Designing Gender Equity: Evidence from Hiring Practices and Committees*

Tatiana Mocanu†

October 30, 2021

Job Market Paper

Click here for most recent version

Abstract

This paper analyzes how different screening practices and who screens job applicants affect gender equity in hiring. I transform highly dimensional and unstructured text records from Brazil's public sector into hundreds of thousands of selection processes with detailed information on candidates, evaluators, screening tools, and scores. Exploiting exogenous variation in the mix of screening methods used by employers induced by a federal policy, I find that increasing screening impartiality improves women's scores and the probability of being hired relative to men. To isolate and quantify the role of screening precision, evaluator and tool bias from individual hiring practices, I develop a framework exploiting the existence of multiple treatment types induced by the reform. I find that the most effective changes to increase women's hiring odds involve adding blind written stages to a hiring process that uses subjective methods or converting these subjective rounds into only blind written tests. However, when employers remove subjective stages from a screening mix, even blinding the remaining objective tests has no implications to gender equity, likely due to screening precision loss. Finally, higher shares of women in hiring committees are associated with male evaluators becoming more favorable toward female candidates in subjective stages, decreasing aggregate evaluation bias. This suggests that the presence of minority colleagues in the committee makes bias expression more costly.

*I am deeply grateful for the guidance and support from David Albouy, Alex Bartik, Dan Bernhardt, and Russell Weinstein. This paper has benefited greatly from feedback and discussions with Vittorio Bassi, Eliza Forsythe, Andy Garin, Dmitri Koustas, Day Manoli, Pedro Tremacoldi-Rossi, Sergio Urzua, Owen Zidar, as well as seminar participants at the University of Illinois. 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.

†Ph.D. candidate, University of Illinois at Urbana-Champaign. Email: mocanu2@illinois.edu.

1 Introduction

Over the last decades, firms have increasingly devoted resources to grapple with lack of gender diversity and under-representation of women at various levels of the corporate ladder. US firms alone spend more than $10 billion a year in initiatives to reduce implicit and intentional bias in recruiting. Most strategies to improve hiring rates of women and other minorities focus on diversity programs and goals. By most accounts, these initiatives have proven ineffective, as they fail to recognize that the design, implementation, and decision-making during the hiring stage is a main source of low diversity.[1] Certain hiring practices considered important predictors of future productivity may disadvantage a particular group, hiring managers may be biased in difficult ways to observe, or firms may simply fail to attract enough applicants from minority groups.

Even though screening and selecting employees is a central part of every firm and organization, how to design processes that select the best candidates, are bias-free, and improve employee diversity, remains an open question. Employers are reluctant to share hiring practices or details on hiring processes, engaging in lengthy legal battles to keep the information from going public. To complicate things further, even if researchers were able to get detailed data on hiring practices and decisions, generating appropriate variation for causal inference would remain a challenge. 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?"

In this paper, I study how the design of hiring practices and who conducts employee screening determine gender disparities in labor market outcomes. I open the black box of hiring processes by constructing uniquely detailed information on the universe of selection processes from Brazil's public sector. Legal requirements in the country mandate every step of public servant selection to be carefully documented and made publicly available. To access, extract, and transform these records into data, I develop a natural language processing algorithm that distillates over 35 million official government text documents into rich information including all job applicant and hiring manager names, applicant scores by screening method and hiring manager, job openings, job offers, and characteristics of job postings.

Equipped with a large-scale data set detailing job applicant performance and evaluators' decision making process, I first exploit a change in the provisions regulating the selection of public sector employees in Brazil's 1988 Constitution. The reform required government employers at all levels for the first time to conduct impersonal and impartial hiring processes,

[1] Most diversity programs show little to no impact, while some show modest effects (Kalev et al. (2006)), and some can even backfire (Dobbin and Kalev (2016)).

but it was immediately implemented only by federal employers. State and local governments conduct public servant selection independently from central authority, and only started addressing the legal changes necessary to implement impartial hiring processes much later than the federal sector.

The impartiality reform induced variation in the mix of screening tools and hiring practices for federal jobs on multiple fronts. First, using written exams without concealing candidates' identity would be a clear violation of the new rules, leading employers to blind tests. Second, because the reform did not specify which hiring practices had to be implemented to achieve impartiality, multiple treatments to changes in screening methods were generated. Employers modified their mix of hiring tools following occupation-specific historical reliance on certain stages (e.g., oral exams for judges) and customary practices (e.g., typing speed and accuracy for secretaries).

The presence of several complier types allows me to tease out different forces determining hiring rates, as well as understand which design changes in selection processes are effective to increase gender diversity. Should an employer remove screening practices with high discretion even if they may provide employers with important information for screening? Does replacing interviews with objective or standardized tests help or hurt female candidates? Does blinding written exams help at all if no other changes are made in the screening process? To answer these questions, the variation generated by the policy allows me to construct counterfactuals not only with respect to the usual untreated group (no changes in screening practices), but that compare alternative choice paths for screening practices.

To establish the policy take-up, I analyze how the design of hiring processes changed in federal jobs relative to states. Federal employers responded sharply to the impartiality reform. Job announcements started including rules detailing written examinations were to be conducted without information on candidate names, clearly indicating an effort to comply with the impartiality requirement. Relative to the same occupation in state hiring processes, federal jobs became more likely to use written (or multiple-choice) exams, less likely to use a non-written tool, and decreased the number of job processes entirely relying on non-written stages by 25 percentage points.

How did greater impartiality affect male and female job candidates? To study the overall effects of the reform, I employ a difference-in-differences design comparing job processes in the same occupation in federal and state governments and estimate that following the reform, women's final score in job processes increased by 0.07 standard deviation, accompanied by a decrease of similar magnitude in men's scores. This resulted in a drop in the gender score gap of 0.14 standard deviation. Confirming that the policy induced intensive margin changes in scores of only written exams — which had to be blind —, the gender gap in these stages also decreased, while relative scores in non-written tools between men and women had no statistically significant changes.

The decrease in the gender final score gap translates into improved hiring rates of women and a narrower gender hiring gap. Using information on the entire candidate pool, which is rarely available to researchers, allows me to look at applicants' probability of being hired conditional on gender instead of measuring hiring gaps from a sample of hired workers, which confounds employer behavior with application rates. This distinction is important because the design of hiring practices may affect both the minority and majority candidate pool sizes. More broadly, extensive margin responses in job processes with a fixed number of openings may differentially crowd out male and female candidates via increased competition.

I estimate that women became 0.3 percentage points more likely to be hired and men's hiring rates decreased by 0.4 points, implying a reduction in the gender hiring gap in federal jobs of about 44% of the pre-treatment level, even after controlling for job process competitiveness. Interpreting these estimates in light of two advantages to my setting — screening methods for a job process are decided at higher bureaucratic levels and not by hiring managers, and results from all screening stages had to be incorporated into the job offer decision following pre-determined rules — indicate that changes in the underlying mix of screening practices and conducting blind written exams, successfully reduced gender disparities originating from employer's behavior in the federal sector hiring.

The slow progress made by women in the labor market has been sometimes attributed to supply-side explanations.[2] By investigating the separate response of application behavior induced by the policy, I find that application rates of women relative to men increased about 1 percentage point, implying a supply-side response 40% larger than the increase in employer's demand. Taken together, both a larger employer demand for female candidates and application growth result in a 13% increase in gender diversity among employers just few years after the policy came into effect. This is an important finding showing that the way employers screen applicants can increase women representation among applicants.[3]

To guide the interpretation of the effects on gender hiring equity from multiple changes in screening practices, I build on a classic statistical discrimination model by incorporating screening tool characteristics and the role of managers, who conduct the screening on behalf

[2]The gender gap in competition participation is a potential factor in explaining gender differences in career choices and labor market outcomes (Bertrand (2011)). Several studies document gender differences in applying for promotions (Hospido et al. (2019), Bosquet et al. (2019)), as well as selection into competitive environments more generally (Niederle and Vesterlund (2007)). Suboptimal entry by high-performing women can also be costly for employers as it prevents firms from hiring the best candidates.

[3]A few recent studies show how the composition of the job applicant pool responds to simple changes in the wording of job vacancies (Del Carpio and Guadalupe (2021), Abraham and Stein (2020)), as well as to signaling a preference for employee diversity in the content of recruiting information (Flory et al. (2021)).

of the employer. I allow hiring managers to be biased toward a certain demographic group, with the degree of expression of this bias regulated by how much discretion a specific hiring practices enables. Interviews allow for high level of discretion due to their subjective nature, while the results from formal tests are more easily observable to the firm, making bias expression more costly. Independently of manager preferences, screening tools provide a productivity signal with certain precision and potentially mean-biased — where the bias term absorbs group-favoring characteristics of a given practice — generating disparate impact even if managers are unbiased. The conceptual framework pins down different considerations employers face when designing hiring processes with the goal of minimizing inequities without necessarily an efficiency trade-off.

To test the model predictions, I conduct a multi-treatment analysis, which under a set of assumptions that enable the construction of counterfactuals between changes in screening tools that had the same pre-policy mix, I first estimate treatment effects of keeping the pre-existing written exam and blinding it. The estimated reduction in the gender hiring gap of 0.5 percentage points (relative to 1.5 p.p. pre-policy) measures the improvement in diversity from completely eliminating evaluator bias, since now evaluators cannot express disparate treatment, nor rely on statistical discrimination. This result shows that even in a context with likely low levels of evaluator discretion, disparate treatment may still play an important role in determining gender gaps in labor market outcomes.

The next set of comparisons I make starts with a screening process that only used non-written methods — mainly interviews and oral exams —, a context with high discretion and potentially large influence of evaluator bias. Starting from a gender hiring gap of almost 17 percentage points, replacing the non-written exam entirely with a blind written test increases female hiring odds 7 percentage points relative to men. The change from a very subjective to a blind, objective tool implies the most dramatic number of changes in the three factors determining hiring rates. Because the net result from changes in screening precision and tool bias may depress females hiring rates, the large estimated coefficient suggests that either written exams have higher precision or smaller disparate impact than non-written tests, or that the combined magnitude of these channels is small relative to the size of disparate treatment in interviews.

A second treatment type of hiring processes relying on interviews before the policy involves keeping the subjective tools, but adding a blind written exam to increase the overall objectivity of the hiring process. Despite the possibility of introducing disparate impact from a tool such as standardized testing, I estimate an increase in women's hiring rates relative to men's of about 5.9 percentage points, or 35% of the initial gap. By introducing an additional screening stage, employers increase screening precision, which helps minority candidates both directly and by diluting the contribution from the interview signal, which is still influenced by group-based priors and disparate treatment.

These changes to screening methods are all successful in increasing female hiring rates. More importantly, two of them — blinding a pre-existing written exam and introducing a blind written test in a process with an interview — can only maintain or increase the average productivity of hired employees. The lack of a equity-efficiency trade-off in these cases is intuitive: redesigning these practices implies reducing biased procedures that kept equally qualified minority applicants below the hiring threshold.

The last set of comparisons illustrates the potential role of well-intentioned changes in screening practices that have no effects toward greater diversity. Employers who employed both written and non-written tests and both removed interviews from the screening process and blinded the written stage saw no changes in female hiring rates relative to men. In this case, even though the employer completely removed evaluator bias from the process, the loss in screening precision depressed female hiring rates or, potentially, removing interviews helped men if women were favored by the screening tool. The other counterfactual in this case, which keeps the interview stage and only blinds the written exam also produces no discernible effect on female hiring rates.

There are several lessons from analyzing the different treatment groups. First, gender disparities in hiring come both from screening practices — either by their differences in precision or the existence of disparate impact — and decision makers. Second, decision makers matter even in instances when the tools being employed provide relatively objective signal and limit bias expression. Third, concealing candidate identity in an existing test benefits the less-favored group unambiguously without introducing implying in efficiency loss. However, blinding alone may not be enough to improve gender diversity. Fourth, introducing blind tests helps even further, as long as bias-free screening precision gains offset potential majority-favoring biases in other stages of the job process. On the other hand, removing subjective tests and blinding exams fall short of improving women's outcomes, suggesting that employers should carefully weigh precision loss and net gains from bias reduction.

In the final part of the paper, I study a complementary approach to improving gender equity in hiring, changing decision makers. While changing screening tools to limit discretion is an intuitive idea and, as my results show, can lead to significant advances toward diversity, redesigning the mix of evaluation stages may be complicated in practice if employers have limited information on relative disparate impact and precision between written and non-written exams. Indeed, some changes in screening tools that appear reasonable, might be ineffective, as when employers remove interviews without introducing another less biased stage.

Even if employers keep the same screening design, they can still limit bias expression by evaluators. Rather than focusing on de-biasing or other training methods, I expand on the idea that hiring managers face an increasing cost when expressing bias in more objective tools by incorporating a penalty function that depends on the hiring committee composition. The analysis follows the same logic from a series of corporate and public policies incentivizing or enforcing more diverse committees.[4]

Exploiting more recent data from Brazil's public sector job processes, I leverage information on candidate scores by exam type and hiring committee member to first study how changes in the gender composition of committees affect female and male candidates and hiring managers. Even though most job processes include a mix of blind-written and non-written tools, women have a slightly lower final score and hiring probability than men. Decomposing the final score into each evaluation round reveals that female candidates receive identical scores to resumes and blind exams, but are scored on average 4 percentage points less than men in non-written exams.

To separate out the confounding effect of individual differences in skills between the two examination types (or disparate impact from the tools) and disparate treatment in interviews and oral tests, I compare the difference between non-written and blind written scores of the same candidate across job processes with different committee gender compositions. Estimated effects show that the non-written penalty for female candidates decreases when there are more women in the committee, as well as their final scores and chances of job offers.

To better understand the forces driving less biased evaluations to female candidates when the hiring committee has more female members, I analyze how the the same female evaluator scores female candidates differently when she participates in hiring committees with different shares of female colleagues. I find little evidence of any change in behavior from female committee members, who only marginally improve non-written scores relative to blind exams they give to female applicants. On the other hand, male evaluators increase their non-written scores given to women relative to men when they have more female colleagues in the hiring committee. As expected, this effect does not appear in blind exam scores, and imply in a decrease in evaluator bias of about 1.4 percentage points.

One possible interpretation is that by adding more minority members to a group, two effects contribute to the change in men's behavior. The first is that even though hiring members evaluate each candidate independently, they are allowed (and often do) to share their opinions on candidates' performance, potentially changing their scores before final submission. While this process plays no role in objective measures like written tests, because interviews or oral

[4]For example, Norway passed a law in 2006 which mandated a gender quota of 40% on corporate boards, with other European countries following lead (Bertrand et al. (2019)).

examinations are usually conducted with all hiring members present, this can change behavior and perception during the evaluation and after, when hiring members share opinions about candidates. If members of the majority group belief other majority members are more likely to hold the same preferences and display similar behavior over the minority group, having members of the minority group observing their decisions may impose a censoring effect.

This paper makes several contributions to the literature. First and most closely related to my work, a body of observational studies examines how screening practices impact equity outcomes. One line of this research has focused on the effects of hiding candidates' identity, starting with the important work by Goldin and Rouse (2000) who show that blind auditions in orchestras increase the likelihood that women musicians are hired. More recent papers include the study of anonymized CVs (Behaghel et al. (2015), Krause et al. (2012), Åslund and Skans (2012)).[5] Other strand has looked at how introducing testing in low-skill jobs helps minorities (Autor and Scarborough (2008) and Hoffman et al. (2018)).[6]

My setting allows me to have a uniquely detailed analysis of screening practices and decisions with plausibly exogenous source of variation in the use of different practices. Screening in Brazil's public sector also involves multiple occupations across all skill levels, most of them with identical or similar counterparties in the private sector, and have job processes structured in a similar way as in other sectors. More importantly, different combinations of tools enable me to separate the effects of screening precision, evaluator and tool bias which require different policy responses to address low diversity.

Second, this paper adds to the large literature investigating discrimination in hiring using audit and correspondence (AC) studies (Neumark et al. (1996), Bertrand and Mullainathan (2004), Kline et al. (2021)) or natural experiments (Goldin and Rouse (2000)). A common feature across these papers is the impossibility of observing hiring practices, manager behavior, and step-by-step results within job processes. Employee selection behavior is effectively a black-box, and the contribution of screening methods or decision makers to observed racial or gender gaps is unknown. By tracking candidate and evaluator results for each screening stage and tool, my paper provides novel evidence that can instruct employers and policymakers on

[5] Several other papers have studied various consequences of partially concealing or including some information about job applicants on labor market outcomes, e.g., age (Neumark (2021)), credit information (Bartik and Nelson (2021)), criminal records and history checks (Holzer et al. (2006), Agan and Starr (2018), Doleac and Hansen (2020)), and drug testing (Wozniak (2015)).

[6] In experimental evidence, Bohnet et al. (2016) examine joint vs. separate evaluation of candidates and find that evaluators are more likely to use gender stereotypes when evaluating one candidate separately, and more likely to base their decisions on individual performance in joint evaluations, as decision makers have more information to update their possibly biased beliefs.

how to achieve gender equity in hiring. I find that biased hiring managers generate disparate treatment even when the evaluation measure is relatively objective.[7]

A second drawback from experimental studies in discrimination is that they usually can only measure call back rates in the early stage of screening processes instead of the final hiring decision. While audit and correspondence studies have generally found racial gaps at this first screening stage, conclusive evidence on gender disparities is much less apparent (Bertrand and Duflo (2017)), with studies leading to conflicting results (e.g., Kessler et al. (2019), Booth and Leigh (2010)). Even if the range of estimated callback gender differentials provided consensus, the measure may not even predict systematic biases in the hiring decision stage (Cahuc et al. (2019)). My estimates capture gender disparities at the various stages of hiring processes both for performance evaluation and job offers, providing a complete picture of disparities in hiring.[8]

My results also shed light on the role and contribution of decision makers in generating gender disparities and how to design hiring committees and implement screening tools that curb evaluators' bias expression. These results contribute to the existing empirical evidence on the importance of evaluator's gender available from specific settings (Broder (1993)) and Card et al. (2019) for grant and journal reviewers, academic promotion committees in Italy and Spain (De Paola and Scoppa (2015), Bagues et al. (2017), public examinations of the Spanish judiciary (Bagues and Esteve-Volart (2010)), and teachers (Lavy (2008), Breda and Ly (2015))), which have offered results ranging from the gender of who evaluates having no effect on candidate results, to female and male evaluators judging women more harshly or less harshly.

This paper finds that hiring manager behavior interacts with the choice of screening tools, as practices with higher degrees of discretion allow for differential degrees of bias expression. When considering hiring committees where evaluation decisions are made individually, but members engage in discussions and their decisions are made publicly, having a similar gender ratio in the committee is likely to balance potential bias in male evaluators.

[7] This is in line with Sarsons (2019), who finds that referring physicians judge surgeons with the same objective performance record differently depending on their gender.

[8] A more subtle point with respect to why AC-type studies may fail to detect gender discrimination in hiring is that the set of employers available to researchers to send resumes tends to be limited to low-skill, entry level jobs. These jobs are more likely to accept email or online applications and offer less chances of detection of fictitious resumes (Neumark (2018)). However, because these occupations are commonly female oriented, female candidates might be preferred to otherwise identical men. Moreover, without measurements of higher-level jobs, where women tend to be under-represented compared to entry-level occupations, it is difficult to trace some of these estimates to the economy more generally, or to connect callback gender gaps to hiring or wage gaps. Further issues include that control attributes in resumes might signal characteristics other than race or gender (Guryan and Charles (2013)) and that detected group disparities measure average, not marginal discrimination (Heckman (1998), Neumark (2018)) — in reality minorities do not apply to jobs in a random way.

Fourth, this paper also relates to the growing literature on personnel economics of state (Finan et al. (2017)) that has studied how governments can change the applicant pool, and the characteristics of individuals that are an important determinant of performance in the public sector (e.g. Dal Bó et al. (2013), Deserranno (2019)). However, given the large public sector premium in many countries and that most government jobs tend to be over-subscribed (Finan et al. (2017)), the type of employees who are hired will ultimately depend on how candidates are chosen, since inadequate screening procedures can undo positive selection.[9]

Fifth, this paper 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., media slant (Gentzkow and Shapiro (2010)), policy uncertainty (Baker et al. (2016))). In many applications, however, researchers are interested in extracting actual structured data from text, a task 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 messages containing numeric data, without being constrained by the shape of raw text. This query-based approach offers another text analysis tool new methods being used in economics, like the method from Shen et al. (2021), which exploits structure patterns to identify text regions in complex layouts.

2 Institutional Details & Setting

2.1 Overview

Brazil's public sector is a vital part of the country's economy. Federal, state, and local governments employ about 13% of the Brazilian workforce, a similar share to OECD countries, including the US. Brazil's government offers an expansive array of services, from universal healthcare to free pre-K to 12 and college education, controls thousands of state-owned enterprises and agencies from oil exploration to banking services, among many others. The hiring stage of public servant selection in the country is particularly important, as public sector employees receive automatic life-time tenure after being hired and termination is only possible vis-à-vis serious misconduct provisioned in a narrow set of rules, such as peculate or other forms of corruption, lobbying, and post abandonment. Wages are fixed and generally compound on a time-in-office basis and mechanically by inflation.

[9]Some papers have studied how patronage affects allocation of public sector positions (Xu (2018), Colonnelli et al. (2020), Brollo et al. (2017)), and the effects of civil service reforms transitioning from discretionary appointments to meritocratic systems (Estrada (2019), Moreira and Pérez (2021), Moreira and Pérez (2199)). This paper examines how changing screening methods within a meritocratic system affects labor market outcomes.

2.2 Public Servant Selection

Brazil was the first country in Latin America to establish a formal, merit-based career civil service, considered a primary example of meritocratic and legally professionalized civil service system (Grindle (2012) and Figure (A.4) for the complete history of meritocracy implementation and public servant selection rules). Over 70% of public sector jobs are allocated through a mandatory legal device known as "*Concurso Público*" (Public Tender), a highly competitive and structured process, referred by Brazilians simply as *Concurso*. The entire *Concurso* must be conducted and reported transparently, with every step of the process recorded and published in a designated daily government gazette (similar to the Federal Register).[10]

Every job selection process follows the same general steps depicted in Figure (1). The first posting regarding a hiring process — i.e., the job announcement — is called *Edital de Concurso*. This is a legally-binding set of rules that must describe in detail all pertinent information about the job posting, how the hiring steps are organized and conducted, the composition of the hiring committee, as well as other rules and guidance. Specific job announcement details are job and employer dependent, potentially varying within the same employer. However, both every job process must follow general guidelines prescribed in the Constitution and must integrally respect the rules laid out in the job announcement.[11]

The same *Concurso* may aim to hire multiple applicants for one job title and opening, multiple openings or job titles, for the same or distinct locations. When job announcements are posted and whether an employer conducts multiple separate hiring processes to fill out open positions or only one broad *Concurso* are determined by a complex bureaucratic process. This process requests that the government employer manifests intent in filling out or expanding specific job titles to the appropriate oversight budget and comptroller offices, which then may or not greenlight the job posting.

The hiring process then proceeds as follows. Candidates apply to the job opening, have their applications screened based on announced requirements (e.g., be a Brazilian citizen, have a valid medical license, 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, with information on full names and often some personal identification such as date of birth, individual taxpayer identifier, or identity card number. The authority organizing the hiring process then publishes individual performance/scores on each selection stage as the hiring

[10] Some public sector jobs are exempt to the formal civil service selection procedure, including temporary jobs, positions of trust, and commissioned posts. These jobs are particularly common in occupations closely related to politicians like congressional staff.

[11] Because wages are fixed and determined by law, job announcements always detail the entry wage and benefits, skill required for a candidate's application to be officially accepted in the hiring process, hours worked etc.

process unfolds, such as interviews or tests, identifying candidates that ultimately are offered jobs, wait-listed, and hired.

2.3 External Validity

Brazil's public sector job processes provide a relevant laboratory to study hiring decisions and practices used well beyond public sector. *Concursos* employ a combination of screening tools widely used in the private sector. Exam types can be divided into written or multiple choice tests, resume analysis, interview, oral examinations, and practical evaluation. When job processes include the use of several screening methods, initial stages usually use tests and written exams of general or specific knowledge (e.g., math, language, laws) to filter out candidates, with more subjective methods being applied at later evaluation stages. Practical exams commonly test job-related skills, such as typing speed and accuracy for secretaries, foreign language conversations for translators, circuit driving for drivers, teaching presentations for teachers and professors etc.[12]

Because I observe the universe of public sector hirings, the occupations and skill distribution of these job selection processes also offers a direct comparison to identical or similar occupations outside public administration. In practice, public sector employers compete for candidate pools with the private sector in most occupations: cleaning personnel, janitors, lawyers, accountants, teachers, police or private security services, doctors, secretaries, telephonists.

The structure and characteristics of public servant recruitment and Brazil also shares similarities with public sector hiring of several other countries, including France, Italy, Spain, and India, in addition to organizations, like the European Central Bank (Hospido et al. (2019)). Generally, applicants to public sector positions in these countries are also subject to competitive exams with rigid rules, and government jobs offer similar amenities (life-time tenure, fixed pay and career progression structure).

More generally, because hiring managers know that the output of their decision making process is publicly available, one could assume that the expression of intentional bias is less likely relative to settings where hiring practices and decisions are privately observed only by the firm. On the other hand, since firing or forced transfers in the public sector are extremely difficult, hiring managers may particularly care about highly subjective traits, such as "culture fit".

[12] Before that, under-qualified candidates (defined as having less schooling or credentials as indicated in the job announcement) have their application requests immediately rejected.

2.4 Setting Advantages & Limitations

in terms of measuring the tools and people. Several features in the organization and implementation of hiring processes in Brazil's public sector make it an almost ideal setting to study the role that different screening tools and who screens. First, committee members usually have no say in choosing which tools are used in a selection process, since the screening method mix is determined at a higher organizational level or by public technicians and legal specialists, usually in line with strong historical and customary use depending on the occupation.[13] Moreover, because the type, order, and weight of each exam is described in the *Concurso*'s job announcement — the *Edital* —, these are all legally binding rules that if not enforced, result in the invalidation of the entire process or any potentially resulting hires.[14]

As a consequence of the structured and quasi-exogenous nature of selection processes, evaluators cannot ignore or disregard scores, with the final job offer decision being entirely determined by candidate ranking, based on exam scores, and the number of job openings.[15] These features provide me a clean setting to measure both the effects of screening tools and who screens on applicant outcomes.

Unfortunately, every observational study is subject to a data missing or incomplete information problem. Some of the historical scanned government gazettes documents were severely damaged and unreadable through any optical character recognition (OCR). In other cases, part of a job processes information was missing due to water marks or stains, in which case I drop the job process altogether. Another inconvenience related to older documents is that most states do not maintain online repositories with long time series of documents, even if these documents are stored in government libraries and official agencies.

3 Opening the Hiring Black Box: Data Extraction

3.1 Raw Text Sources

The raw data used in this paper come from dozens of millions of official journal pages of federal, state, and local governments in Brazil (known as "*Diário Oficial*") from 1970 until 2020. These gazettes are similar to the Federal Register in the US and publish the universe of public

[13]For example, oral examinations have been used to select judges, prosecutors, public defendants, and other judiciary members for centuries, class-style presentations are extremely common to screen teachers and professors as well as written exams for auditors and analysts.

[14]In case the announcement contained minor errors or typos, or certain non-fundamental information changes (e.g., the specific room where examination take place), the publishing agency must post a new notice in the government gazette communicating possible corrections.

[15]I test for the empirical validity of hiring probability conditional on final scores in Figure (4).

notices spanning public procurement processes, executive orders, and information on public servants. Such notices on public sector personnel include the entirety of every public sector employee hiring process (as shown in Figure (1)) and professional relevant events of current employees (e.g., promotions, licenses, sanctions). Every government branch maintains its own decentralized repository with daily scanned issues of official journals, which I first scrape and retrieve in order to assemble a dataset with specific government-level journals over time. Table (A.1) shows a complete list of the separate government entities used to retrieve the government gazettes, as well as when issues first become available online.

The next — and most challenging — step is to extract the hiring data from these documents. To organize ideas, consider the following sequence of tasks necessary to automate the construction of a comprehensive large-scale applicant-reviewer panel:

1. *Filter out all text contained in official government documents unrelated to hiring steps*

2. *Define the boundaries of 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 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 similar nature, filtering out extraneous information that does not belong to a specific job process is also challenging. To make matters worse, different stages of the same job process have no exclusive identifier (e.g., a hiring process code) and subsequent postings rarely mention the date 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 one 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 predetermined 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.[16]

3.2 A New Approach to Transform Unstructured Text Into Data

To address the challenges above, 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 and link all different postings related to the process, and finally transform unstructured text into data. This algorithm generalizes a search query with learning and can be applied to a wide variety of empirical settings that follow the same general structure I have in this paper's data. Here, despite differences in layout and the manner in which information is displayed in the text, all relevant text belong to the same set of temporally ordered documents (i.e., government legal gazettes published daily).

The first step how to define textual matching attributes to link text snippets with highly confounding information and no connecting identifiers. The second, how to transform unstructured text into structure data, instead of using the underlying data as a signal of a latent process, as usually done in the literature.

Motivating the approach. 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.

Standard text analysis algorithms increasingly popular in economics are poor tools for connecting different text corpus based off *exact* text vectors. Even more 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 cali-

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

bration 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.

Conceptually, the problem 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.

Defining Textual Matching Attributes. 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 {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^i 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". In case there are 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. In order 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 comprise all published notices of the job selection process.

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.

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.

Transforming Unstructured Text into Structured Data. After linking hiring rounds across government gazette issues, the next challenge to leverage the richness of the Brazilian public sector hiring information 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 as in Atalay et al. (2020), every text snippet has a fairly similar structure which greatly facilitates mining.

Furthermore, 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 the message content. That is, researchers map text (raw or represented by a numerical array) onto a discrete set of measures $\mathcal{T} \to \{M_1(\tau), M_2(\tau), \cdots, M_K(\tau)\}$, 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 $\{x_1, x_2|x_1, \cdots, x_K|x_1\}$. This is usually an extremely time-intensive task, highly dependent on the particular context that relies heavily on strong prior information about the potential variations of text structure across $\mathcal{T}$. 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.

It is possible to simplify and create a generalizable procedure by taking into account the relation between several of the desired variables and one fixed variable, which I denote by x_1. Let x_1^i correspond to candidate i's exam score, x_2^i her name, x_3^i the exam type, x_4^i the committee member who gave score x_1^i and so forth. In order to deal with the unstructured nature of the text, I start by targeting text tokens containing numbers. 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. 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 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 $i \neq g$.[17]

By choosing one variable to which most or many of the other desired variables relate, I leverage the semantic structure in language that differs across public announcements, but that organizes 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 research will need to implement some form of ex-post harmonization.

The procedure above retrieves immense amounts of text snippets and data from the raw PDF files. There are over 900,000 unique texts identified by my matching attribute search A^t, of which about 110,000 were unique job processes. From these, I successfully link processes with enough information to match on and that start and end (some processes are cancelled or interrupted due to candidates' legal action). Some job processes publish the same post more than one time to give enough visibility to the public, which I further filter out. At the end, I identify 89,000 unique job processes from 1970 to 2020.

[17] 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, every 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.

4 Impact of Increasing Hiring Impartiality on Gender Equity

This section introduces my first set of results, focusing on how greater impartiality in hiring practices impacted hiring odds and application behavior of male and female candidates. I begin by discussing a 1988 reform in Brazil's Federal Constitution that introduced an impersonality requirement in public sector hiring as the main source of variation to the mix of hiring methods used by employers. The impartiality requirement was immediately adopted at the federal government level, but states only began passing the legal framework to equate their public servant selection processes to the new federal norms years later.

By comparing hiring processes of the same occupation in federal and a group of state controls, I analyze how the hiring impartiality requirement affected candidates' scores for different screening stages, gender hiring gap, and applicant behavior, distinguishing both supply and demand channels. To study the first-order reduced-form effects from the policy, I consider the treatment to be binary, in the sense that federal government employers had to promote changes to their mix of tools by removing this and that. Later in the paper I consider all different treatment types induced by the policy.

4.1 The Impartiality Reform: Description

In October 1988, Brazil passed its new Federal Constitution in the wake of the end of several decades under military regime. Policymakers sought an overhaul of civic and legal legislation previously enacted during dictatorship. The new Constitution also modified its provisions instructing how the selection of public servants via *Concurso* should occur. The new text kept all requirements introduced by the previous Constitution in the 1960s, which mandated 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), that is, meritocratic hiring, but included the following amendment: "*hiring must obey the principles of legality, impersonality, morality, transparency, and efficiency*" (1988 Constitution of Brazil, Section 3, Ch. 7, Section 1, Article 37).

These principles are poorly defined legal terms not explicitly laid out in the Constitution's text, although Brazilian jurisprudence at the time already offered interpretations for *legality* — following the letter of the law by not adopting practices explicitly stated as illegal —, *efficiency*, which meant that in order to begin a *Concurso* there should be a clear need for the hire and that the screening cost should be adequate, and *transparency*, which made it official that both job postings, screening stages, and results should be made public, a practice already in place for decades. Note that these requirements introduced by the 1988 Constitution are either maintaining previous practices or of little consequence to the screening process. With respect to *morality*, the principle has been broadly interpreted by courts and legal analysts to make it illegal for candidates or evaluators to display unethical or disloyal behavior, as cheating in screening tests, another practice previously illegal according to job announcement rules.

The most important principle in the 1988 Constitution, *impersonality*, disallowed any practices in public servant hiring that would allow for a specific candidate, or someone from an identifiable group, to gain improper advantage. In the case of written exams or multiple-choice tests, identifying a candidate's name would be a clear violation of the rule, therefore resulting in the blinding of these exams. But how to appropriately handle other screening tools was less straightforward.

Despite the apparent contradiction between conducting interviews and having a hiring process that is impersonal, non-written tests that allowed evaluators to observe and interact with candidates continued to be used in several occupations. Policymakers and government lawyers considered that some common practices were important screening tools for several public servant careers, and that as long as their use was combined with purely impartial tools, such as blind tests, they could still be used observing the other principles (e.g., interviews had to be open to the public and not closed-door). For example, it was common practice to perform oral exams in the judicial system, a practice that remained after 1988. Nonetheless, as I show later, on average, feral sector employers decreased to reliance on non-written stages, either by reducing their relative number with respect to blind practices or removing them completely.

In principle, the provisions in the new Constitution applied public sector hiring in all government levels. However, because public servant selection is conducted by states and municipalities independently of central authority, states had to pass the appropriate legal framework to comply with the new federal government rules. Compliance could be enforced either by passing specific public sector legislation or by passing a new Constitution, similar to the federal government's decision in 1988. In reality, the same reason that prompted Brazil's federal government to pass a new Constitution — the exit from a military regime and return to democracy — imposed the need on other federation entities to also introduce their own updated constitutions. This led to a delay of several years for the sharp shift in federal employer behavior with respect to hiring to trickle down to state agencies and governments.[18]

[18]It was common to observe states hiring for several occupations only using interviews (which complied with the previous constitution requirements as these were personality and character "exams", while out of thousands of job processes at the federal level post policy, I found no occurrences of hiring based solely on interviews. More importantly, it was common to find lengthy discussion pieces in federal gazettes on how federal agencies were adjusting their hiring processes and practices to comply with the impartiality and other guidelines in the Constitution.

4.2 Contemporaneous Changes & Threats to Identification

Among many other changes, the 1988 Constitution re-organized political constituencies, reinstated popular vote for the executive branch, and ended media censorship that was instated during military regime. The Constitution also expanded the bill of rights and public services, most of which took several years before being offered to the population. Although these changes affected civil society and the political landscape, the only changes to public sector hiring were the new principles that I discussed.

Around 1990, gender attitudes in Brazil were broadly in line with Chile, South Korea, and China, but more egalitarian than India. When asked whether they agreed withe the statement "*when jobs are scarce, men have more of a right to a job than women*", 25% of US men and women would say yes (Figure (A.1)), and about 37% of Brazilians. By 2010, gender attitudes in the South American country had improved considerably, reaching the same numbers as the US in 1990 and several points ahead of South Korea, China, India, and Egypt. The country's female share of the labor force followed a similar convergence path to developed nations during the period, but around the impartiality reform stood at two-thirds of the US rate (Figure (A.2)).

4.3 Sample Selection & Data Patterns

For the analysis centering on the impartiality reform policy, I restrict my estimating sample to the years 1986 through 1991 and to the states with official gazette issues available online for the period: 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, in addition to the federal government. 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.

I center the analysis around the 1986-1991 period since states began jointly passing new state-level Constitutions with similar guidelines to the Federal rules at the end of 1990. 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 still hiring solely based on interviews, for example. Figure (2) shows the gender distribution of applicants by occupation and skill level.

4.4 First Stage: Did the Reform Change Screening Practices?

Due to the nature of the shock to hiring practices I study, it is crucial to begin main empirical analysis by evaluating whether and to which extent the introduction of the impartiality requirement led to a reaction from federal employers relative to untreated hiring processes. I test for a series of different take-up or compliance measures in federal jobs relative to states by running:

$$y_{ct} = \delta_{o(c)} + \alpha \text{Post}_{o(c),t} + \gamma_t + u_{ct} \tag{1}$$

where outcomes y_{ct} for a job process c of occupation o are regressed the on variable of interest, $\text{Post}_{o(c),t}$, which takes the value of one if the job process is conducted by the federal government after 1988. Comparing similar occupations between treated and control groups is important to net out composition differences between aggregate jobs posted at different government levels. In Brazil, healthcare services are usually provisioned at the local and state levels, while bureaucracy tends to be concentrated in the federal sector (e.g., tax compliance offices).

Table (1) shows "first-stage" results given by equation (1). Columns (1) and (2) test how likely treated job processes are of having at least one written round after the policy (which then become blind exams) relative to state job processes used as control. To gauge the importance of composition effects, the first column only uses year fixed effects and compares all occupations, with a precisely estimated coefficient of zero. After controlling for occupation in column (2), the coefficient becomes large and statistically significant, indicating that treated jobs become 25 percentage points more likely to have at least one written stage as part of the screening process.

Columns (3) through (4) conduct a similar exercise, but now testing whether the impartiality reform induced treated employers to reduce the probability of having at least one *non*-written exam. Conditional on occupation, column (4) shows a negative but imprecisely estimated effect. Finally, column (5) finds that treated job processes were 48 percentage points more likely to use a unique screening tool, comprised by a written (blind) exam, and column (6) shows a 25 percentage-point decrease in the probability that a job process uses only non-written screening methods.

First-stage estimates indicate sharp changes in the mix of screening tools used in federal sector hiring processes relative to processes in the control group. Although I discuss in detail the different treatment groups giving rise to each estimated effect in Table (1) later in the paper, it is useful to shed some light on what underlying responses each of these estimates capture. For example, maintaining all rounds as non-written in a federal job process would be a direct violation to the principle of impersonality (column (6)). Similarly, to increase impartiality, employers previously using a mix of written and non-written tools might remove subjective stages and blind the written stage (columns (4) and (5)). Instead of removing non-written stages, employers could add a written blind round (1).

These different combinations of changes in screening methods toward more impartiality contribute to a non-perfect compliance rate in each individual regression. Taken together, these estimated effects all represent policy-compliant changes, largely determined by occupation, as I show later in the paper.[19]

4.5 Binary Difference-in-Differences Design

My empirical strategy exploits the immediate compliance of federal government employers with the introduction of impartiality in public servant selection, but a lagged and slow adoption by state-level employers, to assess the effects of the policy on gender gaps in several labor market outcomes:

$$y_{it} = \delta_{o(i)} + \beta \left(\text{Post}_{o(i),t} \times \text{Female}_i\right) + \gamma_t + u_{it} \tag{2}$$

where y_{it} represents candidate i's job process outcomes, $\text{Post}_{o(i),t} \times \text{Female}_i$ measures the differential effect of greater hiring impartiality on women relative to men, while controlling for year and job announcement's occupation fixed effects. In all specifications, standard errors are clustered at the job process level. I only consider job processes with at least one male and female applicants, with known job offers, and that consistently appear before and after the policy in both groups. I assign candidates' gender using Brazil's Census Bureau Gender of Names database, which contains nearly 200,000 unique first names and their correspondent gender. The match precision is above 98%.

4.6 Effect on Candidate Scores

An advantage of having Brazil's public sector as a setting is that, in addition to observing job offers to candidates, I have detailed performance scores from each screening stage. Assuming more impartiality is women-favoring, depending on the magnitude of the effect, women's final scores may increase and yet hiring gaps remain unchanged if the marginally not hired female candidate was too far behind the marginal hired man in measured performance. Therefore, with a less coarse outcome such as scores, more subtle effects of the policy can be measured.

Before using final scores, however, I check whether they actually determine job offers. Figure (4) compares hiring odds across the distribution of final score results within a job process (i.e., the final ranking of candidates determined by sorting highest to lowest final scores). Only candidates in the highest score decile in each job process have a non-zero probability of being hired, with top scorers having about 60% chances of receiving an offer. This is not surprising — according to the rules, hiring decisions are made exclusively in accordance to the ordering of

[19] Figure (A.5) shows an enforcement example of blind exams in a selection process for federal judges published on September 4, 1989 in the job announcement rules (*Edital*). The rule states that candidates identifying themselves in any exam (written or multiple-choice) will be excluded from the hiring process.

candidates' final results, and even top scorers are not guaranteed job offers since job openings are generally fixed.

Table (2) begins by comparing final scores in the hiring process received by female and male candidates. Scores are standardized within each hiring committee, so that they are comparable across different job processes. Women's final score increases 0.07 standard deviations after the policy, with men's final scores decreasing by a slightly larger magnitude. Combined, these effects imply a 0.14 standard deviation narrowing of the gender score gap. These separate effects by treatment and control are shown in Figure (3), where final score gender gaps remain unchanged for candidates in state job processes and the gap significantly narrows for federal jobs. Figure (5) displays dynamic effect versions of the pooled estimate in column (3), first comparing the evolution of the evaluation score gender gap in federal and state hiring processes and then plotting the difference-in-differences estimate of the two series.

To lend further credibility to the change in final scores as a consequence of the reform, note that depending on the mix of screening methods used in each job process, the final score is determined by some weighted average of these tools. As shown in Table (1), albeit occupations complied with the impartiality requirement in different ways, one would expect the increase in the final score to be driven on the intensive margin by an increase in the score of written exams of women relative to men. As I show in my conceptual framework in Section (6), this expected increase in relative score is attributed to the elimination of evaluator bias, or disparate treatment, after blinding written exams.

Column (4) of Table (2) shows that men's written scores decrease by about 0.10 standard deviations, contributing to an overall increase in women's written scores once exams are blinded relative to men of 0.13, almost the entire magnitude of the improvement in final scores. One interpretation of the decrease in men's scores is that prior to conceiving candidates' identity in written exams, men were being over-scored. Next, absent substitution effects — i.e., evaluators strategically adjusting scores in non-written exams as a response to blind written texts —, changes in relative scores of non-written should be close to zero or at least small in magnitude. This is confirmed in columns (7) through (9).

4.7 Effect on Hiring

Having determined that the impartiality reform increased women's final scores relative to men, combined with the fact from Figure (4) that final scores determine hiring offers, my next analysis answers whether the performance improvement was sufficient to increase women's hiring rates. I run regression (2) with a dummy for whether the candidate received a job offer as the outcome. Recall that these job offers represent the official conclusion of the *Concurso*, in which part of the candidate pool that ranks above a final score threshold (when it exists) is considered "adept", that is, could legally be hired, and the number of top candidates matching the number of job openings is offered the job offer, known as *convocação*. When candidates decline or cannot accept the job offer (e.g., because of death), the next enabled candidate outside the initial list receives the offer.

Columns (1) and (2) in Table (3) show that the probability of being hired for women and men, respectively, go opposite directions after the impartiality policy takes effect. Women become 0.3 percentage points more likely to be hired and men's hiring rates decrease by 0.4 p.p. Interpreting these coefficients in light of the variation used the empirical strategy, consider the following example. A woman (man) applying to an accountant job in the federal government is more (less) likely to be hired after the policy compared to a woman (man) who applied to another accountant job in a state. Taken together, these hiring probability estimates imply a 0.7 percentage-point decrease in the gender hiring gap on average. Thus, the policy made women more competitive candidates because of higher final scores, and the improvement in performance was sufficient to result in higher hiring rates.

4.8 Gender Hiring Gap: Disentangling Supply and Demand

The gender hiring gap is determined by a sequence of decisions of both job seekers and employers. First, potential candidates decide whether to apply, and second, conditional on being an applicant, there is some probability of getting a job offer and being hired. Systematic differences at these stages between genders in turn determine the broader hiring gap. My previous estimates focused on the second factor, which is often unobservable in other settings, since calculating the conditional hiring probability requires observing *all* applicants, not only the hired pool.

Knowledge on the applicant pool is important for a complementary reason. Employers and policymakers may also be interested in the initial individual decision of whether to apply to a job or not. Intuitively, drawing more candidates from a minority pool should increase the overall hiring rate of that group if qualified individuals refrain from applying. With respect to gender, previous studies have presented evidence on several fronts suggesting that women may be less likely to apply to promotions and less likely to enter tournaments than men 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. (2021)), sort into female environments to avoid competing against men (Gneezy et al. (2003)), or even being nudged to apply when job ads indicate a preference for diversity (Flory et al. (2021)).

Employers' choice of screening tools is likely to interact with the factors above, further imposing barriers to extensive-margin responses of female candidates. Implementing more impartial screening may motivate more female candidates to apply, as even the perception of fairer treatment can be consequential in shaping minority's behavior (Small and Pager (2020)).[20] Supply-side factors are important because suboptimal entry by high-performing women are costly to firms and without observing application rates, the effectiveness of enforcement of anti-discrimination laws is thwarted.

To understand this point more formally, let the share of women hired by an employer or from a job selection process be defined as $\mathbf{Pr}(\text{Female}|Hired = 1)$, in which researchers observe the pool of hired candidates and then calculate the makeup of female hires. Observing low hiring rates for women in this case masks two different effects: the potential propensity of the employer or hiring committee to discriminate (under a set of assumptions) against women and lower quality or fewer women applying for the job. Because researchers can usually only observe the rate at which women are hired, measured hiring rates are conditioned on an endogenous variable — the hiring decision based on the available candidate pool — and therefore cannot distinguish between employer and applicant behavior.

When policymakers enforcing gender discrimination laws rely on observed hiring rates, non-discriminatory employers may be inadvertently punished when observed gaps are driven by differential gender sorting across employers or other supply-side factors. Brazil's public sector hiring processes enable me to decompose the female hiring rate into two components: a demand channel, capturing differences in the odds of female candidates winning and a supply channel, which measures differences in application rates as:

$$\mathbf{Pr}(\text{Female}|Hired = 1) = \underbrace{\mathbf{Pr}(Hired|\text{Female} = 1)}_{\text{Demand}} \times \underbrace{\mathbf{Pr}(\text{Female} = 1)}_{\text{Supply}}$$

The demand component, $\mathbf{Pr}(Hired|\text{Female} = 1)$, which was estimated in column (1) of Table (3), indicates an employer-driven response improvement in female candidates being hired of 0.3 percentage points, which coupled with a similar-sized decrease in hiring odds to men represents a 0.7 percentage-point decrease in the overall hiring gap. Column (4) in Table (3) estimates the supply response of women applying to jobs after the impartiality policy ($\mathbf{Pr}(\text{Female} = 1)$). The estimate shows that women application rates grew by 1 percentage point. To benchmark these magnitudes, consider that the hiring gap for federal jobs pre-policy was about 1.6 percentage points (net of occupation effects), the drop in the hiring gap implied

[20] Brands and Fernandez-Mateo (2017) argue that previous experiences with an employer's practices leads women to place greater weight than men on fair treatment and negatively affects their perceptions of fairness of treatment.

by the demand channel corresponds to about 44% of the pre-treatment level, while the supply effect measured as gender application gap amounts to around 62%.

To gain further insight into the general forces behind supply movements, in the results in Table (4), I run job process level versions of the binary DiD model, first confirming that the share of women hired grows after the impartiality requirement, as well as the share of female candidates in the applicant pool. Note that these specifications are equivalent to observing aggregate data on $\mathbf{Pr}(\text{Female}|\mathit{Hired} = 1)$ in column (1) and $\mathbf{Pr}(\text{Female} = 1)$ in column (2). Moreover, note that latter effect — about 6% growth in the female application rate — is statistically indistinguishable from the estimated in column (4) of Table (3), around 7% of the pre-treatment level.

Next, column (3) shows that the number of applicants decreases after the policy — a movement that may be driven both by candidates' perception of how costly the job process may be, for example because of more screening rounds, or based on the job openings, which positively correlate with candidate pool and are determined by budgetary and personnel management —, but the reduction in the number of men applying is 10 percentage points larger than women's, accounting for the increase in the probability of a candidate being female.

5 Conceptual Framework

This section sets up a theoretical framework to explore the impact of introducing and removing different screening practices on hiring rates, and builds on the canonical models of statistical discrimination by Phelps (1972), Aigner and Cain (1977), with important modifications by Autor and Scarborough (2008). I model managers and screening practices allowing for them to manifest several dimensions employers face when designing selection processes.

The first ingredient in the model is hiring manager bias. Evaluators have the task of selecting employees with a mix of screening tools delegated by the employer. I allow managers to have a systematic bias for certain demographic group. The term could be interpreted as taste-based discrimination, as it effectively captures utility disamenity from hiring some group, implicit or any other source of unintentional bias. But their bias can only be expressed to the extent that the screening tool used for screening allows for discretion. The intuition is simple: expressing bias vis-à-vis an objective test is more costly than evaluating a candidate in an interview because detection is easier. Because managers base their prior on candidates' productivity based on their group membership, managers always statistically discriminate, unless the tool used prevents managers from observing candidates' identity.

The second addition allows screening tools to be biased. Independently of hiring manager's behavior, certain screening practices may disadvantage a particular group. Finally, by maintaining screening precision in the model, the role of tool bias is equivalent to adding noise to the productivity signal provided to managers, favoring less productive applicants of the favored group.

With these basic forces — manager bias, tool bias, and precision — interacting, my goal is to derive reduced form predictions of gender hiring gaps for five cases of changes in screening tools I empirically observe. In addition to being interesting in their own right, these cases will reveal the relative importance of tools and managers toward gender equity.

5.1 Environment

An employer (the principal) delegates the screening of a pool of job applicants to a some number of hiring managers or evaluators. The candidate pool comprises individuals from two demographic groups, $x = \{m, f\}$, corresponding to a minority and majority group, female and male, respectively. As usual, I use the term minority in its socio-economic dimension, so that for now the gender make-up of the candidate pool is unrestricted. The employer bases the hiring decision on some indicators of productivity $\theta = \{s, \eta\}$, observable only by hiring evaluators, which coarsely 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.[21]

The employer's objective is to hire a proportion K of workers that maximize expected productivity. But evaluators' objectives are imperfectly aligned with those of the firm. Evaluators care both about productivity and their bias toward a group, which must be jointly maximized when hiring job applicants by

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

where π_j is evaluator j's bias, c_θ is a cost function disciplined by the usual properties and defined over $c \in [0, 1]$. This component captures the cost evaluators face by expressing bias, i.e., informing a value of a candidate's measured performance that differs from the signal provided by the screening tool. Intuitively, this cost increases in the objectivity of the screening signal.

[21] Aigner and Cain (1977), Lundberg and Startz (1983), Cornell and Welch (1996), and Bartik and Nelson (2021) model signal precision depending on group membership. Similar to Autor and Scarborough (2008), I assume it to be independent of group membership to focus the analysis on the new features I introduce in the model.

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 friendly or an "appropriate fit".

The cost of expressing bias plays a central role in the model. The term connects an intrinsic property of a screening tool — which I call d_θ — to represent a screening practice's degree of discretion (or subjectivity), which loads on the bias term and determines its relative role in the manager's utility. Later, I impose additional structure on c_θ where the cost of being caught will depend not only on how much discretion is granted to the manager by the tool, but also on the composition of the hiring committee along x.

To keep the notation tractable and match the model predictions to our empirical setting, consider the screening tool choices available to employers before and after the impartiality reform in Brazil's public sector. The full choice set and why the following are the relevant cases are discussed in Section (6). Before the reform, 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.[22] After the policy, employers in the federal government are constrained to screen candidates using only a blind written test or a combination of non-written and blind written tools.

Before I begin to evaluate hiring rates under these different choices of screening practices, a final ingredient in the model is the ability of any screening tool to favor a group from x. This screening tool bias, which I call disparate impact, mean-shifts a candidate's true productivity based on that individual's group membership.[23]

6.1.1 *Before Policy: Hiring Rates With Written Exam*

Starting with the case of selecting candidates based on a written test, allowed to be biased, the distribution of written signal is:

$$
\begin{aligned}
&s^* = y + \nu_s(x) + \varepsilon_s,\ \varepsilon_s \sim N(0, 1/h_s) \\
&s^* \sim N(y + \nu_s(x), 1/h_s)
\end{aligned}
$$

[22]Within the model, I do not distinguish whether employers use one or multiple tests of the same type, as these will provide signals with identical properties the employer. 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.

[23]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 the productivity as $y = y_1 + y_2$, where in the spirit of Frankel (2021) y_1 could represent soft skills and y_2 hard skills.

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 does not depend on group membership x. $\nu_s(x)$ represents disparate impact of the screening tool, which favors men when $\nu_s(m) > \nu_s(w)$.

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

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

with the updated degree of precision $h_0 + h_s$ and the updated mean:

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

With only the written test as the screening technology, 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 threshold that yields a hiring rate of K, and is given by

$$\frac{s - \mu_0(x)}{\frac{1}{h_0} + \frac{1}{h_s}} > \underbrace{\frac{k_s - \nu_s(x) - \mu_0(x) - d_s\pi_j(x)}{\sigma_0\rho_s}}_{z_s^*(x)} \tag{3}$$

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 established by using written exams. The probability that an applicant from group x is hired is $1 - \Phi(z_s^*(x))$. Detailed derivations are shown in Appendix (**??**)

6.1.2 Before Policy: Hiring Rates With Non-Written Exam

Similarly, with only a non-written test as screening tool, the distribution of non-written signals is:

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

where $\nu_\eta(x)$ represents the possible disparate impact of non-written tests and η is the unbiased non-written signal: $\eta = y + \varepsilon_\eta$. Additionally, non-written tests also differ in the discretion allowed to evaluators, with $d_\eta > d_s$.

In this case, an applicant screened with a non-written exam is hired if

$$\frac{\eta - \mu_0(x)}{\frac{1}{h_0} + \frac{1}{h_\eta}} > \underbrace{\frac{k_\eta - \nu_\eta(x) - \mu_0(x) - d_\eta \pi_j(x)}{\sigma_0 \rho_\eta}}_{z^*_\eta(x)} \tag{4}$$

with the corresponding probability for a candidate from group x of being hired calculated as $1 - \Phi(z^*_\eta(x))$.

6.1.3 *Before Policy: Hiring Rates With Written and Non-Written Exams*

The third pre-reform possibility of screening practices is a combination of written and non-written exams. Given the two signals previously determined, η^* and s^*, and the perceived group productivity, $\mu_0(x)$, the firm updates its assessment of expected productivity in the following way:

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

with the updated degree of screening precision $h_0 + h_\eta + h_s \equiv h_T$ and the updated posterior:

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

With both written and non-written screening tools used, the hiring decision is then:

$$\frac{\mu(x, \eta, s) - \mu_0(x)}{\sigma_0 \rho_T} > \underbrace{\frac{k_T - \frac{h_s}{h_T}\nu_s(x) - \frac{h_0 + h_\eta}{h_T}\nu_\eta(x) - \pi_j(x)(d_\eta + d_s) - \mu_0(x)}{\sigma_0 \rho_T}}_{z^*_T(x)} \tag{5}$$

and the probability of a candidate being hired is $1 - \Phi(z^*_T(x))$.

6.1.4 *After Policy: Hiring Rates With Blind Written Exam*

After the impartiality reform of 1988, federal employers using written exams as screening tools had conceal candidates' identity. Within the model, blinding makes 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. Thus, the blind written signal is

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

with the same screening precision h_s and the same disparate impact $\nu_s(x)$ as the written test previously modeled. The first effective difference in the model arises in the evaluator's objective function:

$$u_j\left(y, \pi\left(x\right)\right) = y + \underbrace{\left(1 - c_b\right)}_{=0} \pi_j(x) \equiv y + \underbrace{d_b}_{=0} \pi_j(x)$$

where discretion is entirely removed from the screening tool. Additionally, blinding the written test also affects how the evaluator updates the 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:[24]

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

An applicant from group x is hired with a blind written test signal if:

$$\frac{s - \mu_0(x)}{\frac{1}{h_0} + \frac{1}{h_s}} > \underbrace{\frac{k_b - \nu_s(x) - \mu_0(x)}{\sigma_0\rho_s} - \frac{h_0\rho_s}{2h_s\sigma_0}\left(\mu_0(y) - \mu_0(x)\right)}_{z_b^*(x)} \tag{6}$$

with the corresponding probability of being hired for an applicant from group x expressed as $1 - \Phi(z_b^*(x))$.

6.1.5 *After Policy: Hiring Rates With Blind Written and Non-Written Exams*

Finally, the second type of screening combination post-reform includes blind written and non-written exams. Given the two signals, η^* and b^*, the posterior is:

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

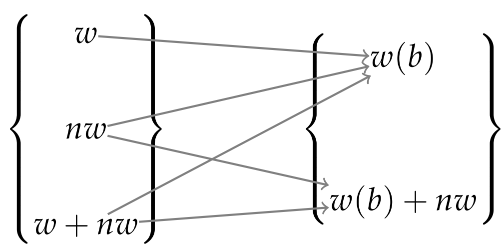
[24]For simplicity and without loss of generality, I assume that each group comprises half of the candidate pool. Another reason for using identical gender distributions is to keep application behavior from the pool of qualified workers outside the model.

with the updated mean: $\mu(x, \eta^*, b^*) = \frac{h_s s + h_\eta \eta + h_0 \mu_0(x) + \nu_s(x) h_s + \nu_\eta(x)(h_0 + h_\eta)}{h_T}$. With this combination of screening tools, given the zero discretion in the blind written test, the evaluator hires an applicant from group x if:

$$\frac{\mu(x, \eta, b) - \mu_0(x)}{\sigma_0 \rho_T} > \underbrace{\frac{k_{\eta b} - \nu_s(x)\frac{h_s}{h_T} - \nu_\eta(x)\frac{h_\eta + h_0}{h_T} - d_\eta \pi_j(x) - \mu_0(x)}{\sigma_0 \rho_T}}_{z^*_{\eta b}(x)} \tag{7}$$

5.2 Reduced Form Predictions

Each one of the previous five combinations of screening methods, before and after the impartiality reform, determines hiring rates for each group and thus the hiring gap. The screening thresholds, $z^*_\theta(x)$ and resulting hiring rates are in turn determined by functions of tool bias — disparate impact —, screening precision, and evaluator bias, as governed by discretion. Formally, denote a written exam by w, non-written as nw, and written-blind $w(b)$, the reform induced employers to change screening tools in the following ways:



Before Policy After Policy

We are therefore interested in how hiring rates for men and women change given each one the transitions above, which I study in detail below.

6.2.1 From Written to Blind-Written Exams ($w \longrightarrow w(b)$)

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? Estimating the effect of blinding a written exam measures disparate treatment by those who implement this screening tool. In Figure (6), it is clear that by eliminating discretion, evaluator bias, $\pi_j(x)$, cannot be expressed and therefore taste-based and implicit discrimination are eliminated, if they existed. More subtly, statistical discrimination is also removed since now candidates' identity and thereby group membership are concealed. Also note that in absolute terms, a screening process comprising only written tools is already likely to allow for low degrees of evaluator bias expression, as these practices offer less discretion due to being more objective. However, the bias of written exam that is independent of the evaluator, $\mu_s(x)$, remains and can still have a disparate impact on the group that it disadvantages.

By inspecting expressions (3) and (6) and considering that written tests — whether blind or not — have the same screening precision and disparate impact, since they are otherwise identical except 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[25]

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

The expression above captures with simplicity 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 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., hiring managers favor men, or if is sufficiently small due to either low discretion or low bias, the hiring rate for women increase and the hiring rate of men decrease after blinding the written exam. Alternatively, if the evaluator is biased in favor of women, blinding the exam curbs the evaluator's ability to balance women's penalty from statistically discriminating with personal bias, potentially decreasing the female hiring rate.

6.2.2 *From Written & Non-Written to Blind-Written & Non-Written Exams* $(w + nw \longrightarrow w(b) + nw)$

[25] This implicitly assumes that individual's measured performance (without bias) remains the same irrespective of the conditions of the examination. While testing this assumption is difficult, the following exercise helps understanding how it could be factored into the model. Suppose women actually perform better when a written test is blind relative to non-blind, perhaps because they feel less stereotype pressure. This would imply a different disparate impact for the blind exam — in this case, $\nu_{sb}(w) > \nu_s(x)$, which would only reinforce the effect of removing disparate treatment. A more subtle point here is how the possibility of differential performance affects whether one interprets what loads on the evaluator bias term as "discrimination". While this confounds the source of the issue — whether evaluator bias or blinding the exam — a broader view of discriminatory practices that also encompasses practices that unintentionally generate disparate impact would still prescribe blinding.

Even though this change also blinds a pre-existing written exam, here candidates' identity is still known during non-written screening stages, so that group membership information is used in the employer's posterior. As a consequence, only disparate treatment from the written exam is removed, while statistical discrimination remains. If evaluators favor male candidates, blinding the written stage in a mix of non-written tools also increases women's hiring rate, but by less than in the $w \to w(b)$ transition.

Furthermore, exams in Brazil's public sector job selection processes have pre-determined non-stochastic weights, for which I implicitly assume an equally-weighted screening tool mix. In practice, if blind written exams receive less weight by principals than prior to the policy, blinding may have limited or undetectable effects on decreasing the gender hiring gap. The same attenuation effect would happen if written exam weights were to remain the same on average before and after the policy, but contribute little to the final score.

6.2.3 *From Non-Written to Blind-Written Exams* ($nw \longrightarrow w(b)$)

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 (4), 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. On the other hand, as discussed before, screening solely based on written exams is likely to provide a setting with low disparate treatment.

Because the two screening tools involve various different parameter values, I first make the following assumptions to focus on the effects of decreasing discretion and removing group-based priors. Let written and non-written signals have the same screening precision, $h_s = h_\eta$, and the same disparate impact, $\mu_s = \mu_\eta$. It follows that the hiring threshold for men is higher with the blind-written signal compared to the non-written signal, $z^*_\eta(m) < z^*_b(m)$, as long as evaluators favor men $\pi_j(m) > 0$ or, alternatively, if it satisfies the following

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

which allows for sufficiently small evaluator bias toward women. Because this implies a higher threshold for hiring male candidates which increases selectivity for men, and given a constant total hiring rate, K, the gender hiring gap decreases.

Next, I conduct the same exercise but now allowing 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:

$$\frac{h_0\rho}{h_s\sigma_0}(\mu_0(f) - \mu_0(m)) < \frac{d_\eta(\pi_j(m) - \pi_j(f)) + (\Delta\nu_\eta - \Delta\nu_s)}{\sigma_0\rho} \tag{8}$$

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 which 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 above inequality. Alternatively, another way to interpret the inequality (8) is to rewrite it as

$$\frac{h_0\rho}{h_s\sigma_0}\mu_0(f) + \frac{d_\eta\pi_j(f)}{\sigma_0\rho} + \frac{(\nu_\eta(f) - \nu_s(f))}{\sigma_0\rho} < \frac{h_0\rho}{h_s\sigma_0}\mu_0(m) + \frac{d_\eta\pi_j(m)}{\sigma_0\rho} + \frac{(\nu_\eta(m) - \nu_s(m))}{\sigma_0\rho} \tag{9}$$

where the left-hand side represents the perceived productivity of female applicants, equal to true productivity plus bias, either from evaluator or screening tool, and 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, we can relax the assumption that screening precision of written and non-written exams are identical. If written tests have a higher precision, $h_s > h_\eta$, 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 (9) holds. On the other hand, 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 (9) is satisfied.

6.2.4 *From Non-Written to Blind-Written and Non-Written Exams* $(nw \longrightarrow w(b) + nw)$

This case maintains the use of non-written exams but in order 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 washes down the weight that the 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.

Begin by assuming 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 increases women's hiring rates if they are the group with the lower perceived productivity: $\mu_0(f) + d_\eta \pi_j(f) < \mu_0(m) + d_\eta \pi_j(m)$, due to the fact that the change in hiring probability with respect to screening precision is:

$$\frac{\partial \left[1 - \Phi(z^*_{\eta b}(x))\right]}{\partial \rho_T} = \phi(z^*_{\eta b}(x)) \left[\frac{z^*_{\eta b}(x)}{\rho_T} - \frac{\partial k_{\eta b}/\partial \rho_T}{\sigma_0 \rho_T}\right] > 0 \tag{10}$$

since $\phi(\cdot) > 0$ 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.

6.2.5 *From Written & Non-Written to Blind-Written* $(w + nw \longrightarrow w(b))$

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 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: $z_T^*(m) < z_b^*(m)$ which implies

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

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.

On the other hand, 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:

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

which 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, and how it compares to bias of the written tool. Consider the following cases.

Fist, consider the case when the written tool favors women, $\Delta\nu_s < 0$, and 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 $h_T = h_0 + h_\eta + h_S > h_s$, and as a consequence 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, thus decreasing their hiring rate.

Third, if both the written and non-written tools favor men, $\Delta\nu_s > 0, \Delta\nu_\eta > 0$, then, irrespective 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, on the other hand, 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.

6 Impact of Changes in Screening Tools on Gender Equity

While so far I have studied the introduction of the impartiality requirement under the canonical, binary difference-in-differences research design, I can leverage the fact that the policy generated multiple treatments to gain further insight into how different screening tools change women's labor market outcomes. In the previous section, I showed that different combinations of screening tools capture various levels of precision, manager bias, and tool bias, and that depending on the compounded effect of changes in the mix of practices, gender hiring gaps may diminish or worsen.

In this section, I take these multiple types of screening tool combinations and changes to the data to analyze how five different changes in screening methods affected hiring rates. I first formalize the treatment space generated by the policy, the assumptions necessary for identification, and then estimate the effects of counterfactual changes in screening methods on final scores, hiring rates, and female participation in job processes. I conduct an analysis of the differential impact by the setting in which hiring processes would otherwise be conducted. Second, there are two possible ending points: blind written, a blind written and non-written combination. Altogether, this creates six possible treatments or transitions/changes in the screening methods used.

6.1 Treatment Space

I group examination types used in job selection processes into two broad categories: *written* and *non-written*. The *written* group encompasses both actual written and multiple-choice exams. Non-written exams include oral and practical examinations, and interviews. I leave out resume analysis stages for now. Job processes may use any number of written or non-written exams, including combining screening tools from both groups, resulting in varying degrees of discretion. This broad grouping is useful as the impartiality reform should affect written exams by making their implementation blind and potentially curb the use of non-written exams, which could be considered intrinsically not impersonal.

In the last section, I studied the effects on the gender hiring gap by five different changes induced by the impartiality policy. To understand why these are the empirically relevant cases in our context, I trace out in Figure (7) the potential treatment space for job processes under the following combinations of screening tools: written (w), non-written (nw), written and non-written ($w + nw$), blind written ($w(b)$), and blind written and non-written ($w(b) + nw$). Cases shaded in gray are ruled out by assumption in a sharp DiD design (perfect compliance). Subgroups of these options would be subject to the monotonicity and exclusion restriction assumptions in the standard IV case (e.g., Kline and Walters (2016), Feller et al. (2016)).

Of the six remaining transition cases, $w \longrightarrow w(b) + nw$ is implausible under the design of the impartiality reform, since even though employers would blind the pre-existing written exam, by introducing a non-written tool, the overall level of discretion in the mix increases, therefore making it hard to claim a shift toward greater impersonality. Indeed, this combinations accounts for less than 1% of transitions in the data. We are thus left with five possible treatments, capturing the following general changes in screening practices:

1. *Only Blinding (No Change in Screening Tools):* $w \longrightarrow w(b)$ and $w + nw \longrightarrow w(b) + nw$

2. *Blinding and Replacing Screening Tools:* $nw \longrightarrow w(b)$

3. *Blinding and Adding Screening Tools:* $nw \longrightarrow w(b) + nw$

4. *Blinding and Removing Screening Tools:* $w + nw \longrightarrow w(b)$

What does each of these treatments measure? Informed by the conceptual framework of the previous session, blinding or modifying screening tools implies changes in evaluator bias (disparate treatment), screening precision, and disparate impact from different tools. For $w \longrightarrow w(b)$ and $w + nw \longrightarrow w(b) + nw$, the only change to the design of the screening process is blinding the written exam, which completely eliminates discrimination associated with evaluators for $w \longrightarrow w(b)$, and decreases disparate treatment in $w + nw \longrightarrow w(b) + nw$, but this effect may be modest toward a candidate's final score if the weight on the written exam is small. In both treatment types, screening precision and disparate impact remain the same, albeit the disparate impact in $w + nw \longrightarrow w(b) + nw$ combines the tool bias from both exam types.

Replacing a non-written exam with a blind-written test ($nw \longrightarrow w(b)$) involves the most dramatic number of changes to the forces determining hiring gap rates. First, disparate treatment of all sources is not only eliminated, but its absolute change could be sizable since one moves away from the tool with highest discretion to the case with no discretion. Screening precision may either remain the same — in the case both tools provide equally accurate productivity signals — and therefore have no impact on the hiring gap, or help women if written tests have higher precision than interviews and female candidates have lower perceived productivity. To conclude, as long as the disparate treatment from interviews favored men more than the change in disparate impact from switching the tools, the hiring gap also decreases.

Adding a blind-written exam to an interview stage ($nw \longrightarrow w(b) + nw$) improves screening precision, which raises women's hiring rates if they are the group with less perceived productivity. By adding the blind exam, there is no introduced disparate impact, and employers also reduce the reliance on evaluator bias in the interview stage since now the additional tool dilutes evaluation weight from the non-written tool. However, with a new tool, an additional disparate impact source is introduced, either increasing the potential group-favoring property of the non-written exam or attenuating it.

Finally, $w + nw \longrightarrow w(b)$ removes disparate treatment from the interview and non-written test, resulting in an hiring process free from evaluator bias other than the disparate impact from written tests, which remains the same. However, by eliminating the non-written stage, its disparate impact is also removed from the process and the total precision in the hiring tool mix decreases, which adversely impacts women (or the minority group more generally).

The exposition above reveals important sources of variation in the use of different combinations of screening tools — the strata in Figure (7) — induced by the policy's increase in hiring impartiality. Coupled with the reduced form prediction in the previous Section, this framework will inform the interpretation and unveil the forces driving the effects of changes in screening tools that I estimate next.

6.2 Assumptions & Identification

Let any treatment type D be defined over the support $\mathbb{D} = \mathbb{D}_{+} \cup \{0\}$, where job process i receives treatment (dose) D_i, with potential outcomes in period $s = \{t-1, t\}$ given by $Y_{is}(d)$. Assume further no anticipation, so that $Y_{it-1} = Y_{it-1}(0)$ and $Y_{it} = Y_{it}(D_i)$. By relaxing the binary treatment assumption, i.e., $D = 0$ or $D = 1$, we allow for any dose or treatment level $d \in \mathbb{D}_{+}$. Using the results from Callaway et al. (2021), in order to identify the average effect of treatment d among job processes experiencing the treatment with:

$$\underbrace{\mathbb{E}[Y_t(d) - Y_t(0)|D = d]}_{ATT(d|d)} = \mathbb{E}[\Delta Y_t|D = d] - \mathbb{E}[\Delta Y_t|D = 0]$$

the standard parallel trends assumption is sufficient. Given all possible treatment types induced by the impartiality reform,

$$\mathbf{g} = \left\{ \begin{array}{c} w \longrightarrow w(b) \\ w + nw \longrightarrow w(b) \\ w + nw \longrightarrow w(b) + nw \\ nw \longrightarrow w(b) \\ nw \longrightarrow w(b) + nw \end{array} \right\},$$

I estimate the following versions of the baseline DiD model for each g

$$y_{git} = \delta_{o(g,i)} + \beta_g \left(\text{Post}_{o(g,i),t} \times \text{Female}_i \right) + \gamma_t + u_{git} \tag{11}$$

in which I compare outcomes (y_{git}) for female candidates relative to men (Female_i) participating in job processes for the same occupation ($\delta_{o(g,i)}$) that had screening practices changed (g) in federal jobs but not in state-level processes $\left(\text{Post}_{o(g,i),t}\right)$.

The following example illustrates the variation used to identify $\widehat{\beta}_{g=w\longrightarrow w(b)}$. The regression in (11) compares a job selection process for, say secretaries, in the federal government that used a written exam before the policy to screen candidates, to another process selecting secretaries to state governments also only using a written test. Under the previous DiD assumptions, the parameter of interest in the example measures the causal effect of blinding the written exam in the selection of secretaries for federal jobs, using the fact that state-level job processes continued using a non-blind written exam. The interpretation of $\widehat{\beta}_{b=w\longrightarrow w(b)}$ is informed by the reduced form predictions from Section (5): the effect represents the change in the outcome as a consequence of eliminating discrimination of any source in the hiring process.

Like in the binary DiD in Section (4), including occupation fixed effects is important to net out composition effects from supply fluctuations in public sector postings, whether determined by political cycles, economic growth, personnel retirements etc. However, when taking into account the multi-valued nature of treatment, comparing job processes with a given treatment type within the same occupation becomes even more important.

But when conditioning on occupations, there is almost no variation in treatment types. That is, job processes for doctors used to employ a mix of written and non-written exams before the policy and almost every *Concurso* for medical doctors after the policy in the federal government uses a combination of blind-written and non-written exams, so that within the occupation, 92% of job processes are treated by $w + nw \longrightarrow w(b) + nw$. While there are no technical reasons for this robust pattern in the data, customary norms and centralized decisions in the federal government may explain why different employers hiring for the same occupation both employ similar screening methods and respond to the policy by adopting similar changes to the screening mix.

The strategy above identifies effects off changes in screening methods given each treatment type — it measures how increasing or decreasing levels of disparate treatment, tool bias, and screening precision given each screening mix changes the outcome. Following Callaway et al. (2021), to identify non-local treatment effects, $ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)]$, a stronger version of parallel trends is necessary. The assumption involves not only untreated potential outcomes, but also all potential outcomes under all the different treatments. That is, for all treatments d, the change in outcomes over time across all units if they had been assigned that treatment is the same as the change for all units that experienced that dose:[26]

$$\mathbb{E}[Y_t(d) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d]$$

[26]The assumption above is however weaker than assuming that all treatment groups would have experienced the same path of outcomes if they were assigned the same dose, which would imply that $ATE(d) = ATT(d|d)$. In contrast, the strong parallel trends assumption allows for some selection into a particular treatment.

While in my setting, most comparisons between estimated effects across **g** are not informative or even ill-determined (e.g., there is no plausible "counterfactual" to compare $w + nw \longrightarrow w(b) + nw$ to $nw \longrightarrow w(b)$ because the two treatments have different starting points), in some cases they might be useful. For example, had treatment $w + nw \longrightarrow w(b)$ been $w + nw \longrightarrow w(b) + nw$ instead, would there be a different effect on the gender hiring gap? This comparison, expressed as $ATT(d|d) - ATT(d'|d') = (ATT(d|d) - ATT(d'|d)) + (ATT(d'|d) - ATT(d'|d'))$, requires the strong parallel trend assumption to warrant causal interpretation, otherwise the selection bias of some occupations removing the non-written exam and others keeping it implies $ATE(d) - ATE(d') \neq \mathbb{E}[Y_t(d) - Y_t(d')]$. In those pairwise comparisons, while I cannot test directly for strong parallel trends, I am assuming the following set of assumptions, shown by Callaway et al. (2021) to be equivalent to assuming strong parallel trends:

$$
\begin{aligned}
\mathbb{E}\left[Y_t(0) - Y_{t-1}(0)\right] =& \mathbb{E}\left[Y_t(0) - Y_{t-1}(0)|D = 0\right] \\
\mathbb{E}\left[Y_t(d) - Y_{t-1}(0)\right] =& \mathbb{E}\left[Y_t(d) - Y_{t-1}(0)|D = d\right] \\
\mathbb{E}\left[Y_t(d') - Y_{t-1}(0)\right] =& \mathbb{E}\left[Y_t(d') - Y_{t-1}(0)|D = d'\right]
\end{aligned}
$$

Violations to the first assumption can be tested by conducting a standard parallel trends "inspection", where each pairwise comparison from **g** should have pre-trends with similar behavior (note that these comparisons have the same pre-treatment mix of screening tools).[27]

6.3 Estimation & Results

I now proceed to estimate model (11) in five separate regressions for gender final score gaps and gender hiring gaps. Figure (8) shows treatment effects of final scores. For conciseness, I center my discussion in Figure (9), which conducts the same analysis using the gender hiring gap as outcome. Since job offers are entirely based off final scores and job openings, any improvement in women's hiring rates relative to men's implies a decrease in the gender final score gap.

Figure (9) analyzes in three groups the five treatment types induced by the policy. Each group has the same baseline or pre-policy screening tool mix — w, nw, and $w + nw$ — for

[27] If the standard parallel trends assumption holds, then

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0)|D = 0\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0)|D = d\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0)|D = d'\right]$$

and thus the first condition of strong parallel assumptions follows, since it is equivalent to the decomposition

$$\mathbb{E}\left[\Delta Y_t(0)|D = 0\right] = \mathbb{E}\left[\Delta Y_t(0)|D = d\right] \frac{P(D = d)}{P(D = d) + P(D = d')} + \mathbb{E}\left[\Delta Y_t(0)|D = d'\right] \frac{P(D = d')}{P(D = d) + P(D = d')}.$$

which I then estimate treatment effects depending on each complier type. To benchmark the following coefficient magnitudes, the initial hiring gap in the federal sector for each case is 1.5%, 17%, and a slight gender hiring advantage in the $w + nw$ case of 0.5% (although the sample average statistic is non-significant). Note that, at least observationally, the hiring gap is much larger in job processes relying solely on non-written stages.

Starting with how the gender hiring gap changed when job processes within the same occupation switched from a written test to a blind-written, the estimated decrease in the gender gap of 0.5 percentage point cleanly measures disparate treatment, or the impact of complete removal of all evaluator bias sources, for a given level of disparate impact and screening precision of written tests. The pre-policy use of written tests provided the smallest (significant) effect on gender hiring gap, likely due to the low discretion from the screening type. The estimated impact corresponds to a decrease in the gender gap of about 33%.[28]

Next, I analyze how two different treatments to a screening strategy using non-written interviews affected the gender gap. In this case, because interviews are are high-discretion tools and leave employers susceptible to bias, they may be interested in removing or replacing non-written stages. However, they may also recognize that interviews could have higher screening precision if managers are better (privately) informed than they are biased. Carefully weighing these considerations is important to ensure a hiring system more equitable and that selects the most productive candidates.

Under the assumptions stated before, the two treatment types provide counterfacults in a similar sense as Mountjoy (2021) in the IV case. When job processes switch from nw to a written-blind, the gender hiring gap decreases by almost 7 percentage points. When benchmarked against the initial gap level, the estimated magnitude implies a decrease in the gender gap of 41%, a larger relative response than $w \longrightarrow w(b)$. In light of the model in the previous section, this treatment type involves the most dramatic changes to all forces determining hiring rates. Suppose evaluators favor men (which is the case in the first estimated effect), the pure disparate treatment channel will help women. Because the net result from changes in screening precision and tool bias may depress female's hiring rates, the large estimated result suggests that either written-exams have higher precision or smaller disparate impact than non-written, or that the combined magnitude of these channels is small relative to the size of disparate treatment in interviews.

The next treatment type and alternative "counterfactual" to the previous case involves adding a blind-written exam to a pre-existing non-written stage. This is an interesting case in

[28]Under the additional assumption that the contents of written tests are exactly the same before and after the policy (I find no evidence to the contrary), blinding the test measures discrimination in terms of disparate treatment by partitioning out the effect of bias from decisions makers, and from disparate impact of written exams.

light of growing criticism over requiring standardized or written tests which could disadvantage women or minorities (introduce disparate impact). My estimates suggest that the potential negative effects from these evaluation methods is more than compensated by gains in screening precision, which helps the minority group. The estimated effect is about 5.9 percentage points, or 35% of the initial hiring gap from using nw. From the reduced form results in the previous section, this treatment improves screening precision without introducing additional disparate treatment by adding $w(b)$, which favors women. Moreover, with another productivity signal, the final score and therefore hiring threshold relies less on the group mean and less weight is given to evaluator bias still remaining in nw, further helping women. However, the introduction of $w(b)$ adds a tool with potentially different screening precision and bias. If the written exam has no disparate impact or favors female candidates, it will also have a positive impact on women's hiring rate.

The next two estimates compare alternative treatments of a screening process containing a mix of written and non-written tools, $w + nw$. The first case is particularly interesting since it could be interpreted as an employer induced to drop a high-discretion screening tool (nw) and altering the other to ensure no disparate treatment. This can appeal as an approach to employers interested in reforming hiring practices by removing stages seen as potential barriers to increasing diversity, fine-tuning the remaining practices, but not replacing the removed tool with any other signal.

The lack of a statistically significant effect — albeit being precisely estimated — indicates that potential gains from removing disparate treatment and potential disparate impact from interviews and evaluator bias from the written test are offset by loss in precision from nw. How much does this precision loss matter? Assuming that evaluators favor men, so that removal of disparate treatment helps women (both via blinding and removing nw), either the precision loss of non-written exams is large enough to offset the complete elimination of evaluator bias and disparate impact of interviews (if they favor women), or interview signal precision matters less because interviews favor women (on the disparate impact margin).

The next treatment estimated starts with the same screening mix $w + nw$, but only blinds the written stage. Compared to the example before, this captures an employer who only fine-tuned some existing practices (by blinding), without removing any stages. I estimate another null effect, although estimates in this case are less precise. Note that the reduced form prediction for $nw + w \longrightarrow w(b) + nw$ is that the female hiring gap increases relative to men due to partial removal of evaluator bias.

To conclude this section, Table (5) compares female participation shares in applicant pools for each treatment type. Consistent with the idea from section 4.8 that perceived discrimination or unfair treatment during hiring may prevent minorities from applying in the first place, from section , column (1) finds that by blinding the written stage, the participation of women in the applicant pool increases by 2%. Column (2) shows that by switching from a non-written exam to a written stage women did not increase application rates relative to men. With a completely different screening method, women may think that the process is fairer, but may be uncertain about potentially allocating more time to prepare for the test. Alternatively, men could interpret the new testing method as a more competition-driven environment, eliciting more male candidates to apply.

Somewhat in line with this cost of application explanation, column (3) shows that when employers introduced an additional screening requirement, even if that could potentially appeal to female candidates as resulting in a less biased process, the higher application cost might be the dominant force, since the estimated effect is zero, so that it does not incentivize or disincentivize applications from women more compared to men. On the other hand, by removing an interview from $w + nw$, column (4) provides an increase in the female share of applicants of 5%. Similar to column (1), blinding a written exam in the treatment type $nw + w \longrightarrow w(b) + nw$ improves female application rates relative to men. Finally, keeping all screening tools but fine-tuning written exams by making them blind also increases female participation.

7 Impact of Who Hires

In the previous sections, I have focused on the redesign of screening tools (blinding written exams, removing subjective stages, among others) to improve gender equality in labor markets. This approach recognizes that certain hiring practices leave firms more susceptible to biased decisions and may impact minority groups differently. I now turn my analysis to another determinant of unequal treatment: hiring managers. There are two main reasons to study the effect of who hires as a complementary strategy to improve minority's outcomes in hiring processes.

First, the changes in screening tools I studied involve limiting evaluator bias expression via more or less discretion, considering as given managers' potentially biased behavior. Blinding exams removes any manager-related bias expression, and altering the screening tool mix to make it less discretionary removes part of the weight from evaluators' private signal or bias. However, some of these choices come with consequences. My estimates indicate that the removal of interviews or other non-written screening tools from a mix containing written and non-written methods does not appear to improve women's final scores or hiring rates. This suggests that removing discretionary screening tools is not necessarily helpful to increase gender equity, as loss in precision may offset potential relative gains from removing disparate treatment and impact.

Second, employers and policymakers may have an intrinsic preference for promoting less biased decisions, and not necessarily be willing to modify screening tools. Evaluator discretion is likely to provide important private information that could both increase efficiency of hires and even help minorities if the disparate impact of standardized or written tests is relatively larger. Intuitively, the usefulness of standardized or blind tests depends on the context: if an employer is hiring a driver, it is unlikely that a math test will select the best candidates at driving. If a short driving test of each candidate seems a reasonable screening option, implementing the stage in a "blind" manner certainly appears challenging. On the other hand, my previous estimates show that women face disparate treatment from evaluators even in screening stages with little discretion.

To investigate how committee composition contributes to biased hiring decisions, I take advantage of several features in Brazil's public sector data. First, I compare blind written and non-written scores a candidate receives from the same exam which allows me to compare blind written and non-written scores given by the same evaluator to a candidate

7.1 The Role Of Diverse Committees

A common strategy to improve diversity of employees being hired is to make who hires more diverse. A diverse hiring committee should bring various viewpoints into the search process, providing more nuanced evaluations of applicants with a different sets of characteristics. In the case of gender, a "critical mass of women" in a team (Kanter (1977)) may correlate with group performance (Woolley et al. (2010)) and influence behavioral changes in male colleagues (Adams and Ferreira (2009)).

I now use more recent data pooled from Brazil's federal and state governments. The data spans 1999-2019 and broadly corresponds to the same setting in the impartiality reform. The only difference, of course, is that written exams in the sample are blind. Table (6) shows descriptive statistics by candidate gender for various job process evaluation scores. Women receive slightly lower scores for resumes than men and have slightly higher scores in blind written exams, albeit both of these differences are not statistically significant. However, female candidates receive 4 percentage points less in non-written exams than men, which results in a non-written blind-written score gap for women, while men have virtually the same performance in both exam types on average. Finally, despite having higher resume and blind written scores, women's final scores are 2 percentage points lower than men's because of penalties in non-written stages.

While the differences above do not necessarily reflect evaluator bias in non-written exams, Table (7) reveals an interesting pattern related to the gender composition of hiring committees. Hiring odds of female candidates in committees with less than 30% of women are much lower than men's. As the gender ratio of the committee starts to balance, hiring changes of both groups begin to align, until female candidates become slightly favored when the committee is female-majority.

The patterns above indicate that, either due to lower skill or actual performance in non-written exams, or due to some factor related to the higher discretion degree in these stages, women's lower scores in interviews, practical exams, oral presentations, penalize their final score, despite better or similar performance in blind exams relative to men. Moreover, when committees become more female-dominated, men's hiring rates decline, and women's increase.

To tease out the effect of committee gender composition on gender gaps, I implement a difference-in-difference-in-differences approach where I net out differences in individuals' skills between written and non-written exams:

$$\underbrace{\text{Score}^{nw}_{icj} - \text{Score}^{w(b)}_{icj}}_{\Delta S_{icj}} = \beta\left(\text{Female}_i \times \%\text{Female Evaluator}_c\right) + \gamma_c + \mu_i + \varepsilon_{icj}$$

where the difference between a candidate's blind written score and in a non-written exam is regressed on a female indicator interacted with the share of female committee members. The regression also controls for candidate and committee fixed effects, so that β is identified off candidates who were evaluated by committees with different gender compositions (i.e., individuals who applied to more than one job). Combining the comparison of blind and non-blind assessments with contemporaneous within-individual comparisons across hiring processes deals with the concern that blind and non-blind examinations may not necessarily measure the same skills.[29] The identification assumption in this case is that candidates differences in abilities between blind written and non-written tests is constant across hiring processes.

The identifying assumption in this case is that individual differences in written and non-written exams do not vary based on gender as candidates apply to different jobs.[30] Finally, note that in light of my conceptual framework, another interpretation for differential ability between distinct exam types is that that these could capture disparate impact of each tool.

[29]The within-individual comparison also accounts for differences between blind and non-blind examinations in terms of their potential different disparate impact, namely, whether one type of examination is favoring one group more than other in a way that is not related to productivity.

[30]Studies have used differences between men and women's gaps in blind and non-blind tests to identify discrimination (Blank (1991), Goldin and Rouse (2000), as well as teacher bias in grading:Lavy (2008), Breda and Ly (2015). However, double (or "simple") differences identified off comparisons between individuals and may be biased by gender-specific differences in individual ability between blind and non-blind examinations.

Table (8) shows the results. The first column presents the gender differences in the raw blind gap (ΔS_{icj}). Women have a slightly lower non-written premium relative to men, consistent with the previous discussion. When analyzing this subjectivity premium by committee composition, female candidates evaluated by committees with less than 50% women receive an even lower non-written premium relative to men. Strikingly, this pattern reverses when the committee is female dominated: women receive the same non-written premium as men or even higher. When controlling for individual differences in skills between the two exam types and directly assessing the effect of higher shares of female committee members, columns (4) through (6) show that the non-written premium for female candidates grow with more women in the committee, just as the final score received and the probability of women receiving an offer relative to men.

7.2 Differences in Scores Between Male and Female Evaluators

Why do female candidates receive higher scores when there are more women in the hiring committee? One hypothesis, based on evidence such as Adams and Ferreira (2009), who find that male board member attendance improves when there are more women on the board, or findings in Boyd et al. (2010) which show that male judges are more likely to hand down favorable decisions to plaintiffs alleging gender discrimination when they serve on panels with a female judge or become more likely to hire female law clerks (Battaglini et al. (2020)), is that the presence of female colleagues changes male evaluators behavior diferentially for female and male job applicants.[31]

Even if the presence of female colleagues affect male managers remaining in the committee — perhaps by increasing awareness over implicit biases or because of a censoring effect —, it is not clear whether the new female members would benefit female candidates. Homophilic competition may emerge, in a similar sense to Beaman et al. (2012) who argues about members of an ethnic network facing a trade-off in the context of job referrals due to competition over employment in an occupational niche. In this case, perhaps the change in men's behavior from

[31] In contrast, in the context of academic promotion competitions in Italy and Spain, Bagues et al. (2017) find that male evaluators become less favorable toward female candidates with women on the committee. However, under which hiring practices evaluators interact with applicants is not not taken into account. This relationship is important since hiring practices determine the discretion level to evaluators and consequently the extent that human bias can be expressed. Zinovyeva and Bagues (2011) document either a positive or non-significant relationship between the proportion of female evaluators and success of female applicants, depending on the position, in academic promotions in Spain. On the other hand, for similar academic promotions in Italy, Bagues et al. (2014) find a negative effect of higher share of women on committees and success rate of female candidates. The difference could lie in the hiring practices used in these different settings, as in Spain evaluations involved a research seminar given by the candidate, while in Italy the evaluation relied completely on CVs and publications and it did not require any personal interaction between evaluators and candidates.

the presence of higher numbers of female colleagues accounts for most of the equity benefit toward female candidates.

To understand how different committee members react to the committee gender make-up, I first analyze how female evaluators score female candidates relative to men. Figure (10) shows that female judges are more give women better scores on non-written exams, and surprisingly give lower scores than male evaluators for blind written exams. As a result, female committee members give a 4 percentage-point non-written bonus to women. In Table (9), I analyze how women's scores given by female and male committee members change relative to men's when the hiring committee has higher shares of women. The specification also controls for evaluator fixed effect, so that each estimate captures how the same evaluator of a given gender scores women relative to men when there are different proportions of female colleagues in the hiring committee.

Columns (1) through (4) show that female committee members do not particularly change their scoring behavior when there are more female colleagues. The estimated effect on the non-blind premium is small and almost statistically non-significant. Columns (5) through (8) show a different response by male committee members. The same male evaluator scores female candidates 0.7 percentage points higher when there are more female colleagues in the committee. As expected, scoring in blind tests remains the same despite of committee composition, which then implies a 1.4 percentage-point non-blind bonus to women.

7.3 Raising the Costs to Evaluator Bias Expression

Why would the presence of female colleagues reduce bias from male evaluators toward female applicants? Re-examining the hiring manager's decision problem from Section (6):

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

where the cost of expressing bias, previously fixed, might be determined endogenously based on some underlying characteristic of the job hiring process: $c_\theta(\phi)$. Therefore, by finding an appropriate ϕ such that $c'(\phi) > 0$, employers could successfully reduce the disparate impact of non-written exams.

In practical terms, $c_\theta(\phi)$ represents the cost of detection an evaluator faces when giving a score that deviates from the signal provided by the screening tool. The higher the discretion degree of a tool (e.g., interview), the harder for the employer to observe the deviation from the observed signal. One choice of variable for ϕ could include the transparency level in the screening practice, either by conducting interviews in front of impartial observers (say, compliance officers), or recording the process (which would potentially expose firms to litigation risk).

Job processes in Brazil's public sector implement these measures to some extent, since hiring must be transparent. Yet, even in settings with little discretion, I find evidence of disparate treatment.

To inform on the choice for ϕ, I drew from a large set of existing literature that shows the effects of conditioning evaluator's behavior along some demographic variable, particularly matching the demographic groups being studied. Formally, I let

$$c_\theta\left(\frac{\sum_j f_j}{\sum_j f_j + \sum_j (1-f_j)}\right), \qquad c'_\theta > 0$$

with higher-order derivatives unsigned for now and f_j representing the number of female committee members. The expression above considers that bias expression of an evaluator toward a minority group increases in the share of committee members of that minority group. Put it simply, expressing bias against female candidates becomes increasingly more costly when more women are hiring committee colleagues, which captures the effects I estimate in Table (9).

8 Conclusion

This paper opens the black-box of hiring decisions and studies how screening practices and who screens determine employee gender diversity. Using the universe of Brazil's public sector job processes and uniquely detailed data with candidate performance, screening methods, and evaluator decision at all hiring stages, I show that between improving screening tools or making hiring managers less biased, employers can do both.

Exploiting a policy that generated several treatment types to improve hiring impartiality, I find that switching from an interview to a blind written exam or adding the test to subjective stages raises women's hiring rates. However, removing non-written rounds, even if written tests are conducted blind, has no impact toward greater gender equity. I also show that evaluator bias matters even in hiring processes employing relatively objective measures, and that by increasing the share of female members in hiring committees, male colleagues significantly improve scores given to female applicants in subjective stages.

My results have implications for both firms in the private sector and bureaucracies around the world who rely on competitive examinations to select their civil servants. While I do not empirically test in this paper for effects of various screening tools on efficiency, most changes to hiring practices that increase hiring diversity guarantee employee productivity to at least remain the same or increase after implementation.

References

Abraham, Lisa and Alison Stein (2020) "Words Matter: Experimental Evidence from Job Applications," *Unpublished manuscript*.

Adams, Renée B and Daniel Ferreira (2009) "Women in the boardroom and their impact on governance and performance," *Journal of financial economics*, Vol. 94, No. 2, pp. 291–309.

Agan, Amanda and Sonja Starr (2018) "Ban the box, criminal records, and racial discrimination: A field experiment," *The 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.

Åslund, Olof and Oskar Nordströum Skans (2012) "Do anonymous job application procedures level the playing field?" *ILR Review*, Vol. 65, No. 1, pp. 82–107.

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," *The Quarterly Journal of Economics*, Vol. 123, No. 1, pp. 219–277.

Bagues, Manuel F and Berta Esteve-Volart (2010) "Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment," *The Review of Economic Studies*, Vol. 77, No. 4, pp. 1301–1328.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva (2014) "Do gender quotas pass the test? Evidence from academic evaluations in Italy," *Scuola Superiore Sant'Anna, LEM Working Paper Series*, Vol. 14.

——— (2017) "Does the gender composition of scientific committees matter?" *American Economic Review*, Vol. 107, No. 4, pp. 1207–38.

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

Bartik, Alexander and Scott Nelson (2021) "Deleting a signal: Evidence from pre-employment credit checks," *Working paper*.

Battaglini, Marco, Jorgen M Harris, and Eleonora Patacchini (2020) "Professional Interactions and Hiring Decisions: Evidence from the Federal Judiciary,"Technical report, National Bureau of Economic Research.

Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova (2012) "Female leadership raises aspirations and educational attainment for girls: A policy experiment in India," *science*, Vol. 335, No. 6068, pp. 582–586.

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 (2011) "New perspectives on gender," in *Handbook of labor economics*, Vol. 4: Elsevier, pp. 1543–1590.

Bertrand, Marianne, Sandra E Black, Sissel Jensen, and Adriana Lleras-Muney (2019) "Breaking the glass ceiling? The effect of board quotas on female labour market outcomes in Norway," *The Review of Economic Studies*, Vol. 86, No. 1, pp. 191–239.

Bertrand, Marianne and Esther Duflo (2017) "Field experiments on discrimination," *Handbook of economic field experiments*, Vol. 1, pp. 309–393.

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.

Blank, Rebecca M. (1991) "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review," *The American Economic Review*, Vol. 81, No. 5, pp. 1041–1067.

Bohnet, Iris, Alexandra Van Geen, and Max Bazerman (2016) "When performance trumps gender bias: Joint vs. separate evaluation," *Management Science*, Vol. 62, No. 5, pp. 1225–1234.

Booth, Alison and Andrew Leigh (2010) "Do employers discriminate by gender? A field experiment in female-dominated occupations," *Economics Letters*, Vol. 107, No. 2, pp. 236–238.

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

Boyd, Christina L, Lee Epstein, and Andrew D Martin (2010) "Untangling the causal effects of sex on judging," *American journal of political science*, Vol. 54, No. 2, pp. 389–411.

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.

Breda, Thomas and Son Thierry Ly (2015) "Professors in Core Science Fields Are Not Always Biased against Women: Evidence from France," *American Economic Journal: Applied Economics*, Vol. 7, No. 4, pp. 53–75.

Broder, Ivy E. (1993) "PROFESSIONAL ACHIEVEMENTS AND GENDER DIFFERENCES AMONG ACADEMIC ECONOMISTS," *Economic Inquiry*, Vol. 31, No. 1, pp. 116–127.

Brollo, Fernanda, Pedro Forquesato, and Juan Carlos Gozzi (2017) "To the victor belongs the spoils? Party membership and public sector employment in Brazil," *Party Membership and Public Sector Employment in Brazil (October 2017)*.

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu (2020) "The structure of economic news,"Technical report, National Bureau of Economic Research.

Cahuc, Pierre, Stephane L Carcillo, Andreea Minea, and Marie-Anne Valfort (2019) "When correspondence studies fail to detect hiring discrimination."

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna (2021) "Difference-in-Differences with a Continuous Treatment," *arXiv preprint arXiv:2107.02637*.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri (2019) "Are Referees and Editors in Economics Gender Neutral?," *The Quarterly Journal of Economics*, Vol. 135, No. 1, pp. 269–327.

Coffman, Katherine B, Manuela Collis, and Leena Kulkarni (2021) "When to Apply?" *Working paper*.

Colonnelli, Emanuele, Mounu Prem, and Edoardo Teso (2020) "Patronage and selection in public sector organizations," *American Economic Review*, Vol. 110, No. 10, pp. 3071–99.

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," *The Quarterly Journal of Economics*, Vol. 128, No. 3, pp. 1169–1218.

De Paola, Maria and Vincenzo Scoppa (2015) "Gender discrimination and evaluators gender: evidence from Italian academia," *Economica*, Vol. 82, No. 325, pp. 162–188.

Del Carpio, Lucia and Maria Guadalupe (2021) "More women in tech? Evidence from a field experiment addressing social identity," *Management Science*.

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.

Dobbin, Frank and Alexandra Kalev (2016) "Why diversity programs fail," *Harvard Business Review*, Vol. 94, No. 7, p. 14.

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.

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

Feller, Avi, Todd Grindal, Luke Miratrix, and Lindsay C. Page (2016) "Compared to what? Variation in the impacts of early childhood education by alternative care type," *The Annals of Applied Statistics*, Vol. 10, No. 3, pp. 1245 – 1285.

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, 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.

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.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini (2003) "Performance in competitive environments: Gender differences," *The 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.

Guryan, Jonathan and Kerwin Kofi Charles (2013) "Taste-based or statistical discrimination: the economics of discrimination returns to its roots," *The Economic Journal*, Vol. 123, No. 572, pp. F417–F432.

Heckman, James J (1998) "Detecting discrimination," *Journal of economic perspectives*, Vol. 12, No. 2, pp. 101–116.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li (2018) "Discretion in hiring," *The 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," *The 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," *The Review of Economics and Statistics*, pp. 1–45.

Kalev, Alexandra, Frank Dobbin, and Erin Kelly (2006) "Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies," *American sociological review*, Vol. 71, No. 4, pp. 589–617.

Kanter, Rosabeth Moss (1977) "Some Effects of Proportions on Group Life: Skewed Sex Ratios and Responses to Token Women," *American Journal of Sociology*, Vol. 82, No. 5, pp. 965–990.

Kessler, Judd B, Corinne Low, and Colin D Sullivan (2019) "Incentivized resume rating: Eliciting employer preferences without deception," *American Economic Review*, Vol. 109, No. 11, pp. 3713–44.

Kline, Patrick M, Evan K Rose, and Christopher R Walters (2021) "Systemic Discrimination Among Large US Employers,"Technical report, National Bureau of Economic Research.

Kline, Patrick and Christopher R. Walters (2016) "Evaluating Public Programs with Close Substitutes: The Case of Head Start," *The Quarterly Journal of Economics*, Vol. 131, No. 4, pp. 1795–1848.

Krause, Annabelle, Ulf Rinne, and Klaus F Zimmermann (2012) "Anonymous job applications of fresh Ph. D. economists," *Economics Letters*, Vol. 117, No. 2, pp. 441–444.

Lavy, Victor (2008) "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment," *Journal of public Economics*, Vol. 92, No. 10-11, pp. 2083–2105.

Moreira, Diana and Santiago Pérez (2021) "Civil Service Reform and Organizational Practices: Evidence from the Pendleton Act,"Technical report, National Bureau of Economic Research.

——— , "Who Benefits from Meritocracy?".

Mountjoy, Jack (2021) "Community colleges and upward mobility,"Technical report, National Bureau of Economic Research.

Neumark, David (2018) "Experimental research on labor market discrimination," *Journal of Economic Literature*, Vol. 56, No. 3, pp. 799–866.

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

Neumark, David, Roy J Bank, and Kyle D Van Nort (1996) "Sex discrimination in restaurant hiring: An audit study," *The Quarterly journal of economics*, Vol. 111, No. 3, pp. 915–941.

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

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," *The american economic review*, Vol. 62, No. 4, pp. 659–661.

Sarsons, Heather (2019) "Interpreting signals in the labor market: evidence from medical referrals."

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.

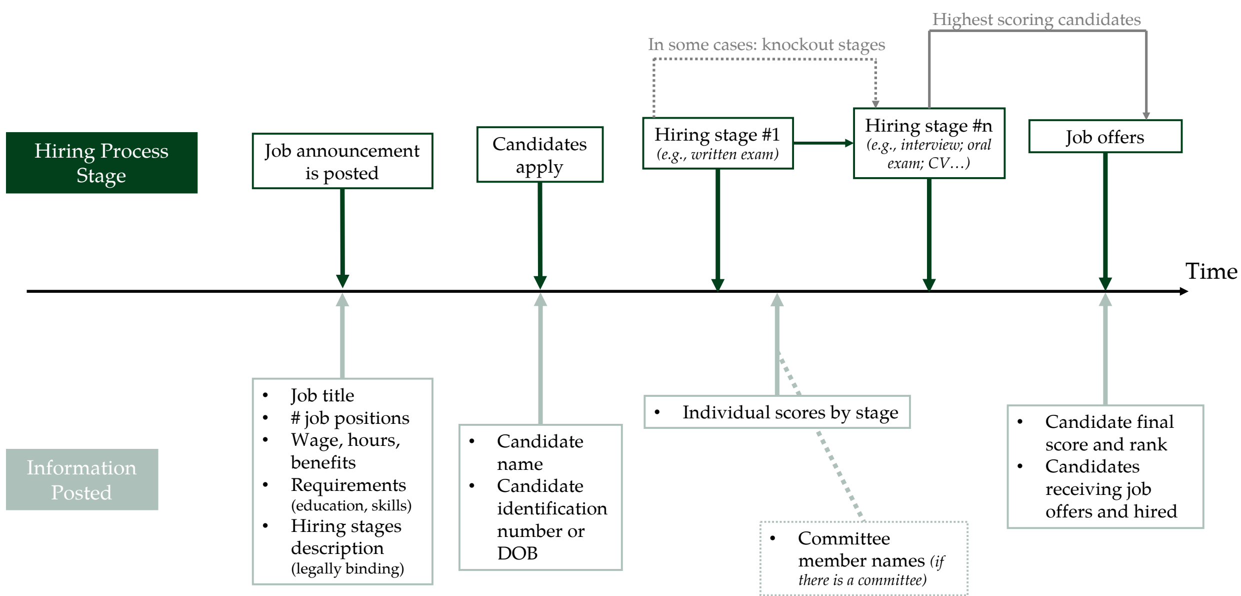
Woolley, Anita Williams, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone (2010) "Evidence for a collective intelligence factor in the performance of human groups," *science*, Vol. 330, No. 6004, pp. 686–688.

Wozniak, Abigail (2015) "Discrimination and the effects of drug testing on black employment," *Review of Economics and Statistics*, Vol. 97, No. 3, pp. 548–566.

Xu, Guo (2018) "The costs of patronage: Evidence from the british empire," *American Economic Review*, Vol. 108, No. 11, pp. 3170–98.

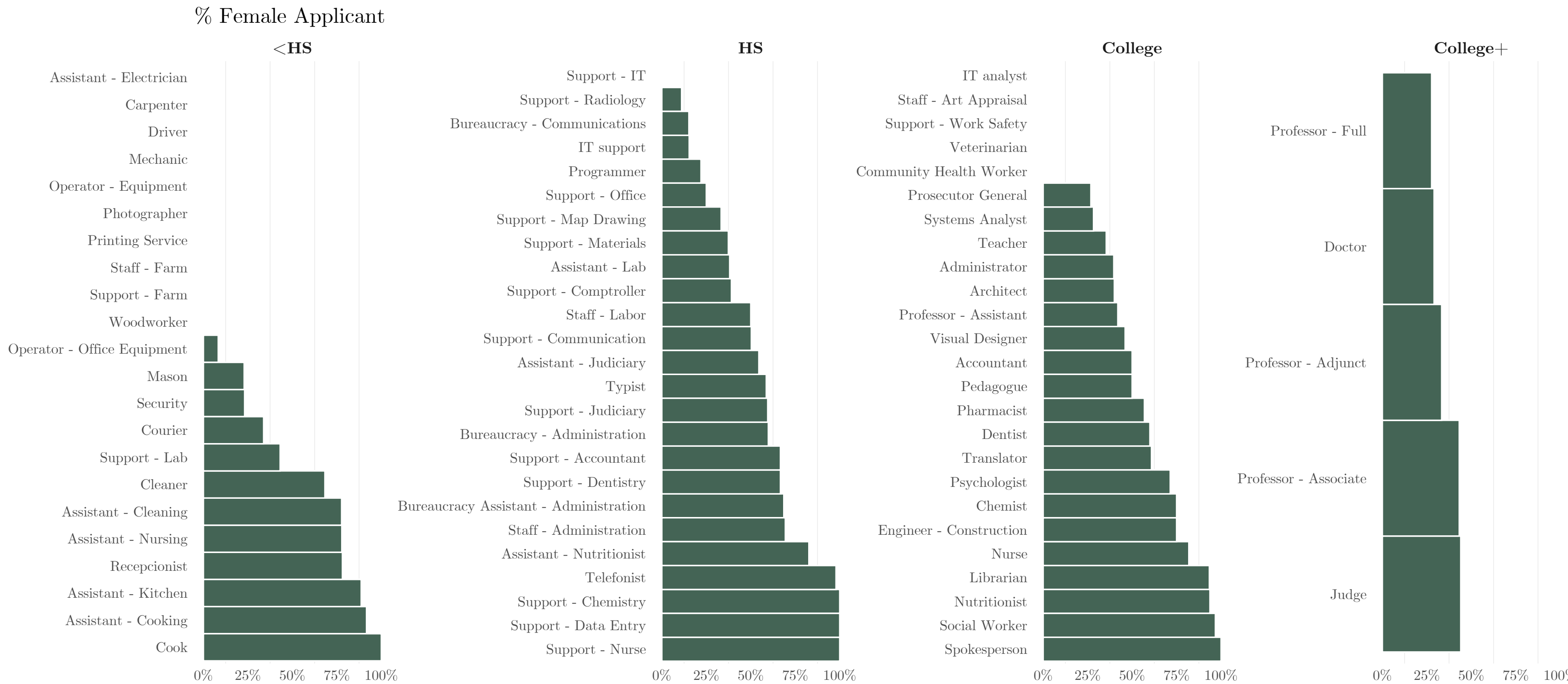
Zinovyeva, Natalia and Manuel F Bagues (2011) "Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment. IZA Discussion Papers 5537," *Institute for the Study of Labor (IZA). Figure I: Kernel density estimates of scores at written and oral tests, by track and gender*.

9 Figures & Tables



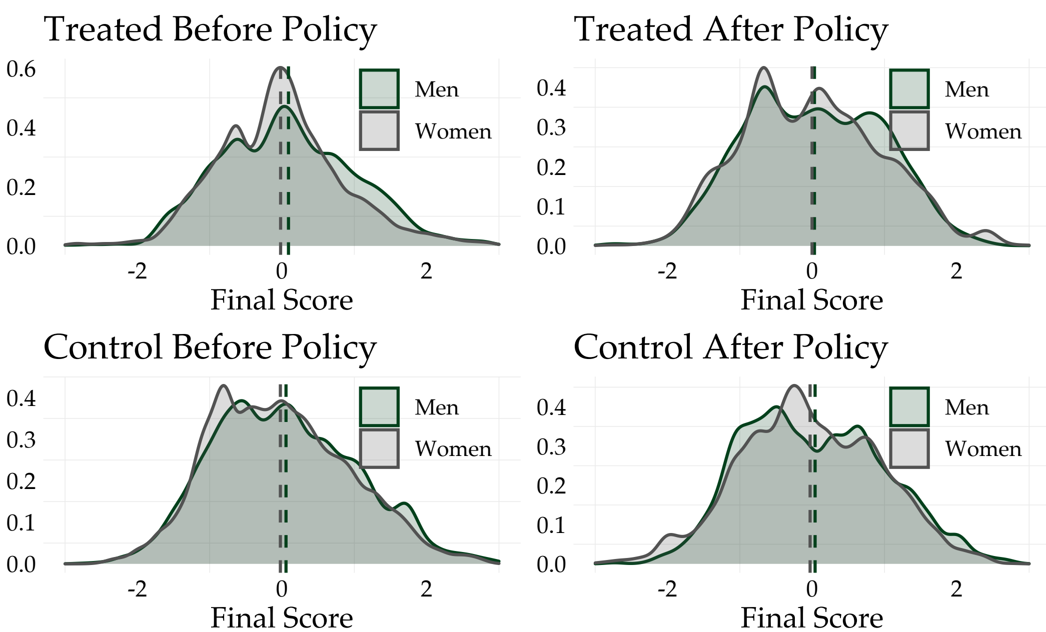
Notes: This shows the stylized structure of a hiring process in the Brazilian public sector posted in raw government publications (official gazettes). Information at the top (dark green) describes the screening dynamics from the moment a job is announced until job offers are sent out. The lower part 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 Hiring Processes



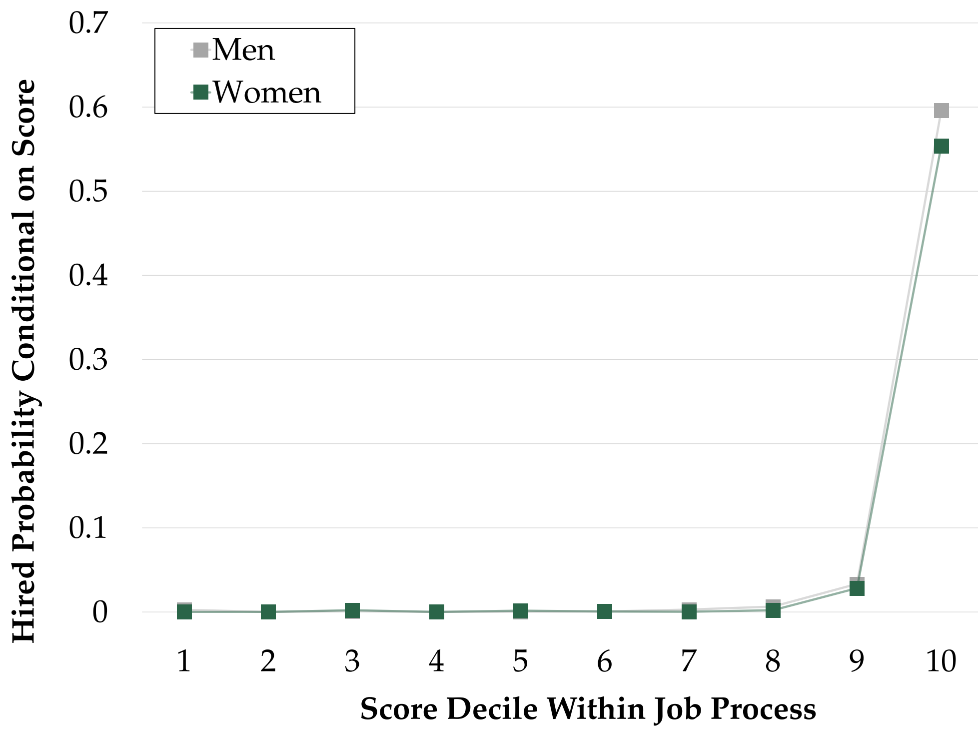
Notes: This figure shows the gender share distribution of job applicants to various occupations and skill levels in Brazil's public sector from 1986 to 1991. Occupation titles in the data follow employer-specific career titles given each organizational structure. These titles are translated from Portuguese 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 2: Distribution of Female Applicants by Occupation and Skill Level



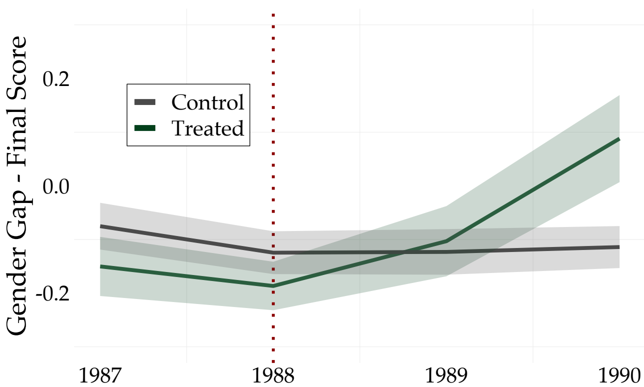
Notes: These panels show standardized final scores of male and female job applicants in federal and state job processes. Federal (treated) and states (control) before and after the impartiality reform are displayed in each panel. To compare magnitudes across densities, tails are censored between 2 standard deviations right and left.

FIGURE 3: Final Scores Distributions

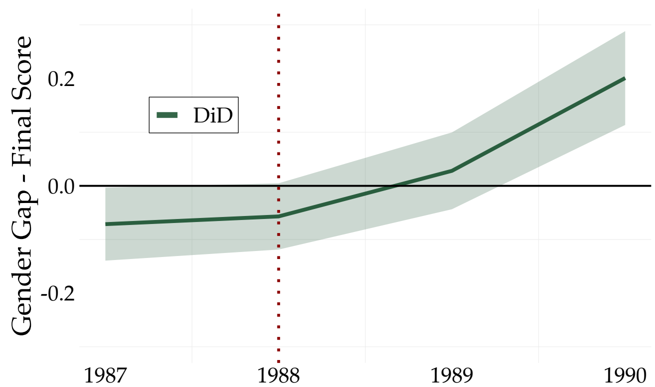


Notes: This figures illustrates how final scores completely determine candidates' probability of bein hired in public sector job processes. In accordance with the law requiring that the highest scoring candidates across all evaluation stages are offered jobs until all openings are fulfilled, candidates with scores in the highest decile in their job process have a 60% probability of receiving a job offer. Top performing women have a slightly lower probability of receiving a job offer than top performing men (across all job processes). Not all top candidates receive offers because public sector jobs are oversubscribed.

FIGURE 4: Candidate Final Scores Determine Hiring Chances



(a)



(b)

Notes: The first figure plots $\widehat{\gamma}$ estimates of the regression

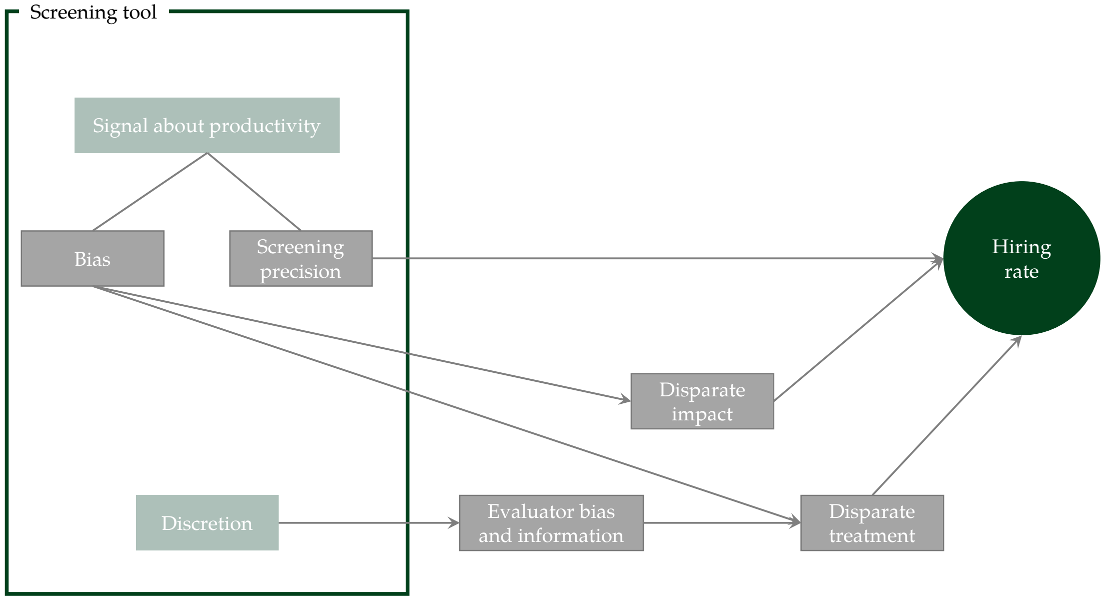
$$\text{Final Score}_{it} = \delta_{o(i)} + \gamma \text{Female}_i + u_{it}$$

for control (state governments) and treated (federal government) groups in each year, where i denotes the job selection process and o the occupation or job title. The second shows dynamic effects of the baseline DiD model

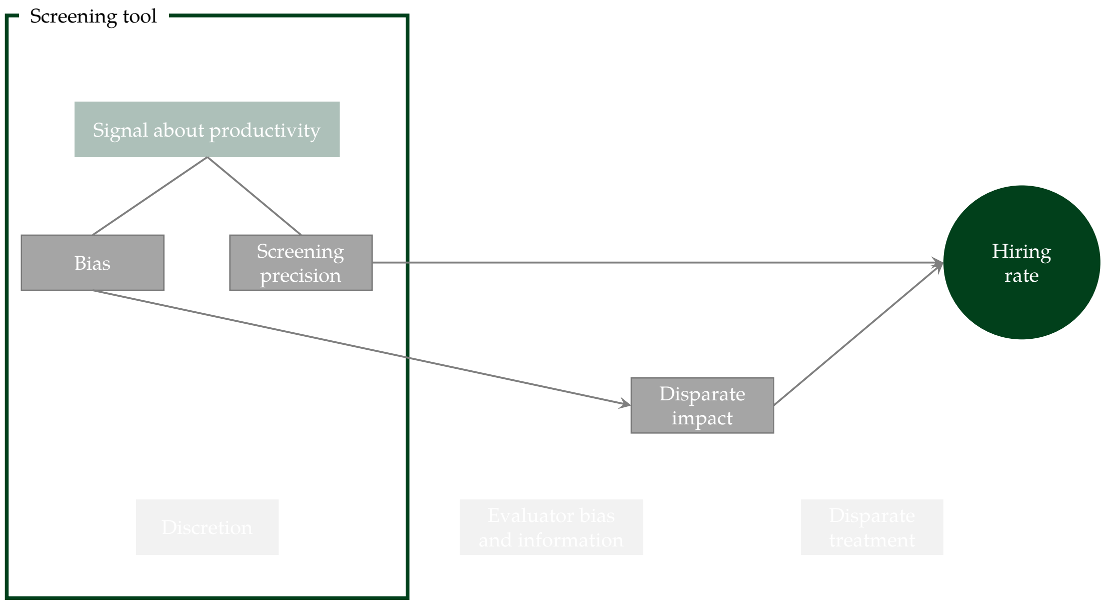
$$\text{Final Score}_{it} = \delta_{o(i)} + \beta \left(\text{Post}_{o(i),t} \times \text{Female}_i \right) + \gamma_t + u_{it}$$

where $\text{Post}_{o(i),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level. Shaded areas are 95% confidence intervals. Pooled DiD estimates are shown in Table (2).

FIGURE 5: DiD Dynamic Effects of Impartial Hiring on Final Scores



(a) Standard Screening Tool



(b) Blinding Screening Tool

Notes: This figure represents the main forces captured in my conceptual framework that determine hiring rates of candidates evaluated using a screening tool, like a test or an interview. A screening tool provides a productivity signal with certain precision, but the signal can have a bias that favors a specific demographic group. In the model, this bias term receives the interpretation of a disparate impact. The other property of a screening tool is the degree of discretion it enables. More subjective practices allow hiring manager's evaluation to deviate from the signal provided more easily. When evaluators are biased toward a group, the screening practice also allows disparate treatment. By concealing candidates' identities — when possible or desirable — in the screening tool, managers cannot express bias or statistically discriminate, leaving only precision and tool bias to determine hiring rates.

FIGURE 6: Hiring Rate Determinants: Conceptual Framework

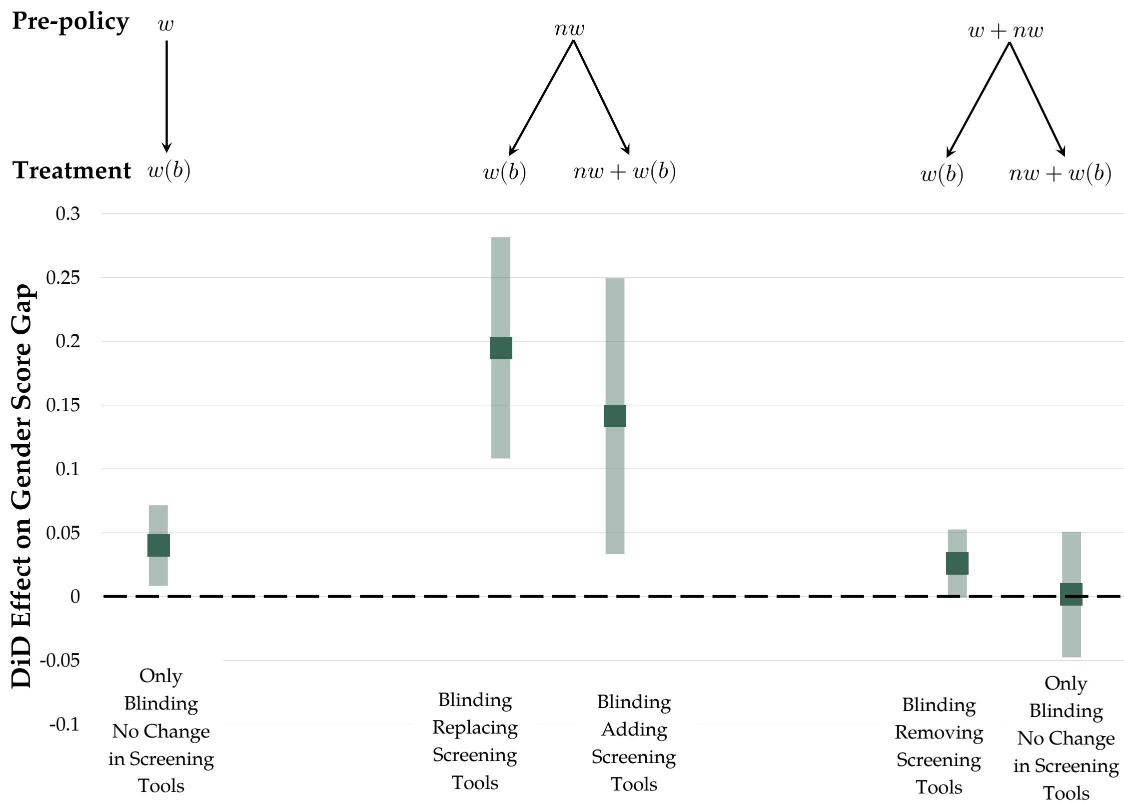
$z = 0$

$z = 1$

	Written	**Non-written**	**Written & Non-written**	**Written Blind**	**Written Blind & Non-written**
Written	Always Written				
Non-written		Always Non-written			
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 1988 Impartiality Reform on federal jobs in Brazil's public sector. Areas shaded in gray are ruled out by standard DiD assumptions and 5 out of the 6 allowed treatments is consistent with the variation induced by the policy: a 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 ($w + nw \longrightarrow w(b) + nw$). The potential treatment $w \longrightarrow w(b) + nw$ is invalided by the research design since this would go in the opposite direction of decreasing overall partiality in the hiring process. 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 7: Potential Treatments Induced by Reform

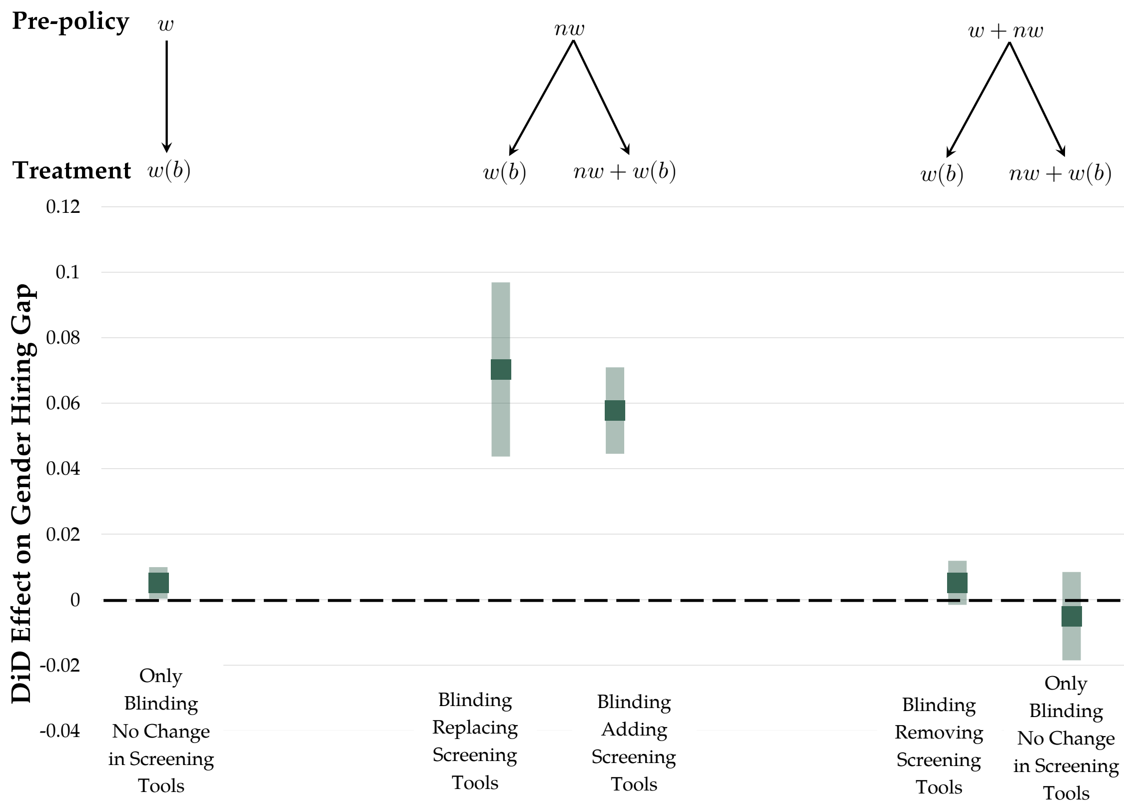


Notes: This figure plots treatment effects for each treatment type g induced by the 1988 Impartiality Reform in Brazil's public sector. Each bar central point estimates a version of the DiD regression

$$\text{Final Score}_{git} = \delta_{o(g,i)} + \beta_g \left(\text{Post}_{o(g,i),t} \times \text{Female}_i\right) + \gamma_t + u_{git}$$

where $\text{Post}_{o(g,i),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Treatment type g 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 ($w + nw \longrightarrow w(b) + nw$). Standard errors are clustered at the job process level. Bars are 95% confidence intervals.

FIGURE 8: Treatment Effects of Changes in Screening Tools: Gender Final Score Gap

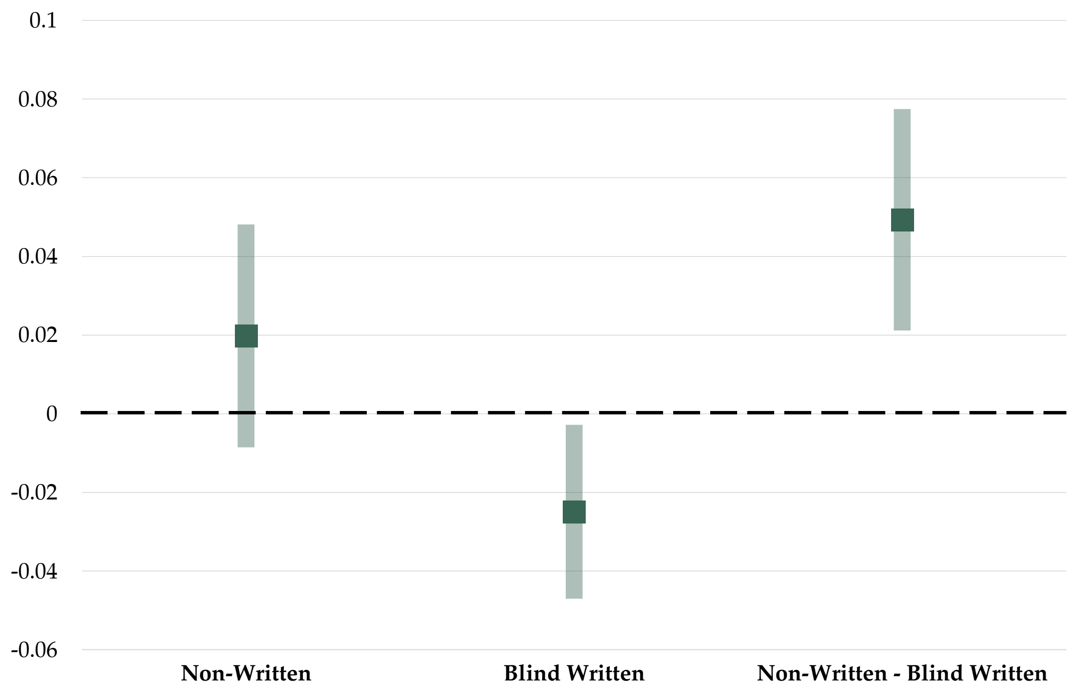


Notes: This figure plots treatment effects for each treatment type g induced by the 1988 Impartiality Reform in Brazil's public sector. Each bar central point estimates a version of the DiD regression

$$\mathbf{Pr}(Hired = 1)_{git} = \delta_{o(g,i)} + \beta_g \left(\text{Post}_{o(g,i),t} \times \text{Female}_i\right) + \gamma_t + u_{git}$$

where $\text{Post}_{o(g,i),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Treatment type g 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 ($w + nw \longrightarrow w(b) + nw$). Standard errors are clustered at the job process level. Bars are 95% confidence intervals.

FIGURE 9: Treatment Effects of Changes in Screening Tools: Gender Hiring Gap



Notes: This figure plots treatment effects for each treatment type g induced by the 1988 Impartiality Reform in Brazil's public sector. Each bar central point estimates a version of the DiD regression

$$\mathbf{Pr}(Hired = 1)_{git} = \delta_{o(g,i)} + \beta_g \left(\text{Post}_{o(g,i),t} \times \text{Female}_i\right) + \gamma_t + u_{git}$$

where $\text{Post}_{o(g,i),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Treatment type g 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 ($w + nw \longrightarrow w(b) + nw$). Standard errors are clustered at the job process level. Bars are 95% confidence intervals.

FIGURE 10: Treatment Effects of Changes in Screening Tools: Gender Hiring Gap

TABLE 1: Estimated Reaction to Impartiality Policy

	At Least One Written Stage		At Least One Non-Written Stage		Only One Round & Written	All Rounds Non-Written
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{Post}_{o(j),t}$	-0.006	0.252***	-0.581***	-0.145	0.476***	-0.252***
	(0.078)	(0.090)	(0.060)	(0.092)	(0.081)	(0.090)
Occupation FE		X		X	X	X
Year FE	X	X	X	X	X	X
Job Processes	6,554	6,554	6,554	6,554	6,554	6,554

Notes: This table displays regression coefficients of the model $y_{jt} = \delta_{o(j)} + \alpha \text{Post}_{o(j),t} + \gamma_t + u_{jt}$, where outcomes in columns (1) through (6) at the job process level j are regressed on an indicator for treated post 1988 in federal jobs, $\text{Post}_{o(j),t}$, controlling for occupation title and year fixed effects. Each regression compares the effect of the impartiality reform on the outcome for the same occupation in the federal sector and states. Standard errors are clustered at the job process level.

TABLE 2: DiD Estimates of Screening Impartiality on Cadidate Scores

	Final Score			Written Score			Non-Written Score		
	Women	Men	Gap	Women	Men	Gap	Women	Men	Gap
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\text{Post}_{o(i),t}$	0.067**	−0.075*		0.024	−0.109*		−0.010	0.020	
	(0.030)	(0.037)		(0.044)	(0.059)		(0.071)	(0.099)	
$\text{Post}_{o(i),t} \times \text{Female}_i$			0.141***			0.134*			−0.031
			(0.048)			(0.074)			(0.122)
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 DiD estimates of the form $y_{it} = \delta_{o(i)} + \gamma \text{Post}_{o(i),t} + u_{it}$ only with female candidates in columns (1), (4), and (7), only with male candidates in columns (2), (5), and (8), and $y_{it} = \delta_{o(i)} + \beta \left(\text{Post}_{o(i),t} \times \text{Female}_i\right) + \gamma_t + u_{it}$ in the remaining columns. The outcome y represents either a candidate's final score, written score (written exams), or non-written score (interview, oral, practical exams). $\text{Post}_{o(g,i),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level.

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

	$\mathbf{Pr}(Hired \mid \text{Female})$	$\mathbf{Pr}(Hired \mid \text{Male})$	Hiring Gap	$\mathbf{Pr}(\text{Female})$
	(1)	(2)	(3)	(4)
$\text{Post}_{o(i),t}$	0.003**	-0.004***		0.010**
	(0.001)	(0.001)		(0.005)
$\text{Post}_{o(i),t} \times \text{Female}_i$			0.007***	
			(0.002)	
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: $\mathbf{Pr}(Hired = 1)_{it} = \delta_{o(i)} + \gamma \text{Post}_{o(i),t} + u_{it}$ which is ran only on female individuals, column (2) runs the same regression but in the male candidate subsample, column (3) runs $\mathbf{Pr}(Hired = 1)_{it} = \delta_{o(i)} + \beta \left(\text{Post}_{o(i),t} \times \text{Female}_i\right) + \gamma_t + u_{it}$. Finally, column (4) regresses the specification $\mathbf{Pr}(\text{Female} = 1)_{it} = \delta_{o(i)} + \beta \left(\text{Post}_{o(i),t} \times \text{Female}_i\right) + \gamma_t + u_{it}$. $\text{Post}_{o(g,i),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level.

TABLE 4: DiD Estimates of Screening Impartiality on Job Process Outcomes

	% Women of Hired	% Female Candidates	Log # Candidates		
			All	Women	Men
	(1)	(2)	(3)	(4)	(5)
$\text{Post}_{o(j),t}$	0.134**	0.061*	-0.245	-0.212	-0.316
	(0.0692)	(0.042)	(0.322)	(0.382)	(0.251)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Obs.	54,892	32,067	86,959	86,959	86,959

Notes: The table shows selection process level regressions $y_{it} = \delta_{o(i)} + \gamma \text{Post}_{o(i),t} + u_{it}$, where $\text{Post}_{o(g,i),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level.

TABLE 5: Treatment Effects of Changes in Screening Tools: % 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)
$\text{Post}_{o(c),t}$	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 g induced by the 1988 Impartiality Reform in Brazil's public sector. Each column estimates a version of the DiD regression

$$\text{\% Female Applicants}_{gct} = \delta_{o(g,c)} + \beta_g \text{Post}_{o(g,c),t} + \gamma_t + u_{gct}$$

where $\text{Post}_{o(g,c),t}$ is an indicator for whether the job process is for a federal-level position post Impartiality reform ($t \geq 1989$). Treatment type g 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 ($w + nw \longrightarrow w(b) + nw$). Standard errors are clustered at the job process level.

TABLE 6: Summary Statistics Hiring Committee Analysis

Panel A. Job Applicant Statistics

	Resume Score	Blind Written Score ($w(b)$)	Non-Written Score (nw)	Score Gap $nw - w(b)$	Final Score
Female Applicants	0.863	0.860	0.816	-0.044	0.871
Male Applicants	0.880	0.855	0.852	-0.004	0.892
Female $\neq$ Male?	No	No	Yes**	Yes***	Yes*
Obs.	51,809	51,809	51,809	51,809	51,809

Panel B. Job Process Statistics

	% Female Evaluators	# Candidates	# Female Candidates	# Evaluators
Job Process Average	46.1%	4.58	2.52	3.24

Notes: This table shows summary statistics of job applicants used in the hiring committee analysis in the paper. $w(b)$ represents blind written exams, nw represents non-written exams (interview, oral examinations, practical exams). Female $\neq$ Male? reports whether the sample statistics between men and women are statistically different than zero.

TABLE 7: Raw Hiring Probabilities By Committee Gender Composition

	$\mathbf{Pr}(Hired = 1)$			
	$<$ 30% Female Evaluators	$<$ 50% Female Evaluators	$>$ 50% Female Evaluators	$>$ 70% Female Evaluators
Female Applicants	0.25	0.46	0.33	0.33
Male Applicants	0.67	0.49	0.25	0.14
Female $\neq$ Male?	Yes**	No	No	Yes*
Obs.	30,701	38,004	22,500	10,800

Notes: This table reports raw hiring probabilities (raw data) of female and male candidates for various gender make-ups in the evaluation committees they face. Female $\neq$ Male? reports whether the hiring probabilities are different between the two groups.

TABLE 8: Effect of Committee Gender Composition on Gender Equity

	$\text{Score}^{nw} - \text{Score}^{w(b)}$					
	Overall	$<$ 50% Female	$>$ 50% Female	Overall	Final Score	$\mathbf{Pr}(Hired = 1)$
	(1)	(2)	(3)	(4)	(5)	(6)
Female_i	-0.023^{*}	-0.069^{***}	0.034^{*}			
	(0.014)	(0.017)	(0.018)			
$\text{Female}_i \times$				0.407^{***}	0.163^{***}	0.414^{***}
$\%\text{Female Evaluator}_c$				(0.102)	(0.052)	(0.138)
Job Applicant FE				X	X	X
Job Process FE	X	X	X	X	X	X
Obs.	60,504	38,004	22,500	9,901	9,901	9,901

Notes: This table plots estimates of the model $\text{Score}^{nw}_{icj} - \text{Score}^{w(b)}_{icj} = \beta\text{Female}_i + \gamma_c + \varepsilon_{icj}$ in columns (1) through (3). The regression captures the non-written penalty female candidates receive in the full sample (1), when being evaluated by committees with male-majority (2), and women-majority (3). Column (4) runs the augmented model $\text{Score}^{nw}_{icj} - \text{Score}^{w(b)}_{icj} = \beta\,(\text{Female}_i \times \%\text{Female Evaluator}_c) + \gamma_c + \mu_i + \varepsilon_{icj}$, where β captures evaluator bias in non-written exams. Columns (5) and (6) run the same model but with candidate final score and probability of receiving a job offer as outcomes. Standard errors are clustered at the job process level.

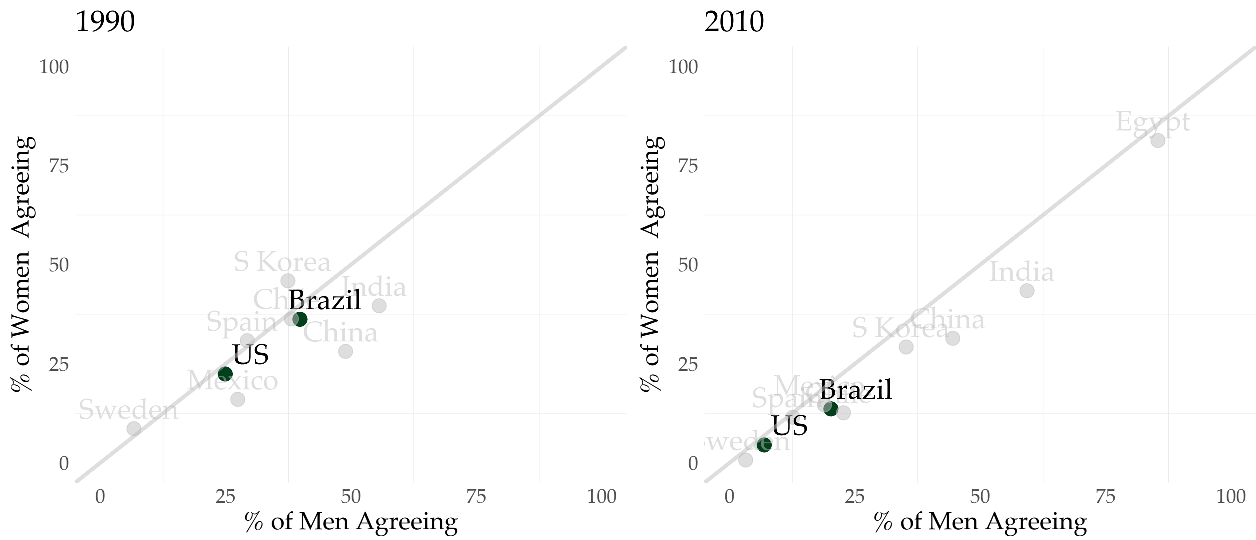
TABLE 9: Do Male Committee Members React to More Female Colleagues?

	Scores from Female Committee Member					Scores from Male Committee Member			
	nw	$w(b)$	$nw - w(b)$	Final Score		nw	$w(b)$	$nw - w(b)$	Final Score
	(1)	(2)	(3)	(4)		(5)	(6)	(7)	(8)
Female$_i$ ×	0.026	−0.003	0.008*	−0.012		0.073*	−0.035	0.139***	−0.010
%Female Evaluator$_c$	(0.042)	(0.032)	(0.042)	(0.029)		(0.039)	(0.031)	(0.039)	(0.024)
Committee Member FE	X	X	X	X	X	X	X	X	X
Obs.	60,504	60,504	60,504	60,504		60,504	60,504	60,504	60,504

Notes: This table compares, from columns (1) through (4), how female committee members score female candidates depending on different levels of female composition in the hiring committee. Columns (5) through (8) perform the same exercise but with scores from male committee members. Standard errors are clustered at the job process level.

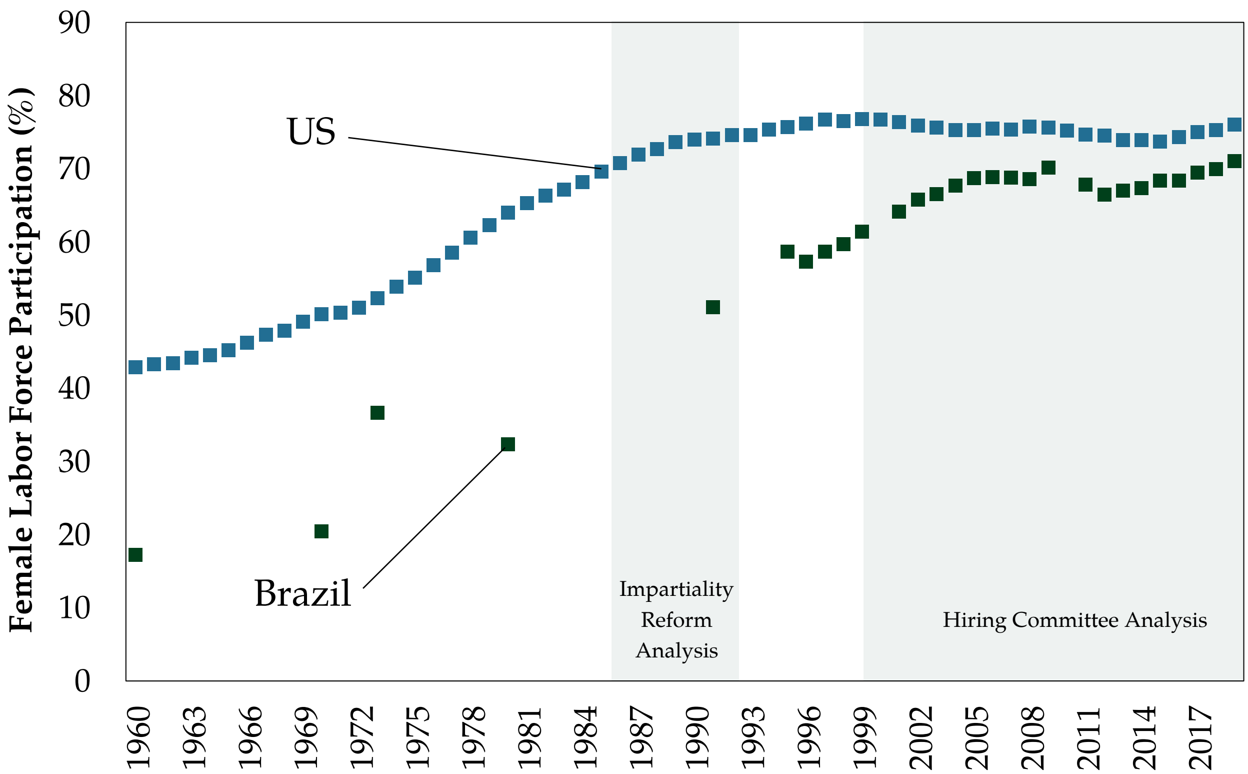
Appendix

A Appendix Tables and Figures



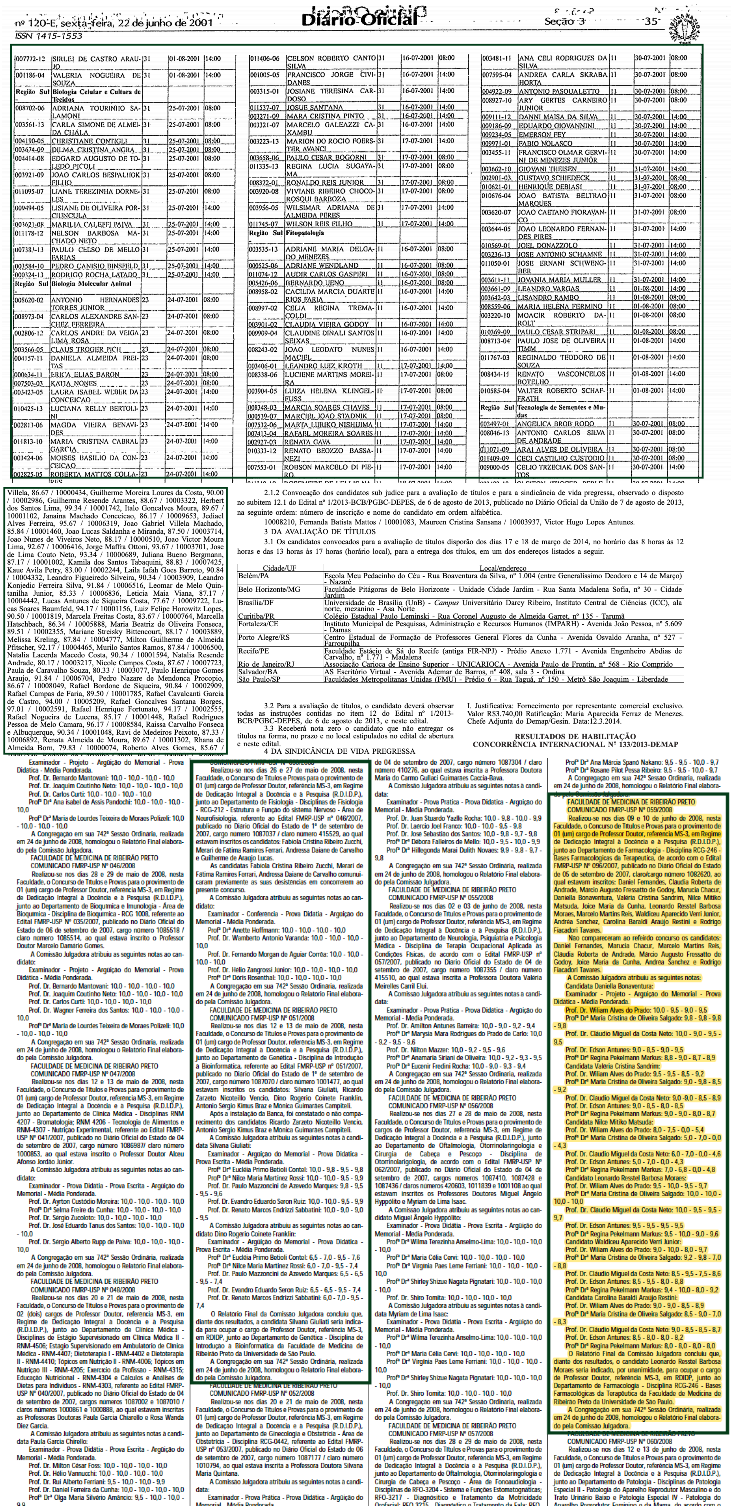
Notes: International Value Survey (IVS) answers for the 1990-1994 and 2010-2014 waves of women and men agreeing with the statement "*when jobs are scarce, men have more of a right to a job than women*". Countries plotted: South Korea, China, India, Brazil, US, Mexico, Spain, Sweden, and Egypt (for 2010 only).

FIGURE A.1: Gender Attitudes Across Countries



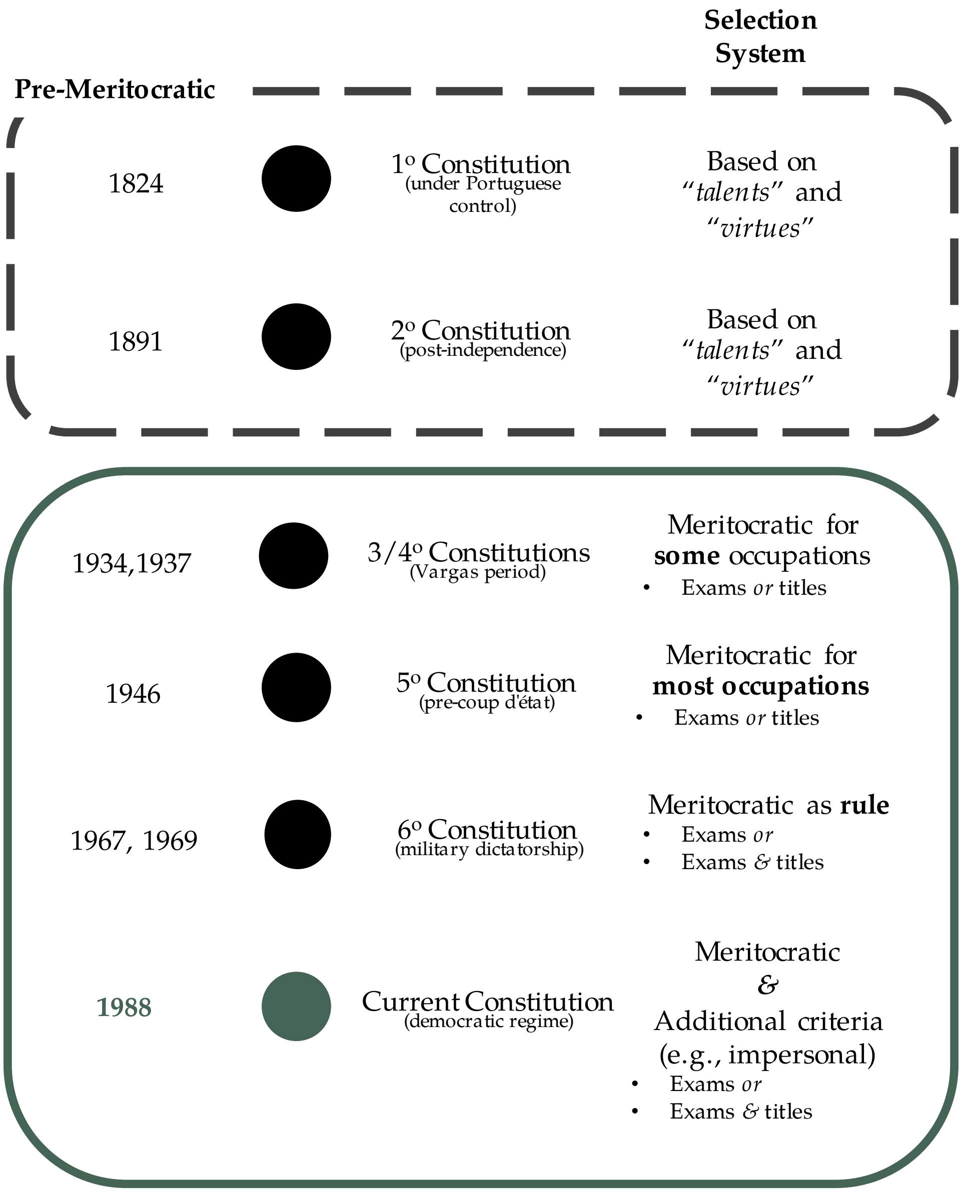
Notes: Female labor force participation rates (aged 25-54) for Brazil and the US. Shaded areas represent periods for different empirical analyses in the paper.

FIGURE A.2: Female labor force participation in Brazil and U.S.



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 during almost 50 years.

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



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.4: History of Changes in the Selection of Public Servants in Brazil

4.5. As provas escritas e prática terão a duração 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.5: Enforcing Blind Exams After Reform

TABLE A.1: 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**
Porto Velho	2007	Municipality (State Capital)
Manaus	2000	Municipality (State Capital)
Rio Branco		Municipality (State Capital)
Campo Grande	1998	Municipality (State Capital)
Macapá	2018	Municipality (State Capital)
Brasília		Municipality (State Capital)
Boa Vista	2010	Municipality (State Capital)
Cuiabá		Municipality (State Capital)
Palmas	2001	Municipality (State Capital)
São Paulo	**1975**	**Municipality (State Capital)**
Teresina	**1986**	**Municipality (State Capital)**
Rio de Janeiro		Municipality (State Capital)
Belém	2005	Municipality (State Capital)
Goiânia		Municipality (State Capital)
Salvador		Municipality (State Capital)
Florianópolis		Municipality (State Capital)
São Luís		Municipality (State Capital)
Maceió		Municipality (State Capital)
Porto Alegre		Municipality (State Capital)
Curitiba		Municipality (State Capital)
Belo Horizonte		Municipality (State Capital)
Fortaleza		Municipality (State Capital)
Recife		Municipality (State Capital)
João Pessoa		Municipality (State Capital)
Aracajú		Municipality (State Capital)
Natal		Municipality (State Capital)
Vitória		Municipality (State Capital)

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.2: Do Male Committee Members React to More Female Colleagues?

	Scores from Female Committee Member				Scores from Male Committee Member			
	nw	$w(b)$	$nw - w(b)$	Final Score	nw	$w(b)$	$nw - w(b)$	Final Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female$_i \times$	-0.095**	-0.038	-0.068*	-0.010	0.094***	0.006	0.108***	0.029
%Female Evaluator$_c$	(0.043)	(0.028)	(0.041)	(0.027)	(0.039)	(0.031)	(0.033)	(0.023)
Obs.	60,504	60,504	60,504	60,504	60,504	60,504	60,504	60,504

Notes: This table compares, from columns (1) through (4), how female committee members score female candidates depending on different levels of female composition in the hiring committee. Columns (5) through (8) perform the same exercise but with scores from male committee members. Standard errors are clustered at the job process level.