 50\%$	$> 50\%$	$< 50\%$	$> 50\%$	$< 50\%$
Occupation FE	X	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X	X

Notes: This table plots treatment effects for treatment type d where jobs process transition from using a mix of written and non-written tools to blinding the written ($w + nw \longrightarrow w(b) + nw$). The model implemented is the triple-differences specification $y_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right) + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$. If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$. The outcome y represent either a candidate's final score, non-blind (non-written) score, or probability of being offered a job. Columns (2), (4), (6) and (8) condition on the weight on the blind written test in a job process to be $> 50\%$, and columns (3), (5), (7) and (9) for the weight on the blind written test to be less than 50%. Standard errors are clustered at the job process level.

TABLE A.12: Treatment Effects of Changes in Screening Tools: Skill Robustness

	$\mathbb{P}(Hired \mid \text{Applied} = 1)$				
	$w \longrightarrow w(b)$	$nw \longrightarrow w(b)$	$nw \longrightarrow nw + w(b)$	$w + nw \longrightarrow w(b)$	$w + nw \longrightarrow nw + w(b)$
	(1)	(2)	(3)	(4)	(5)
$\text{Federal}_j \times \text{Post}_{j(t)}$	0.005**	0.070***	0.058***	0.003	-0.004
$\times$ Female_i	(0.002)	(0.013)	(0.007)	(0.004)	(0.008)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Occupation Skill FE	X	X	X	X	X
Obs.	12,066	9,343	18,570	44,964	32,898

Notes: This table shows treatment effects for each treatment type d implementing the triple-difference specification $\mathbb{P}(Hired = 1)_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right) + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$, with an included set of fixed effects accounting for occupation skill. If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE A.13: Treatment Effects of Changes in Screening Tools: Feminization Robustness

	$\mathbb{P}(Hired \mid \text{Applied} = 1)$				
	$w \longrightarrow w(b)$ (1)	$nw \longrightarrow w(b)$ (2)	$nw \longrightarrow nw + w(b)$ (3)	$w + nw \longrightarrow w(b)$ (4)	$w + nw \longrightarrow nw + w(b)$ (5)
$\text{Federal}_j \times \text{Post}_{j(t)}$ $\times \text{Female}_i$	0.005** (0.002)	0.070*** (0.013)	0.057*** (0.007)	0.005 (0.003)	-0.005 (0.007)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Occ. Feminization FE	X	X	X	X	X
Obs.	12,066	9,343	18,570	44,964	32,898

Notes: This table shows treatment effects for each treatment type d implementing the triple-difference specification $\mathbb{P}(Hired = 1)_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right) + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$, with an included set of fixed effects accounting for the degree of feminization of the occupation (female or male dominated, with neutral as the omitted group). If a job process occurred $t \geq 1989$, $\text{Post}_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, $\text{Federal}_j = 1$. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

TABLE A.14: Treatment Effects of Changes in Screening Tools by Occupation Skill

	$\mathbb{P}(Hired \mid \text{Applied} = 1)$				
	$w \longrightarrow w(b)$	$nw \longrightarrow w(b)$	$nw \longrightarrow nw + w(b)$	$w + nw \longrightarrow w(b)$	$w + nw \longrightarrow nw + w(b)$
Panel A. Less Than College	(A.1)			(A.4)	(A.5)
Federal$_j$ × Post$_{j(t)}$ × Female$_i$	0.004*** (0.001)			0.012* (0.006)	0.006 (0.011)
Panel B. College or More	(B.1)	(B.2)	(B.3)	(B.4)	(B.5)
Federal$_j$ × Post$_{j(t)}$ × Female$_i$	0.005* (0.002)	0.085*** (0.020)	0.055*** (0.006)	−0.006 (0.006)	−0.017 (0.011)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X

Notes: This table shows treatment effects for each treatment type d implementing the triple-difference specification $\mathbb{P}(Hired = 1)_{ijt} = \delta_1 \text{Female}_i + \delta_2 \text{Federal}_j + \delta_3 \left(\text{Post}_{j(t)} \times \text{Federal}_j\right) + \delta_4 \left(\text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_5 \left(\text{Federal}_j \times \text{Female}_i\right) + \beta \left(\text{Federal}_j \times \text{Post}_{j(t)} \times \text{Female}_i\right) + \delta_{j(o)} + \theta_t + u_{ijt}$. If a job process occurred $t \geq 1989$, Post$_{j(t)}$ is a dummy equal to 1, and if it was for a federal job, Federal$_j$ $= 1$. Panel A uses a subsample of candidates applying to jobs requiring less than a college degree, while Panel B estimates off job applicants to college positions. All specifications contain year and occupation fixed effects and have standard errors are clustered at the job process level.

Appendix B Transforming Unstructured Text Into Data

To organize ideas, consider the following sequence of tasks necessary to automate the construction of a large-scale applicant-performance :

1. *Filter out all text contained in official government documents unrelated to hiring steps*
2. *Define the boundaries of the relevant text*
3. *Identify the underlying job process of a certain relevant text*
4. *Link different postings belonging to the same process.*
5. *Transform text in each posting into data.*

Due to the layout of Brazilian official journals, each step above presents a host of issues. First, there are no boundaries between the text of a job posting and other information — say another job posting or a list of government contractors suspended — so that defining the domain of relevant information ex-ante is difficult. Then, because surrounding text may be of a similar nature, filtering out extraneous information that does not belong to a specific job process is also challenging. Further complicating matters, the different stages of the same job process have no exclusive identifier (e.g., a hiring process code) and subsequent postings rarely mention the date that the *Edital* (job announcement) was published. Taken together, these issues underscore the limitations of relying on any text-selection method based on existing content structure to automate steps 1 through 4.

Given that one could identify and link the precise text domain of all stages of a hiring process, extracting data from the raw text presents an even bigger challenge. There is no pre-determined layout or set of rules instructing how postings in the *Diários* should display information. Some postings may present candidate results in tables, others in continuous text; scores may be organized by exam type or committee member, or a combination of both; exam types are sometimes informed near candidates and scores and other times at the beginning of the journal posting. While there is certainly some commonality across official postings, after all, these have legal content and enforcement and are often submitted by specialized bureaucrats on behalf of the employer, these similarities are subtle and offer little aid to scrape-like tools that rely on well-defined patterns.

A Conceptual Implementation

While all steps of hiring processes in the Brazilian public sector are carefully documented and publicly available, there are two major challenges to systematically using these raw data sources. The first is that published notices within the same hiring process are not directly linked. In practice, it is non-trivial to assign a list of candidate scores posted in a certain journal issue to a previously-published job announcement information. Off-the-shelf text analysis tools that connect text bodies based on proportionality and similarity, like the term frequency-inverse document frequency (tf-idf) and cosine similarity, are not useful in this context since information in legal publications is highly confounding. The same page of an official gazette might contain sections with a hiring round of eye surgeons at a certain hospital and a section with another job selection process of brain surgeons at the same hospital. In other cases, the same hospital might be hiring eye surgeons through more than one public notice.

A.1 Defining Textual Matching Attributes

Conceptually, the problem of linking steps of the same job process boils down to connecting a number of T text snippets by matching on N text attributes. Both T and N are ex-ante unknown. A job selection process might have any number T of published texts and it is unclear which and how many N lexical structures one might need to properly connect such announcements. How should N be chosen? Consider that a sequence of $t = 1, ..., T$ connected text documents can be summarized by the set of attributes A^t:

$$A^t = \{\text{message keyword, sender, release date, message keyword feature}\}$$

In the case of a specific hiring process, these attributes take the correspondence

$$\{\text{job, employer, release date, job feature}\}$$

where job feature might refer to the place of work, position title, or any dimension that distinguishes A^t from A^j given $A^t \backslash \{\text{job attribute}\} = A^j \backslash \{\text{job attribute}\}$, $t \neq j$. The motivation for defining A^t stems from its search-query use. For each government gazette issue, I search for a job posting notice, using a combination of words in the same paragraph (formally defined as some text string neighborhood) comprised of "announcement", "job", "hiring", and "posting". When there is one or more hits, I bound the relevant text to each job announcement and extract attributes A (the implementation of relevant text boundaries is detailed below). Only the *release date* is ex-ante known, since I know when each journal issue is published. To correctly identify the terms containing the other attributes in A, I rely on ad hoc dictionaries and allow them to expand by "learning" new terms.

More precisely, I construct a list with all public entities from government webpages and a dictionary of occupations that provide a fairly broad library to search for full or partial matches in job announcement texts. After I identify a job (message keyword) and employer (sender) pair, I update these dictionaries used in the search query for the same keyword. For example, my initial occupation library contains "Professor" and adds terms like "Assistant Professor", "Associate Professor", and so on as I progressively incorporate richer versions of the message keyword "Professor". After building the set of attributes that uniquely identifies a job hiring process, I search in all documents published after the release date for occurrences of A^t. The collection of T text excerpts containing A^t thus comprises all published notices of the job selection process.[34]

Suppose a researcher wants to use the New York Times online archives to collect data on murder rates in major US cities since 1890. In this case, the message keyword could be "murder rate", a list with the desired city names would inform different values for the sender, the release date is the issue's date, and the message keyword feature could be a year matching the release date. Instead of going through multiple manual searches in the archived texts for each combination of city and year, the results to the approach above would give the relevant text snippets for the next stage: transforming the text into data.

A.2 Data Extraction

After linking hiring rounds across government gazette issues, the next challenge is the lack of structure in the published notices. Hiring rounds might be displayed in tables of varying dimensions, in free text, or in a combination of both. In most text analysis applications (e.g. Atalay et al. (2020)) every text snippet has a fairly similar structure, which greatly facilitates mining. Moreover, even in cases with free text as in Bybee et al. (2020), the underlying text structure is relevant only to the extent that it conveys information to identify a predictor based on

[34] Note that while still relying on some dictionaries to discipline the domain of the message keyword and sender types, this approach takes an agnostic view with respect to the information derived from the underlying text contents and its potential use to connect text snippets, as well as the need for computationally-intensive updating of the initial search libraries. Indeed, in most applications, researchers may not even need to update their initial search parameters.

the message content. That is, researchers map text (raw or represented by a numerical array) onto a discrete set of measures $\mathcal{T} \rightarrow \{M_1(\tau), M_2(\tau), \cdots, M_K(\tau)\}$, where τ is a transformation of the underlying raw text. Such mappings include sentiment-based approaches as in Gentzkow et al. (2019), where the true sentiment of a message is transformed into a function of a latent quantity.

In many applications, however, researchers might be interested in extracting exact information from text and converting that into a database by distilling $\mathcal{T}$ into a pre-determined list of variables $\{x_1, x_2, \cdots, x_K\}$. This is usually an extremely time-intensive task, highly dependent on the particular context and that relies heavily on strong prior information about the potential variations of text structure across $\mathcal{T}$. To solve this issue, I leverage the semantic structure implied by the relationship between listed variables, so that a fixed variable x_1 conditions all other data points that a researcher is interested in extracting, $\{x_1, x_2|x_1, \cdots, x_K|x_1\}$.

To see how, let x_1^i correspond to candidate i's exam score, x_2^i her name, x_3^i the exam type and x_4^i the committee member who gave score x_1^i. In order to deal with the unstructured nature of the text, I start by targeting text tokens containing numbers, which is the only morphology that maps onto exam scores. Of course, many numbers within the text might be extraneous and not represent scores. The next step searches for tokens in the neighborhood of every number that match the characteristics of each additional variable x. This both fully defines the other variables that relate to x_1 and filters out numeric elements that are not scores.[35]

Often times the implementation of an automated tool to extract data in these cases is so burdensome that researchers end up hand-collecting the desired variables from a feasible subsample of text documents. But in many cases this is simply not feasible. By choosing one variable to which most or all of the other desired variables relate, this approach avoids reliance on information from semantic structure that differs across public announcements, and focuses on relationships that organize each candidate's relevant information in the same way within a job notice text. The underlying semantic structure thereby informs the selection model about the location of certain variables rather than feed a label grouping, such as political slant or favorability of a review. This step requires the use of few *ad hoc* dictionaries (a list with Brazilian names in the current application and another with different examination types), which are allowed to learn similarly to before with the lists of occupations and employer names.

Returning to the application example of historical murder rates in major US cities, after defining the relevant NYT articles containing murder rates of a city in a given year (step 1), now the researcher implements step 2 to extract the actual number from the text (x_1), which is the murder rate. The process here is simple since *i)* the murder rate number only has two relevant attributes — city and period or year. Of course, numeric values of x_1 may give different scales or measurements of murder figures, for which the researcher will need to implement some form of ex-post harmonization.

[35] For instance, numbers without recognizable names in their vicinity are discarded. Further, the same candidate might have several scores for different exams, which will differ along some dimension (Exam I and Exam 2, Written Exam and Oral Exam, etc.). This attribute will be relevant not only for individual i's score, but also for all other candidates who took the same exam type. Thus, it must be that the relation between x_1^i and x_3^i holds for all $i \neq g$. For example, if the data is organized in a table where a certain column contains each exam type and rows display candidate names and scores, each candidate's score in a given exam will be aligned with the column's name. Another example: if the beginning of recorded scores displays a legend that gives an ordering such as "*Name - ID # - Written Exam - Interview - Final Score - Rank*", every candidate will have scores displayed in the same order.

Appendix C Model Solution

This Appendix shows the complete solution to the model described in Section 5. The environment and structure of hiring decision are as described there.

A Hiring Rates

A.1 Hiring Rate With Written Exams

When the hiring technology only includes written tests, the distribution of written signals, $s^* = y + \nu_s(x) + \varepsilon_s$, $\varepsilon_s \sim N(0, 1/h_s)$, is given by:

$$s^* \sim N\left(y + \nu_s\left(x\right), 1/h_s\right)$$

where s represents the unbiased signal $s = y + \varepsilon_s$, h_s is the inverse of the variance of the written signal, measuring the precision of written testing and independent of group membership x. After observing s^*, the hiring manager updates her assessment of expected productivity of candidates, initially based on group productivity, $\mu_0(x)$, forming the posterior:

$$\mu(x, s^*) = s\frac{h_s}{h_0 + h_s} + \mu_0(x)\frac{h_0}{h_0 + h_s} + \nu_s(x).$$

Given the written signal, s, and the perceived group productivity, $\mu_0(x)$, the evaluator updates her assessment of expected productivity of candidates according to:

$$y \mid_s \sim N(\mu(x, s), 1/(h_0 + h_s)).$$

Here, the updated degree of precision is $(h_0 + h_s)$ and the updated mean equals:

$$\mu(x, s) = s\frac{h_s}{h_0 + h_s} + \mu_0(x)\frac{h_0}{h_0 + h_s} + \nu_s(x).$$

The hiring decision that maximizes the evaluator's objective function satisfies the rule Hire $= I\{\mu(x, s) > k_s\}$, where k_s is the screening threshold that yields a fixed hiring rate of $K \in (0, 0.5)$. Plugging the expression for $\mu(x, s)$ into the hiring rule yields the following:

$$s > \frac{(h_0 + h_s)(k_s - \nu_s(x) - d_s\pi_j(x)) - h_0\mu_0(x)}{h_s}.$$

Since the distribution of s is $N(\mu_0(x), 1/h_0 + 1/h_s)$, the above inequality can be rewritten as:

$$\frac{s - \mu_0(x)}{\frac{1}{h_0} + \frac{1}{h_s}} > \frac{(h_0 + h_s)(k_s - \nu_s(x) - d_s\pi_j(x) - \mu_0(x))}{h_s\sqrt{(\frac{1}{h_0} + \frac{1}{h_s})}}$$

which can finally be expressed as:

$$\frac{s - \mu_0(x)}{\frac{1}{h_0} + \frac{1}{h_s}} > \underbrace{\frac{k_s - \nu_s(x) - \mu_0(x) - d_s\pi_j(x)}{\sigma_0\rho_s}}_{z_s^*(x)} \tag{C.1}$$

where $\rho_s \equiv Corr(\mu(x,s),y) = (1 - \frac{h_0}{h_0+h_s})^{1/2}$ and $z_s^*(x)$ is the hiring threshold for group x determined by using written exams. The expected hiring rate of applicants from group x is $1 - \Phi(z_s^*(x))$.

A.2 Hiring Rate With Non-Written Exam

When the employer screens job applicants solely based on non-written tests, the intuition for the effect of evaluator bias, precision, and tool bias is similar to the case of written tests. However, an important distinction arises as a consequence of different subjectivity degrees between the two practices. Formally, let the distribution of non-written signals be $\eta^* = y + \nu_\eta(x) + \varepsilon_\eta,\ \varepsilon \sim N(0, 1/h_\eta)$, where $\nu_\eta(x)$ represents the possible disparate impact of non-written tests and η is the unbiased non-written signal, $\eta = y + \varepsilon_\eta$. Non-written exams allow discretion d_η to evaluators. Given that interviews or oral exams are more subjective than written tests, the discretion given to the evaluator is higher with non-written than with written tests: $d_\eta > d_s$.

The hiring rate for group x when candidates are screened using non-written tests is obtained following the same steps as in the previous case, observing the different distribution of non-written signals. In this case, an applicant screened with a non-written exam is hired if

$$\frac{\eta - \mu_0(x)}{\frac{1}{h_0} + \frac{1}{h_\eta}} > \underbrace{\frac{k_\eta - \nu_\eta(x) - \mu_0(x) - d_\eta \pi_j(x)}{\sigma_0 \rho_\eta}}_{z_\eta^*(x)} \tag{C.2}$$

The corresponding probability that a group x candidate is hired is given by $1 - \Phi(z_\eta^*(x))$.

A.3 Hiring Rate With Written and Non-Written Exams

Given the two signals previously determined, η^* and s^*, and the perceived group productivity, $\mu_0(x)$, the evaluator updates her assessment of expected productivity according to:

$$y \mid_{\eta^*, s^*} \sim N(\mu(x, \eta^*, s^*), 1/(h_0 + h_\eta + h_s)).$$

From the above, the updated degree of screening precision is $h_0 + h_\eta + h_s \equiv h_T$ and the updated posterior is thereby:

$$\mu(x, \eta^*, s^*) = s\frac{h_s}{h_T} + \eta\frac{h_\eta}{h_T} + \mu_0(x)\frac{h_0}{h_T} + \nu_s(x)\frac{h_s}{h_T} + \nu_\eta(x)\frac{h_0 + h_\eta}{h_T}.$$

Thus, the hiring decision is given by:

$$\mu(x, \eta, s) > k_T - \pi(x)(d_\eta + d_s)$$

$$\frac{h_s s + h_\eta \eta + h_0 \mu_0(x)}{h_T} > k_T - \pi(x)(d_\eta + d_s) - \nu_s(x)\frac{h_s}{h_T} - \nu_\eta(x)\frac{(h_0 + h_s)}{h_T}.$$

Since $\eta = y + \varepsilon_\eta$, $s = y + \varepsilon_s$, and $y, \varepsilon_\eta, \varepsilon_s$ are independent, the left-hand side of the above inequality is distributed as:

$$\begin{aligned}
\frac{h_s s + h_\eta \eta + h_0 \mu_0(x)}{h_T} &\sim N\left(\mu_0(x), \left(\frac{h_s}{h_T}\right)^2 \left(\frac{1}{h_0} + \frac{1}{h_s}\right) + \left(\frac{h_\eta}{h_T}\right)^2 \left(\frac{1}{h_0} + \frac{1}{h_\eta}\right) + 2\frac{h_s}{h_T}\frac{h_\eta}{h_T}\frac{1}{h_0}\right) \\
&\sim N\left(\mu_0(x), \frac{h_s^2 + h_\eta^2 + h_0^2 - h_0^2 + 2h_s h_\eta h_s h_0 + h_\eta h_0}{h_0 h_T^2}\right) \\
&\sim N\left(\mu_0(x), \frac{h_T - h_0}{h_0 h_T}\right) \\
&\sim N\left(\mu_0(x), \sigma_0^2 \rho_T^2\right).
\end{aligned}$$

Further manipulation finally gives the hiring threshold:

$$\frac{\mu(x,\eta,s) - \mu_0(x)}{\sigma_0 \rho_T} > \underbrace{\frac{k_T - \frac{h_s}{h_T}\nu_s(x) - \frac{h_0 + h_\eta}{h_T}\nu_\eta(x) - \pi_j(x)(d_\eta + d_s) - \mu_0(x)}{\sigma_0 \rho_T}}_{z_T^*(x)}. \tag{C.3}$$

With two screening tools, evaluators place less weight on their group priors, which favors the hiring threshold of the minority group if $\mu_0(m) > \mu_0(f)$. Moreover, the overall bias now captures bias from both tools.

A.4 Hiring Rate With Blind Written Exam

Within the model, blinding makes it impossible to assign individual candidates to a group, since hiring evaluators cannot observe whether a certain signal is generated by a male or female candidate. A blind written exam provides the following signal:

$$s(b)^* = y + \nu_s(x) + \varepsilon_s, \;\; \varepsilon \sim N(0, 1/h_s)$$

with the same screening precision h_s and the same disparate impact $\nu_s(x)$ as the written test. When the screening technology includes blind written tests, the evaluator's objective becomes:

$$u\left(y, \pi\left(x\right)\right) = y + \underbrace{\left(1 - c_{s(b)}\right)}_{d_{s(b)}=0} \pi(x) \equiv y,$$

as discretion is entirely removed from the screening tool. Additionally, blinding the written test affects how the evaluator updates perceived candidate productivity, using the written signal, s, and the perceived *population* productivity, $\mu_0 = \frac{\mu_0(x) + \mu_0(y)}{2}$, since group membership is not identifiable:[36]

$$\mu(x, s(b)^*) = s\frac{h_s}{h_0 + h_s} + \frac{\mu_0(x) + \mu_0(y)}{2}\frac{h_0}{h_0 + h_s} + \nu_s(x).$$

[36] For simplicity and without loss of generality, I assume that each group comprises half of the candidate pool. Another reason for using identical gender distributions is to keep application behavior from the pool of qualified workers outside the model.

Manipulating the above gives the hiring threshold for group x:

$$\frac{s-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_s}} > \frac{(h_0+h_s)(k_{s(b)}-\nu_s(x)-\mu_0(x))}{h_s\sqrt{(\frac{1}{h_0}+\frac{1}{h_s})}} - \frac{h_0(\mu_0(y)-\mu_0(x))}{2h_s\sqrt{(\frac{1}{h_0}+\frac{1}{h_s})}}$$

and finally,

$$\frac{s-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_s}} > \underbrace{\frac{k_{s(b)}-\nu_s(x)-\mu_0(x)}{\sigma_0\rho_s} - \frac{h_0\rho_s}{2h_s\sigma_0}(\mu_0(y)-\mu_0(x))}_{z^*_b(x)}. \tag{C.4}$$

The hiring threshold for group x determined by $s(b)^*$ is similar to the expression obtained for written test screening, s^*, with an important distinction. While the signal given by the blind written test is just as informative as in the non-blind case, now evaluators update a group-neutral prior.

A.5 Hiring Rate With Blind Written and Non-Written Exams

Lastly, consider blinding a written exam when the screening process also includes a non-written test. This is similar to the previous case of combining screening signals from both exams, except for the blind written exam having no disparate treatment. However, evaluators still rely on group means and express bias in the overall posterior because of the non-written signal. Given the two signals, η^* and $s(b)^*$, the posterior is:

$$y\mid_{\eta^*,s(b)^*} \sim N(\mu(x,\eta^*,s(b)^*),1/h_T)$$

and the updated mean

$$\mu(x,\eta^*,s(b)^*) = \frac{h_s s + h_\eta\eta + h_0\mu_0(x) + \nu_s(x)h_s + \nu_\eta(x)(h_0+h_\eta)}{h_T}.$$

Since η and s can be rewritten as $\eta = y+\varepsilon_\eta$ and $s = y+\varepsilon_s$, and $y,\varepsilon_s,\varepsilon_\eta$ are independent, it follows that:

$$\mu(x,\eta,s(b)) \equiv \frac{h_s s + h_\eta\eta + h_0\mu_0(x)}{h_T} \sim N\left(\mu_0(x),\sigma_0^2\rho_T^2\right)$$

The hiring decision can then be rewritten as:

$$\frac{h_s s + h_\eta\eta + h_0\mu_0(x)}{h_T} > k_{\eta s(b)} - \nu_s(x)\frac{h_s}{h_T} - \nu_\eta(x)\frac{(h_0+h_\eta)}{h_T} - d_\eta\pi(x)$$

$$\frac{\mu(x,\eta,b)-\mu_0(x)}{\sigma_0\rho_T} > \underbrace{\frac{k_{\eta s(b)} - \nu_s(x)\frac{h_s}{h_T} - \nu_\eta(x)\frac{h_\eta+h_0}{h_T} - d_\eta\pi(x) - \mu_0(x)}{\sigma_0\rho_T}}_{z^*_{\eta s(b)}(x)}. \tag{C.5}$$

B Predictions for Gender Hiring Gaps

B.1 Change in Hiring Rate for $w \longrightarrow w(b)$

Without loss of generality, assume that female candidates are less productive on average than men: $\mu_0(f) < \mu_0(m)$, or in other words, that women are the minority group. By blinding the written exam, how do screening thresholds $z_s^*(x)$ and $z_b^*(x)$ compare and thus how are hiring rates affected? By inspecting expressions (C.1) and (C.4) and considering that written tests — whether blind or not — have the same screening precision and disparate impact, since they are otherwise identical save for hiding a candidate's identity, women face a lower hiring threshold in the blinded exam, $z_s^*(f) > z_b^*(m)$, if and only if

$$d_s \pi_j(f) < \frac{h_0 \rho_s^2}{2h_s}(\mu_0(m) - \mu_0(f)).$$

The expression above captures the following intuition. As long as the evaluator favors male candidates, either through statistical discrimination or evaluator bias, blinding the written exam increases hiring rates for women. The right-hand side is always positive since $\mu_0(f) < \mu_0(m)$, and it represents the improvement in women's hiring odds from removal of the ability to statistically discriminate. Therefore, if the left-hand side — which captures evaluator bias — is negative, i.e., if hiring managers favor men, or if it is sufficiently small due to either low discretion or low bias, then the hiring rate for women increases and the hiring rate for men decreases after blinding the written exam. Alternatively, if an evaluator is biased in favor of women, blinding the exam curbs the evaluator's ability to balance women's penalty from statistical discrimination with personal bias, potentially decreasing the female hiring rate.

B.2 Change in Hiring Rate for $nw \longrightarrow w(b)$

I now analyze the potential change induced by the policy that most dramatically alters the mix of screening tools. To build intuition, consider an employer that solely relies on interviews to screen candidates. From the expression in (C.2), the disparate impact of interviews, their precision, and how much they enable evaluator bias to be expressed all determine an applicant's hiring odds. Only in terms of evaluator bias, under the assumption that interviews offer more discretion than written exams, this pre-policy state contains the highest expression of evaluator bias. In contrast, as discussed before, screening solely based on written exams is likely to provide a setting with low disparate treatment.

Assume $h_s = h_\eta$ and $\mu_s = \mu_\eta$. It follows that the hiring threshold for men is higher with the blind-written signal than with the non-written signal, $z_\eta^*(m) < z_b^*(m)$, as long as evaluators favor men $\pi_j(m) > 0$ or, alternatively, if the following is satisfied

$$\frac{d_\eta \pi_j(m) + (k_b - k_\eta)}{\sigma_0 \rho} > \frac{h_0 \rho}{2h_s \sigma_0}(\mu_0(f) - \mu_0(m)),$$

which allows for sufficiently small evaluator bias toward women. Because the above inequality implies a higher threshold for hiring male candidates under blind written compared to non-written screening, it increases selectivity for men, and, given a constant total hiring rate, K, the gender hiring gap decreases.

Next, conduct the same exercise but now allow for written and non-written exams to have different disparate impacts, $\Delta\nu_s \neq \Delta\nu_\eta$, where $\Delta\nu_s = \nu_s(m) - \nu_s(f)$. In this case, changing from non-written screening stages to blind-written exams increases female hiring rates if and only if:

$$\frac{h_0 \rho}{h_s \sigma_0}(\mu_0(f) - \mu_0(m)) < \frac{d_\eta(\pi_j(m) - \pi_j(f)) + (\Delta\nu_\eta - \Delta\nu_s)}{\sigma_0 \rho} \tag{C.6}$$

Note that the left-hand side of the expression above is negative, so that if evaluators are men-favoring and interviews have a larger disparate impact than written exams, the inequality is satisfied and female hiring rates increase. In other words, if the principal substitutes a hiring tool for one that has a smaller disparate impact and eliminates discretion, the change will raise hiring rates of the minority group. More generally, if either evaluator bias favors men, or if the relative bias of non-written tests is lower than that of written tests, it can still increase female hiring as long as it satisfies the inequality above. Another way to interpret the inequality (C.6) is to rewrite it as

$$\frac{h_0 \rho}{h_s \sigma_0}\mu_0(f) + \frac{d_\eta \pi_j(f)}{\sigma_0 \rho} + \frac{(\nu_\eta(f) - \nu_s(f))}{\sigma_0 \rho} < \frac{h_0 \rho}{h_s \sigma_0}\mu_0(m) + \frac{d_\eta \pi_j(m)}{\sigma_0 \rho} + \frac{(\nu_\eta(m) - \nu_s(m))}{\sigma_0 \rho} \tag{C.7}$$

The left-hand side represents the perceived productivity of female applicants, equal to true productivity plus bias, either from the evaluator or screening tool. The right-hand side represents the perceived productivity of male applicants. Thus, if female applicants are perceived as less productive under non-written screening relative to written screening, then the transition increases their hiring rate.

Finally, relax the assumption of identical screening precisions. If written tests are more precise, $h_s > h_\eta$, switching from non-written screening to written testing raises the hiring rate of the group with lower perceived productivity, that is, it raises the female hiring rate if (C.7) holds. However, if interviews have higher precision, $h_s < h_\eta$, the transition from interviews to written test decreases screening precision and leads to higher hiring rates of the favored group, men. The net effect then depends on the losses from decreased screening precision relative to the gains from lower bias if (C.7) is satisfied.

B.3 Change in Hiring Rate for $nw \longrightarrow w(b) + nw$

This case maintains the use of non-written exams but, to comply with the impartiality policy, the employer adds a blind-written exam to the hiring process. By having an additional evaluation tool, total hiring precision increases, $h_0 + h_\eta + h_s > h_0 + h_\eta$, without introducing disparate treatment, since $d_b = 0$. Adding the blind-written tool reduces the weight that discretion in the non-written test plays in determining hiring rates (recall that $\frac{d_\eta \pi_j(x)}{\sigma_0 \rho_T} < \frac{d_\eta \pi_j(x)}{\sigma_0 \rho_\eta}$). However, introducing a different screening tool potentially incorporates that tool's disparate impact.

To start assume that screening tools do not favor any group, that is $\nu_\eta(f) = \nu_\eta(m)$, $\nu_s(f) = \nu_s(m)$, and that $\nu_\eta = \nu_s$. Then,

$$z^*_{\eta b}(x) = \frac{k_{\eta b} - \nu(x) - \mu_0(x) - d_\eta \pi_j(x)}{\sigma_0 \rho_T} < \frac{k_\eta - \nu(x) - \mu_0(x) - d_\eta \pi_j(x)}{\sigma_0 \rho_\eta} = z^*_\eta$$

That is, the hiring threshold is lower for group f if $\mu_0(f) + d_\eta \pi_j(f) < \mu_0(m) + d_\eta \pi_j(m)$ — women have lower perceived productivity. For the minority group both effects help as long as the same condition holds: $\mu_0(f) + d_\eta \pi_j(f) < \mu_0(m) + d_\eta \pi_j(m)$.

The increase in screening precision and decrease in relative importance of evaluator bias increase women's hiring rates if they are the group with the lower perceived productivity: $\mu_0(f) + d_\eta \pi_j(f) < \mu_0(m) + d_\eta \pi_j(m)$, reflecting that the change in hiring probability with respect to screening precision is:

$$\frac{\partial\left[1-\Phi(z^*_{\eta b}(x))\right]}{\partial\rho_T}=\phi(z^*_{\eta b}(x))\left[\frac{z^*_{\eta b}(x)}{\rho_T}-\frac{\partial k_{\eta b}/\partial\rho_T}{\sigma_0\rho_T}\right]>0 \tag{C.8}$$

Here $\phi(\cdot)>0$ and $z^*_{\eta b}(m)<z^*_{\eta b}(f)$ if the above inequality of women being perceived as the group with lower productivity is satisfied.

Now, allow for screening tool bias to differ between written and non-written tests and to favor one group, $\nu_\eta(f)\neq\nu_\eta(m)$. Then, women benefit from the added precision if the following inequality holds:

$$\mu_0(f)+d_\eta\pi_j(f)+\nu_s(f)\frac{h_s}{h_T}+\nu_\eta(f)\frac{h_0+h_\eta}{h_T}<\mu_0(m)+d_\eta\pi_j(m)+\nu_s(m)\frac{h_s}{h_T}+\nu_\eta(m)\frac{h_0+h_\eta}{h_T}$$

If the written test that is added is bias increasing, $|\Delta\nu_s|>|\Delta\nu_\eta|$, it causes excess hiring of the group that is favored by the bias. Then, if the bias favors men, $\nu_s(m)-\nu_s(f)\equiv\Delta\nu_s>\Delta\nu_\eta\equiv\nu_\eta(m)-\nu_\eta(f)$, the net effect on the female hiring rate depends on the gains from increased screening precision relative to the losses from increased bias. On the other hand, if the bias favors women, and written tests are more biased than interviews, it leads unambiguously to higher hiring rates of women since all three forces have a positive effect.

B.4 Change in Hiring Rate for $w+nw\longrightarrow w(b)$

Removing the non-written signal from a screening mix of written and non-written decreases total screening precision, $h_0+h_s<h_0+h_s+h_\eta$, removes evaluator bias within the non-written test, $d_\eta\pi_j(x)$, and removes the non-written screening tool bias, $\nu_\eta(x)$. In addition, blinding the written test removes evaluator bias within the exam, $d_s\pi_j(x)$, as well as the use of group means (statistical discrimination) in determining the evaluator's posterior.

To begin with, assume $\nu_s=\nu_\eta$, which does not however eliminate the effect of removing the non-written screening tool bias, but just assumes that the type of tool bias reduced is the same in magnitude and sign (favors the same group), as the bias characterizing the written test.

Removing both screening tool and evaluator biases raises selectivity of the favored group and reduces selectivity of the non-favored group: $z^*_T(m)<z^*_b(m)$. Thus

$$\left[\frac{k_T-\nu(m)-\mu_0(m)}{\sigma_0\rho_T}-\frac{k_b-\nu(m)-\mu_0(m)}{\sigma_0\rho_s}\right]-\frac{\pi_j(m)(d_\eta+d_s)}{\sigma_0\rho_T}+\frac{h_0\rho_s}{2h_s\sigma_0}(\mu_0(f)-\mu_0(m))<0$$

where the inequality holds for m if this is the favored group. Thus, removing the non-written tool and evaluator bias, as well as evaluator bias within the written screening tool reduces selectivity of women and thus raises women hiring rates if they are the non-favored group.

However, the decrease in screening precision due to removal of the non-written signal has the opposite effect on hiring rates of the non-favored group:

$$\gamma_f\equiv\frac{\partial\left[1-\Phi(z^*_T(f))\right]}{\partial\rho_T}=\phi(z^*_T(f))\left[\frac{z^*_T(f)}{\rho_T}-\frac{\partial k/\partial\rho_T}{\sigma_0\rho_T}\right]$$

with ρ_T decreasing as $\rho_S < \rho_T$, $\phi(\cdot) > 0$, and $z_b^*(m) < z_b^*(f)$ if:

$$\frac{\mu_0(f) + \nu_s(f) - (\mu_0(m) + \nu_s(m))}{\sigma_0 \rho_s} + \frac{h_0 \rho_s}{h_s \sigma_0}(\mu_0(m) - \mu_0(f)) < 0$$

This later inequality holds if $\mu_0(f) + \nu_s(f) < \mu_0(m) + \nu_s(m)$ (men are the favored group, perceived to have higher productivity). Note that the inequality of men having the higher perceived productivity can hold even if the written test favors women, $\nu_s(f) > \nu_(m)$, if it is small enough: $\nu_s(f) - \nu_s(m) < \mu_0(m) - \mu_0(f)$. So with ρ decreasing, $\gamma_f < 0$ and $\gamma_m > 0$ if men are the favored group. Consequently, the net effect depends on the positive effect on female hiring rates from decreased bias relative to the negative effect from decreased screening precision.

Third, removing the non-written tool also eliminates its bias, ν_η, which affects hiring rates depending on whether the bias favored men or women, as well as the relative size of this bias compared to the written tool bias. Consider the following cases.

Fist, suppose that the written tool favors women, $\Delta\nu_s < 0$, while the non-written favors men, $\Delta\nu_\eta > 0$, where $\Delta\nu_\theta = \nu_\theta(m) - \nu_\theta(f)$. Then, removing the non-written signal is bias-reducing and reduces excess hiring of the group favored by the non-written bias — men — increasing selectivity for the group and increasing the hiring rate for women. More formally, this follows from:

$$\begin{aligned}
(z_T^*(m) - z_b^*(m)) - (z_T^*(f) - z_b^*(f)) &< 0 \\
\frac{h_s \rho_s - h_T \rho_T}{\sigma_0 \rho_s \rho_T h_T}(\nu_s(f) - \nu_s(m)) + \frac{h_0 + h_\eta}{\sigma_0 \rho_T h_T}(\nu_\eta(f) - \nu_\eta(m)) &< 0 \\
(\nu_\eta(f) - \nu_\eta(m)) &< \frac{h_T \rho_T - h_s \rho_s}{(h_0 + h_\eta)\rho_s}(\nu_s(f) - \nu_s(m))
\end{aligned}$$

where the fraction term is positive from $h_T = h_0 + h_\eta + h_s > h_s$. It follows that the right-hand side is also positive and the left-hand side is negative. This implies an increase in women's hiring rates.

Second, if instead the written signal favors men $\Delta\nu_s > 0$, while the non-written favors women, $\Delta\nu_\eta < 0$, then using the same inequality, it follows that removing the women-favoring bias from non-written increases women's selectivity, decreasing their hiring rate.

Third, if both the written and non-written tools favor men, $\Delta\nu_s > 0, \Delta\nu_\eta > 0$, then, regardless of which bias is larger, removing the non-written signal is bias-reducing and thus reduces excess hiring of the group favored by the bias, men, which in turn increases hiring rate for women. If, instead, both tools favor women, $\Delta\nu_s < 0, \Delta\nu_\eta < 0$, then, similarly, the transition is bias-reducing and decreases excess hiring of the favored group, which in this case are women. This increases selectivity for women, which decreases their hiring rate.