Designing Gender Equity: Evidence from Hiring Practices and Committees*

Tatiana Mocanu†

February 15, 2023

Abstract

This paper analyzes how different screening practices affect gender equity in hiring. I transform tens of millions of high-dimensional, unstructured records from Brazil's public sector into selection processes with detailed information on candidates, evaluators, screening tools, and scores. Exploiting a federal provision that required the use of more impartial hiring practices, I find that increasing screening impartiality improves women's evaluation scores, application rates, and probability of being hired. To understand which design choices reduce gender disparities, I combine variation in how job processes complied with the reform requirements with a model of hiring in which evaluator bias, tool bias, and screening precision jointly determine relative hiring outcomes by gender. Screening changes that limit discretion in existing hiring practices or add new impartial screening tools reduce the gender hiring gap by a third, while policies that eliminate subjective screening tools are ineffective because the loss of screening precision outweighs the reduction in evaluator bias. Finally, more gender-balanced hiring committees induce male evaluators to become more favorable toward female candidates in subjective stages.

Keywords: hiring, screening methods, gender discrimination, public sector
JEL Classification: M51, J16, J24, J45, J71.

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

†Stone Centre, University College London. Email: t.mocanu@ucl.ac.uk.

1 Introduction

In recent decades, firms have increasingly devoted resources to grapple with a lack of gender diversity and under-representation of women at various levels of the corporate ladder.[1] Even though screening and the selection of employees is a central part of every firm and organization, how to best design processes that are bias-free, improve employee diversity, and select the best candidates are questions that remain open. Certain hiring practices that are 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.

Answering these questions empirically is challenging because hiring decisions are effectively a black box. Employers are reluctant to share hiring practices or details on hiring processes, often engaging in lengthy legal battles to keep the information from going public. Moreover, 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 them determine gender disparities in labor market outcomes. I open the black box of hiring decisions by constructing uniquely-detailed information on the universe of selection processes from Brazil's public sector from 1980 to 2020. To access, extract, and transform these records, I develop a natural language processing algorithm that distills over 35 million official government text documents into data. This process generates a rich database detailing job applicant performance and evaluators' decision making process, including job openings and offers, applicant and manager identities, and candidate individual scores by screening method and manager. Equipped with these data, my analysis shows in three main parts that the implementation, design, and decision-making during the hiring stage are key sources of gender gaps.

I begin providing answers to the design of hiring practices by exploiting a reform in the provisions regulating the selection of public sector employees in Brazil's 1988 Constitution. The reform required government employers to conduct impersonal and impartial hiring processes, although only federal employers implemented it immediately. State governments conduct em-

[1] US firms alone spent more than $10 billion in 2003 in initiatives to reduce bias in recruiting (Hansen (2003)). Most strategies focus on diversity training programs that include debiasing, networking, and mentoring programs. Kalev et al. (2006) find no relationship between these programs and employee diversity in a sample of over 800 U.S. companies. Diversity training aimed at raising awareness about gender inequality can also backfire due to moral licensing (Bohnet (2016)).

ployee selection independently from central authority, and started addressing the legal changes necessary to implement impartial hiring processes only 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. Using written exams without concealing candidates' identity would be a clear violation of the new rules, leading employers to blind tests. At the same time, 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).

To quantify the overall effects of the reform, I first consider the average effect of the impartiality treatment in a binary difference-in-differences design. To establish the reform take-up, I analyze how the design of hiring processes changed in federal jobs relative to states. Federal employers responded sharply. Job announcements started including rules detailing written examinations that 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 that relied solely on non-written stages by 25 percentage points.

How did greater impartiality affect male and female job candidates? By comparing individual performance in job processes in the same occupation in federal and state governments, I estimate that following the reform, women's final scores increased by 0.07 standard deviation, accompanied by a decrease of a similar magnitude in men's scores. This resulted in a drop in the gender score gap of 0.14 standard deviation. Confirming that the reform induced intensive margin changes only in the scores of written exams — which had to be blinded — the gender gap in these stages also decreased, while relative scores in non-written tools between men and women had no statistically significant changes. This further indicates a lack of strategic response from evaluators conducting screening in exams with greater discretion as a response to the reform.

The decrease in the gender final score gap translates into improved hiring rates of women and a narrower gender hiring gap. 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 fully determine job offer decisions

following pre-determined rules — suggest that changes in the underlying mix of screening practices toward greater impartiality successfully reduced gender disparities originating from employers' behavior in the federal sector hiring.

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 job-seeker pools. For example, hiring practices perceived as unfair may discourage qualified minority candidates to apply. 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 higher employer demand for female candidates and female application rates result in a 13% increase in gender diversity among employees just a few years after the new Constitution came into effect.

The second part of the paper takes advantage of the different ways screening practices were modified to comply with the reform. Bureaucrats and legal aides at high organizational levels put forth occupation-specific guidelines with the changes in screening methods that employers should implement. As a consequence, over 95% of hiring processes within an occupation adopted only one change in screening methods. I then show that occupations with the same pre-reform mix of screening tools that implemented different changes had nearly identical pre-trends in gender gaps. Additionally, being assigned a specific set of new screening steps shows no relationship with characteristics such as degree of feminization, skill requirements, and selection process competitiveness.

My empirical strategy focuses on recovering counterfactuals based on several complier types not only with respect to the untreated group (no changes in screening methods) as in a standard difference-in-differences, but that compare treatment effects from alternative changes in screening methods for the same pre-reform screening mix. This variation allows me to disentangle between different forces driving gender gaps and answer design-relevant questions that include: Should an employer remove screening practices that entail 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?

To guide the understanding of how the key economic forces in each mix of hiring practices determine job applicant outcomes, I build on a classic statistical discrimination model by incorporating screening tool characteristics and the role of managers, who conduct the screening on behalf of the employer. This framework pins down different considerations that employers face when designing hiring processes. 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 levels of discretion due to their subjective nature, while the results from formal tests are more easily observable to the employer, 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 (e.g., Bohren et al. (2022)) — generating disparate impact even if managers are unbiased.

The first change in screening method I analyze is when employers screened candidates only using written tests and to achieve impartiality blind the exams. My estimate implies in a reduction in the gender hiring gap of 0.5 percentage point (relative to 1.5 p.p. pre-reform). In light of my model, this treatment effect isolates the complete removal of disparate treatment, since now evaluators cannot express disparate treatment, nor rely on statistical discrimination. This 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.

Next, I compare two alternative changes to an employer who only screened using non-written methods — mainly interviews and oral exams. In this context with high discretion, the initial gender hiring gap is almost 17 percentage points. For this group to comply with the reform, the first option was to replace the non-written stage with a blind written test. The substitution improves female hiring odds by 7 percentage points relative to men. This large treatment effect suggests that either written exams have higher precision or a smaller disparate impact than non-written tests, or that the combined magnitude of these channels is small relative to the size of the evaluator bias in interviews.

The second change to this initial mix of screening methods involves keeping the subjective tools, but adding a blind written exam to increase the overall objectivity of the hiring process. This increases screening precision, which helps minority candidates both directly — by providing an additional productivity signal — and by diluting the contribution from the existing interview signal — still subject to group-based priors and disparate treatment. Even if the written exam imposes some disparate impact, as long as its tool bias is low relative to that of the pre-existing interview, the minority group should benefit from the addition of the blind test. In line with that, 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.

The last set of comparisons I make illustrate the potential shortcomings of well-intentioned design changes to improve diversity. Employers with an initial mix of written and non-written tests who only blinded the written stage had on average no improvement in female outcomes relative to men, except when the pre-determined weight of the blind exam toward the final score is large enough. Alternatively, employers could blind the written test but remove the interview. This treatment also generates no reduction of gender hiring gaps. By

removing a screening stage, the total precision of the job process decreases, which lowers the hiring rate of women since now evaluators put more weight on their group-based priors about candidate ability. The null result I find indicates that the screening precision of interviews must be sufficiently large in order to offset the removal of evaluator and tool bias, which suggests that statistical discrimination is more important than evaluator bias at least along this margin.

Several practical lessons emerge from the second set of results. 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 where the tools being employed provide relatively objective signals and limit bias expression. However, blinding alone an existing test may not be enough to improve gender diversity. Third, despite its potential disparate impact, the introduction of a blind test generates gains in screening precision that narrow gender gaps the most. Finally, removing subjective tests fails to improve women's outcomes, suggesting that employers should carefully weigh information loss and net gains from bias reduction.

In the third and final part of the paper, I study a complementary approach to improving gender equity in hiring: changing the mix of decision makers. While changing screening tools can lead to significant advances toward diversity, this approach may be impractical in certain cases and possibly backfire when employers have limited information. For example, knowledge on the relative disparate impact and precision between written and non-written exams is important to avoid removing information that benefits the minority group. Blinding may even decrease efficiency if statistical discrimination is accurate and there is no evaluator bias.

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 by incorporating a penalty function that depends on the hiring committee composition.[2] 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 study how changes in the gender composition of committees affect female and male candidates. Committee members and candidates match in a quasi-double-blind mechanism — evaluators are disclosed after candidates apply, but chosen prior to the public announcement. In line with institutional features, I find little evidence of systematic association between the gender composition of committees and blind (written) or CV scores, applicant pool size, or number of female applicants.

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

[2]The analysis follows the same logic as that of a series of corporate and public policies incentivizing or enforcing more diverse committees, such as gender quotas on boards (Bertrand et al. (2019)).

on resumes and blind exams, but are scored on average 4 percentage points less than men on non-written exams. To separate out the confounding effect of individual differences in skills between the two types of screening tools, I use a triple-differences strategy, comparing gaps between non-written and blind-written scores of the same candidate across job processes within the same employer 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, thus improving their final scores and chances of job offers.

To better understand the forces driving the reduction in biased evaluations of female candidates when the hiring committee has more female members, I analyze how the same male evaluator scores women when he participates in hiring committees with different shares of female colleagues. I find that as committees include more women evaluators, men increase their non-written scores given to women relative to male candidates. This effect does not appear in blind exam scores and implies a decrease in evaluator bias of about 1.4 percentage points.

Why do more women in the hiring committee change men's behavior? I rule out two hypotheses — stereotype threat and gender differences in candidate attribute screening — and find evidence consistent with increased norms-based costs induced, for example, by changes in group gender norms (Akerlof and Kranton (2000), Field et al. (2021)). The first hypothesis I rule out, stereotype threat (Steele (1997)), posits that female candidate's actual performance changes depending on the composition of hiring committees. However, contrary to male evaluators whose scoring of women changes depending on the screening context, I find no evidence of behavioral responses of female evaluators as the gender make-up of committees varies.

Another potential explanation is that women screen for a different set of candidate characteristics when giving marks during interviews relative to male colleagues. If these characteristics disproportionately favor female candidates, more women in the hiring committee could change the group discussion dynamics by stressing such abilities then neglected by male evaluators.[3] Contrary to this hypothesis, I find that female evaluators are harsher on female job applicants, scoring women 2 percentage points lower relative to male candidates than male colleagues. This is consistent with previous results in other settings showing that women produce less favorable results to other women when compared to men (e.g., Miller and Sutherland (2022), Bagues and Esteve-Volart (2010)).

In contrast, the previous sets of results are consistent with the idea that expressing bias against female candidates becomes more costly as the committee minority share increases. This could be attributed to a change in gender group norms imposing a norms-based cost. This view is also consistent with the non-monotonic relationship I find between the female share

[3]Even though hiring members evaluate each candidate independently, they are allowed to share their opinions on candidates' performance, potentially changing their scores before final submission.

of committee members and women's scores. As committees become more gender-balanced, male evaluators reduce their gender bias expression in non-written exams, equalizing hiring outcomes between male and female applicants. As long as hiring committees remain gender-imbalanced toward male members, additional women increase the behavioral response from male evaluators beyond simply adding one female evaluator to the group. However, once committees are female-dominated, this reduction in men's bias is eventually offset by lower scores from female evaluators to women applicants.

1.1 Related Literature

The first contribution of this paper is to causally identify the roles of discriminatory individuals and practices and which screening designs mitigate gender disparities in hiring. Previous important work by Goldin and Rouse (2000) and a large number of papers using audit and correspondence studies (e.g., Neumark (1996), Bertrand and Mullainathan (2004), Kline et al. (2022)) has investigated the existence of discrimination in hiring. Either because of data limitations — the hiring black box — or experimental design, this literature primarily focused on documenting the presence and extent of discrimination, falling short of determining the sources of hiring discrimination and how they should be addressed.

My second contribution is to quantify the extent of evaluator bias, statistical discrimination, and screening tool bias and how they interact with job attributes using a wide set of occupations. Different strands in the literature have studied separately the role of specific screening interventions. One line of work focuses on the effects of hiding candidates' identity, starting with Goldin and Rouse (2000) who show that blind auditions in orchestras increase the likelihood that women musicians are hired.[4] Another strand has looked at how introducing testing in low-skill jobs affects minorities (Autor and Scarborough (2008) and Hoffman et al. (2018)).[5] To flesh out the relative importance of different sources driving disparities, one needs a rich set of simultaneous interventions as generated by the impartiality reform. Moreover, the large set of occupations in Brazil's public sector provides lessons on how these effects can differ depending, for example, on the occupation's skill level or feminization.

This paper also contributes to the existing empirical evidence on the importance of evaluator's gender documented in non-hiring settings (Sarsons (2019), Broder (1993), Card et al.

[4]More recent papers include the study of anonymized CVs (Behaghel et al. (2015), Krause et al. (2012), Åslund and Skans (2012)), which have been limited to settings in which firms self-select into these programs, providing mixed results. Several other papers have studied different consequences of partially concealing or including some information about job applicants, e.g., age (Neumark (2021)), credit information (Bartik and Nelson (2022)), criminal records and history checks (Holzer et al. (2006), Agan and Starr (2018), Doleac and Hansen (2020)).

[5]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.

(2020), De Paola and Scoppa (2015), Bagues et al. (2017), Bagues and Esteve-Volart (2010), Lavy (2008)). This body of work has offered results ranging from the gender of who evaluates having no effect, to female and male evaluators judging women more harshly or less harshly. My findings show that knowledge on the underlying degree of discretion allowed to evaluators is crucial to interpret decision outputs from committees. They also provide guidance on how to design hiring committees and implement screening tools that curb evaluators' bias expression.

This paper further relates to the growing literature on personnel economics of state that has studied how governments can change the applicant pool (e.g. Dal Bó et al. (2013), Ashraf et al. (2020), Deserranno (2019)). However, given the large public sector premium in many countries and the fact 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.[6]

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

2 Institutional Details and Setting

2.1 Overview

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 im-

[6]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 (2021a), Moreira and Pérez (2021b)). This paper examines how changing screening methods within a meritocratic system affects labor market outcomes.

portant, as public sector employees receive automatic life-time tenure after being hired and termination is only possible following 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 offer a significant premium compared to the private sector, and generally compound on a time-in-office basis and mechanically by inflation. As a result, public sector jobs are highly competitive, with an average probability of being hired of around 4%.

2.2 Public Servant Selection

Brazil was the first country in Latin America to establish a formal, merit-based career civil service. It is considered a primary example of a meritocratic and legally professionalized civil service system (see Grindle (2012) and Figure A.3 for a 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 to 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).[7]

Each job selection process follows the same general steps depicted in Figure 1. The first posting regarding a hiring process — 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, every job process must follow the general guidelines prescribed in the Constitution and must integrally respect the rules laid out in the job announcement.[8]

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. The timing of job announcements 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 manifest intent in filling out or expanding specific job titles to the appropriate oversight budget and comptroller offices, which then decides whether the job posting should be greenlighted.

[7] Some public sector jobs are exempt from 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.

[8] 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.

The hiring process proceeds as follows. First, 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 in its own government gazette individual performance (scores) on each selection stage as the hiring process unfolds, including interviews and tests, and identifying the candidates who are ultimately offered jobs, wait-listed, and hired.

3 Opening the Hiring Black Box: Data Extraction

3.1 Raw Text Sources

The raw data used in this paper come from over 35 million of official journal pages of federal, state, and local governments in Brazil (known as "*Diário Oficial*") from 1980 until 2020. These gazettes are similar to the Federal Register in the US and publish the universe of public notices spanning public procurement processes, executive orders, and information on public servants. Such notices on public sector personnel include the entirety of every public sector employee hiring process (as shown in Figure 1) and 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 the relevant text.*
3. *Identify the underlying job process of a certain relevant text.*
4. *Link different postings belonging to the same process.*
5. *Transform text in each posting into data.*

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

Given that one could identify and link the precise text domain of all stages of a hiring process, extracting data from the raw text presents an even bigger challenge. There is no 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.[9]

3.2 A New Approach to Transform Unstructured Text Into Data

To address all of these challenges, I develop a two-step natural language processing algorithm that allows me to first define the relevant text portions from highly confounding text, attribute a posting to a unique job hiring process 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 of 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).

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

[9] See Figure A.5 for some examples. I document over 200 different text layouts, with multiple variations within the same broad layout type.

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

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^t stems from its search-query use. For each government gazette issue, I search for a job posting notice, using a combination of words in the same paragraph (formally defined as some text string neighborhood) comprised of "announcement", "job", "hiring", and "posting". When there is one or more hits, I bound the relevant text to each job announcement and extract attributes A (the implementation of relevant text boundaries is detailed below). Only the *release date* is ex-ante known, since I know

when each journal issue is published. To correctly identify the terms containing the other attributes in A, I rely on ad hoc dictionaries and allow them to expand by "learning" new terms.

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

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 leveraging 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.

In addition, 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 $\mathcal{T} \rightarrow \{M_1(\tau), M_2(\tau), \cdots, M_K(\tau)\}$, where τ is a transformation of the underlying raw text. Such mappings include sentiment-based approaches as in Gentzkow

et al. (2019), where the true sentiment of a message is transformed into a function of a latent quantity.

In many applications, however, researchers might be interested in extracting exact information from text and converting that into a database by distilling $\mathcal{T}$ into a pre-determined list of variables $\{x_1, x_2, \cdots, x_K\}$. This is usually an extremely time-intensive task, highly dependent on the particular context and that relies heavily on strong prior information about the potential variations of text structure across $\mathcal{T}$. 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.

To solve this issue, I leverage the semantic structure implied by the relationship between listed variables, so that a fixed variable x_1 conditions all other data points that a researcher is interested in extracting, $\{x_1, x_2|x_1, \cdots, x_K|x_1\}$. To see how, let x_1^i correspond to candidate i's exam score, x_2^i her name, x_3^i the exam type and x_4^i the committee member who gave score x_1^i. In order to deal with the unstructured nature of the text, I start by targeting text tokens containing numbers, which is the only morphology that maps onto exam scores. Of course, many numbers within the text might be extraneous and not represent scores. The next step searches for tokens in the neighborhood of every number that match the characteristics of each additional variable x. This both fully defines the other variables that relate to x_1 and filters out numeric elements that are not scores.[10]

By choosing one variable to which most or all of the other desired variables relate, this approach avoids reliance on information from semantic structure that differs across public announcements, and focuses on relationships that organize each candidate's relevant information in the same way within a job notice text. The underlying semantic structure thereby informs the selection model about the location of certain variables rather than feed a label grouping, such as political slant or favorability of a review. This step requires the use of few *ad hoc* dictionaries (a list with Brazilian names in the current application and another with different examination types), which are allowed to learn similarly to before with the lists of occupations and employer names.

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

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

Implementation. The first step 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 keys, 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 court injunctions on behalf of candidates' legal actions). 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.

I model the algorithm's pipeline that implements step 2 following the sequence in Figure 3. Implementation performance can be thought as an inference problem: all text snippets extracted in the first step contain the true data, but also false positives (incorrect signals). To ensure the final dataset contain only correct information, the algorithm first maximizes stringency and therefore the number of false negatives — it throws away important information that should be part of final data. Then, to recover lost information, I increasingly relax the search parameter's stringency.

Without computation constraints, this iterative process would continue until gains in data extracted between two iterations is negligible. At a given point in the pipeline, my data dimensionality spans anywhere from 8 to 60 billion characters, implying more than 2 trillion matching computations needed every time the NLP algorithm runs. To make extraction feasible, I continue to minimize false negatives in the data output until the final data has an accuracy above 90% and the cost of decreasing false negatives is greater than some increase in the number of observations in the final data (e.g., a manual adjustment to the algorithm that takes 1 hour produces 5 additional job processes that were otherwise discarded). The final estimating sample contains 86,959 candidates. Figure A.6 shows validation exercises with sample statistics generated from the extracted data and official aggregate statistics given by Brazil's federal government.

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.

4.1 The Impartiality Reform: Description

In October 1988, Brazil passed a new Federal Constitution in the wake of the end of several decades under military regimes. 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. In addition, it added 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, such as cheating on screening tests, another practice previously deemed 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 a specific candidate, or someone from a specific 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, resulting in the blinding of these exams. However, determining 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. 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, federal sector employers decreased their reliance on non-written stages, either by reducing their relative number with respect to blind practices or by removing them completely.

In principle, the provisions in the new Constitution applied to public sector hiring at all government levels. However, because public servant selection is conducted by states and municipalities independently of the central authority, states had to pass the appropriate legal frameworks 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. As a result, it took several years for the sharp shift in federal employer behavior with respect to hiring to trickle down to state agencies and governments.[11]

Among 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, their potential consequences to my setting are limited. First, my research design partial out any common time effects. Second, the only changes related to public sector hiring were the new principles that I discussed. Finally,

[11] 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 other practices to comply with the impartiality and other guidelines in the Constitution.

it is important to stress that since the impartiality requirement was not a diversity-oriented reform, but part of a much larger civic redesign.

4.2 Sample Selection and Data Patterns

For the analysis centering on the impartiality reform policy, I restrict my estimating sample to the years 1986 through 1991 to the federal government level and to the states with official gazette issues available online for the period. These states were Amazonas in the country's north region, Pernambuco in the northeast, Distrito Federal, Mato Grosso, and Mato Grosso do Sul in the central region, São Paulo — the largest and richest state — in the southeast, and Rio Grande do Sul in the south. I use all job processes with complete information on job requirements, screening steps, as well as candidate scores, final ranks, and job offers, if any.[12]

I focus the analysis on 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 and Table 1 provides summary statistics on education requirements, screening steps, and job applicants for control and treated groups, before and after the reform.

4.3 Did the Reform Change Screening Practices?

Due to the nature of the shock to hiring practices that I study, it is crucial to preface my empirical analysis by evaluating the extent to which 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 estimating regressions of the form:

$$y_{ct} = \delta_{o(c)} + \gamma \left(\text{Fed}_{o(c)} \times \text{Post}_{o(c),t} \right) + \theta_t + u_{ct} \tag{1}$$

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 after 1988 and $\text{Fed}_{o(c)}$ if it selects employees for the federal government. Comparing similar occupations between treated and control groups is important to net out composition differences between aggregate jobs

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

postings 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 and enforcement agencies). The parallel trends identifying assumption in model (1) is satisfied if, absent the reform, average outcomes y_{ct} would have followed parallel paths over time.

Table 2 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 became blind exams) relative to state job processes used as control. To gauge the importance of the 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) reflect similar exercises, 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.

The above first-stage estimates indicate sharp changes in the mix of screening tools used in federal sector hiring processes relative to those in the control group. Although I discuss in detail the different treatment groups giving rise to each estimated effect in Table 2 later in the paper, it is useful to shed some light on which 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 (column (1)).

These different combinations of changes in screening methods toward greater impartiality generate apparent non-perfect compliance rates in each individual regression. Taken together, these estimated effects all represent policy-compliant changes, and, as I later show, are largely determined by occupation.[13]

[13]Figure A.4 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.

4.4 Binary Difference-in-Differences Design

I begin my empirical analysis by first assessing broadly whether greater impartiality in hiring affected female applicants differentially from men. The starting point exploits the immediate compliance with the introduction of impartiality in public servant selection by federal government employers, together with a lagged and slow adoption by state-level employers. To assess the effects of the reform on gender gaps in several labor market outcomes, I run:

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

where y_{it} represents candidate i's job process outcomes, β estimates the differential effect of greater hiring impartiality on women relative to men, while controlling for year and job announcement occupation fixed effects. In all specifications, standard errors are clustered at the job process level. Only job processes with at least one male and one female applicants, with known job offers, and that consistently appear before and after the policy in both groups are kept in the estimating sample. I assign candidates' gender using Brazil's Census Bureau *Gender of Names* database, which contains nearly 200,000 unique first names and their corresponding gender. The matching precision between job applicant names and gender from this dictionary is above 98%.

4.5 Effects on Gender Hiring and Score Gaps

To better organize the results in this section, I first look at whether the impartiality reform narrowed the gender hiring gap. Specifically, 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 the 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 "adept" candidate outside the initial offer list receives the position.

Columns (1) and (2) in Table 3 show that the probability of being hired for women and men, respectively, go in opposite directions after the impartiality reform 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 in 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. Combined, the point-estimates imply a reduction of the gender hiring gap of 0.7 percentage points, equivalent to 44% of the initial hiring gap. 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.

An advantage from using 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 marginally hired man in measured performance. Therefore, with a less coarse outcome such as scores, more subtle responses to the reform can be captured.

Before using final scores, however, I check whether they actually determine job offers. Figure A.7 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 a 60% chance of receiving an offer. This is not surprising — according to the rules, hiring decisions are made exclusively in accordance with the ordering of candidates' final results, and even top scorers are not guaranteed job offers since job openings are generally fixed.

Table 4 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. The final scores of women increase by 0.07 standard deviations after the new Constitution is implemented, with the final scores of men decrease by slightly more. 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 A.8, where final score gender gaps remain unchanged for candidates in state job processes and the gap significantly narrows for federal jobs.

Figure 4 unveils in more detail how gender score gaps behaved in treated and control groups dynamically. In the first part of the figure, gender gaps in the control group remain flat both before and after the reform, suggesting an absence of spillover effects from the reform adoption in the federal sector on state jobs. Alternatively, this also indicates no changes to the outcome path in state job processes, underscoring the lagged implementation of the impartiality principle by those government levels. The second part of the figure plots difference-in-

differences estimates of the gender final score gap, both showing no pre-trends and a sharp response in the outcome after the reform.

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 Table 2 shows, 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 5, 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 4 shows that the written scores of men 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 concealing 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 the addition of blind written tests — 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).

Overall, greater impartiality in hiring substantially narrowed the observable gender performance gap, primarily because women's written scores improved and men's were significantly decreased once candidate identities were concealed. The improvement in final scores in turn induced more women being hired relative to men.

4.6 Threats to Identification and Interpretation of Estimates

The estimation of parameter β in equation 2 nets out common non-discriminatory effects of greater impartiality on male and female candidates. However, there are two sources of potential issues to causally linking measured effects on gender gaps to a decrease in partiality after the reform is implemented. To see why, let a candidate with gender $x = \{m, f\}$ have ability $\alpha_i(x)$. If the average ability of the candidate pool changes before and after the reform, $\mathbb{E}^{t>1988}[\alpha_i(x)] \neq \mathbb{E}^{t<1989}[\alpha_i(x)]$, the estimated effects attributed to greater screening impartiality could simply reflect quality effects (e.g., higher ability women applying after the reform).

A second, more nuanced potential threat to identification arises if greater impartiality affects women's performance differently than men's, $\mathbb{E}[\alpha_{it>1988}(f) - \alpha_{it<1989}(f)] \neq \mathbb{E}[\alpha_{it>1988}(m) - \alpha_{it<1989}(m)]$, where $\alpha_{it}(x)$ may change over time because of dampening effects. This would be the case, for example, if women become less nervous when tests conceal

gender — perhaps due to stereotype threats — and perform better in written exams after the reform. Observationally, in both examples women would have greater ability improvement than men, leading to an upward bias in the estimated effect attributed to the reform.

How can one address each identification threat? The strategy in the first case is straightforward: re-estimate β while holding the average candidate quality pool constant over time. This can be achieved by including applicant fixed effects to model 2. Table A.2 replicates the results in Table 4 for final scores controlling for time-invariant individual ability. The estimated effect of the increase in women's scores is larger than in the specification that allows for composition changes, while the effects for men are more negative. These results imply that the impartiality reform indeed narrowed the gender gap. Additionally, comparing the coefficients in both specifications, suggests that the average quality of the female applicant pool decreased after the reform, with the opposite effect on men's average quality.

Addressing the second potential issue to identification is more challenging. In Section 7, I indirectly test for the plausibility that the underlying context in which the screening method is implemented affects women differently than men. Performance of female candidates relative to men in blind exams and interviews when the committee gender composition changes remains stable, even when controlling for committee member fixed effects.

Lastly, another consideration involves not the identification of β but its interpretation. Specifically, that my estimated effects could be attributable to gender differences in connections — men having deeper and broader networks than women (Cullen and Perez-Truglia (2022), Ductor et al. (2021)) and being able to exploit them before the reform. Of course, after the new Constitution their ability to leverage these connections would be limited. I consider this channel to be unlikely in practice, however, due to two reasons. First, if connected evaluators can no longer affect the written exam score after blinding, they would likely compensate in e.g. interview stages. Contrary to that, I find in Table 4 that other scoring in non-blind stages remains the same after the reform. Second, if these networks are gendered in nature, one would expect female (male) evaluators to score women (men) more favorably. Again in lieu of that, in Section 7 I show that women score other women more harshly relative to male colleagues.

4.7 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 typically unobservable in other settings, since

calculating the conditional hiring probability requires observing *all* applicants, not only the hired pool.

Knowledge of 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 refrained from applying. With respect to gender, previous studies have presented evidence on several fronts suggesting that women may be less likely to apply for 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. (2023)), 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' use of biased screening tools is likely to interact with these factors, further magnifying barriers to extensive-margin responses of female candidates. Simply implementing more impartial screening may motivate more female candidates to apply, as even the perception of fairer treatment could be consequential in shaping minorities' behavior (Small and Pager (2020)).[14] Supply-side factors are important because suboptimal entry by high-performing women is costly to firms and without observing application rates, the effectiveness of enforcement of anti-discrimination laws cannot be fully assessed.

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: (i) the potential propensity of the employer or hiring committee to discriminate (under a set of assumptions) against women, and (ii) 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

[14] Women are more likely to place greater weight than men on fair treatment, and the perception of fair treatment is more strongly linked to women's than to men's willingness to apply at a previously rejecting firm (Brands and Fernandez-Mateo (2017)).

channel, which measures differences in application rates as:

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

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 reform ($\mathbf{Pr}(\text{Female} = 1)$). 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 1.5 percentage points (net of occupation effects), the drop in the hiring gap implied 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 5, I run job process level versions of the binary difference-in-differences 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 $\mathbf{Pr}(\text{Female}|Hired = 1)$ in column (1) and $\mathbf{Pr}(\text{Female} = 1)$ in column (2). Moreover, note that latter effect — about a 6% growth in the female application rate — is statistically indistinguishable from that estimated in column (4) of Table 3, of around 7% of the pre-treatment level.

Next, column (3) shows that the number of applicants decreases after the new Constitution is introduced, although not a statistically significant effect. This may be driven by candidates' perception of an increase in the cost of the job process, for example because of more screening rounds. I formally investigate this possibility in Section 6, where I distinguish between the treatments that added hiring stages from treatments that only blinded existing ones. Another potential driver is the number of job openings, which is positively correlated with the size of the candidate pool, and is determined by budgetary and personnel management constraints. The reduction in the number of men applying, albeit not statistically significant, is 10 percentage points larger than that for women, accounting for the increase in the probability of an applicant being female.

4.8 Differences Across Skills and Feminization

The magnitude of the treatment effect of the impartiality reform may depend on characteristics related to the occupation. Along educational requirements, higher-skill occupations may employ more screening steps, improving screening precision. Easier productivity observability diminishes reliance on statistical discrimination. Moreover, thinner markets in highly-specialized high-skill jobs make taste-based discrimination more costly. In contrast, women may face more stereotypes for higher-skill jobs at the top of the career ladder, for example, stereotypes against women's leadership abilities.

Table 6 examines how the effect of increasing screening impartiality varies across the skill level required in job postings. Compared to male candidates in high-skilled occupations (college degree or more), female scores increase by 0.2 standard deviation, and their hiring rates by 1.1 percentage points. This represents a magnitude 50% greater than the average effect estimated across all skill levels. Increasing impartiality for low-skill jobs — that require either high school or less as education — does not have an effect on scores or hiring rates of women. This can be due to screening tools varying in productivity signals generated and their precision in high-skill compared to low-skill jobs. For example, screening tools that give evaluators more discretion, such as interviews, may be more valuable in higher-skilled settings as private signals observed by evaluators may be more important to determine worker quality relative to lower-skill settings. As a consequence, the introduction of blind testing and removal of interviews can have different effects across the skill distribution.

Table 7 assesses how the effect of greater impartiality differs by the feminization rate of the occupation. Here, I take a comprehensive view of gender "identity" in an occupation. To assign a job title to one of the three groups: female-dominated, neutral, or male-dominated, I consider the gender segregation of that occupation in the public and private (when applicable) sectors, the total share of women applying to job openings of that occupation, and when the occupation requires a specific college degree (e.g., structural engineering), the national gender make-up in the major. These factors broadly align. I define occupations with less than 40% female participation to be male-dominated, 40-60% neutral, and above 60% female-dominated.

Comparing columns (A1) and (A3) reveals that the final score gap narrows by approximately 0.3 standard deviations for both female- and male-dominated occupations. Investigating the effects for male and female applicants separately in panels (B) and (C) shows that female candidates' scores in male-dominated occupations increase by more than in feminized jobs. This indicates that the impartiality reform was more beneficial to women in male-dominated hiring environments. Conducting the same exercise for male candidates unveils equally striking effects. After the reform, men's scores fall more in feminized occupations than in male-

dominated, but their correction is larger than women's:

$$|\widehat{\beta}^{F}\left[\text{Female}=0\right]| - |\widehat{\beta}^{M}\left[\text{Female}=1\right]| \approx 0.09$$

where $|\widehat{\beta}^{F}\left[\text{Female}=0\right]|$ is the estimated effect on men in feminized occupations. One interpretation to the difference above is that the benefit men derived in female-dominated occupations was larger than the penalty women faced in male-dominated sectors.

The female hiring gap also decreases in male-dominated occupations, but by less than in feminized jobs. Investigating outcomes for each gender separately reveals similar patterns to the final scores. Men are favored more in female occupations than women disfavored in male-dominated ones. But in this case, the improvement in female hiring rates is almost entirely driven by men's hiring odds decreasing in the least feminized jobs rather than women's chances also increasing. These results imply a pass-through rate between final scores and individual hiring odds from increasing hiring impartiality of $\frac{|\widehat{\beta}^{Hired,F}\left[\text{Female}=0\right]| - |\widehat{\beta}^{Hired,M}\left[\text{Female}=1\right]|}{|\widehat{\beta}^{Score,F}\left[\text{Female}=0\right]| - |\widehat{\beta}^{Score,M}\left[\text{Female}=1\right]|} \approx 1/9$. Combining the demand and supply side channels, Table 8 shows that the share of female hires increases more in male-dominated occupations than other types of jobs.

5 Conceptual Framework

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

The first ingredient in the model is hiring manager bias. Managers have the task of selecting employees with a mix of screening tools delegated to them by the employer. I allow managers to have a systematic bias for a certain demographic group. The term could be interpreted as taste-based discrimination, as it effectively captures a utility disamenity from hiring some group, as well as implicit or any other source of unintentional bias. However, this bias can only be *expressed* to the extent that the screening tool used enables discretion. In contrast, statistical discrimination expression in the model is independent from the degree of discretion of screening practices used. Managers base their prior of a candidate's productivity on her

group membership. When candidate identity is concealed, they instead resort to the population mean.

The second addition I make is to model the possibility of screening practices themselves to be biased. Independently of the behavior of a hiring manager, certain screening tools may disadvantage a particular group. For example, if written tests reward risky behavior by penalizing wrong answers without measuring productivity, women may be disadvantaged and the screening practice would lead to a disparate impact.[15] Finally, by maintaining screening precision in the model, the role of tool bias is equivalent to adding systematic 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 types 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 for gender equity.

5.1 Environment

An employer (the principal) delegates the screening of a pool of job applicants to 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. Given that I consider women to be the minority group, women's average productivity is perceived to be lower than men's, $\mu_0(f) < \mu_0(m)$.[16]

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

[16] Aigner and Cain (1977), Lundberg and Startz (1983), Cornell and Welch (1996), and Bartik and Nelson (2022) 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 that I introduce in the model.

The employer's objective is to hire a proportion K of workers that maximize expected productivity.[17] 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 that evaluators face by expressing bias, i.e., reporting to the employer a value of a candidate's measured performance that differs from the signal provided by the screening tool. Intuitively, this cost increases in the objectivity of the screening signal. Scoring a candidate's written test differently than the publicly-observable signal poses a much higher threat of detection than underscoring someone after an interview because the person did not appear to be friendly or an "appropriate fit".

The 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 behaving in biased ways 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 the 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.[18] After the reform, 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.

Hiring Rates With Written Exams When the hiring technology only includes written tests, the

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

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

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

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

where s represents the unbiased signal $s = y + \varepsilon_s$, h_s is the inverse of the variance of the written signal, measuring the precision of written testing and independent of group membership x.[19] $\nu_s(x)$ captures the disparate impact of the screening tool, which favors men when $\nu_s(m) > \nu_s(f)$.[20] After observing s^*, the hiring manager updates her assessment of expected productivity of candidates, initially based on group productivity, $\mu_0(x)$, forming the posterior:

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

The expression above represents a weighted average of perceived productivity of group x and written signal provided by the written test, with weights determined by the relative precision of the signal with respect to productivity dispersion. A direct implication from the updated group mean $\mu(x, s)$ is that when written tests are less informative, hiring evaluators rely more on the group prior.

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. For the detailed solution of the hiring threshold, $z^*_\theta(x)$, as well as all the detailed solution for all cases below, please see Appendix B. Due to the linear form of the signal expression, the hiring threshold for group x decreases when the group mean productivity is higher, the tool's bias favors the group, or when evaluators are biased toward x (given the discretion in written exams).

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

[19] If test signals are mean-biased, suppose that employer's prior is mean consistent with the information given by the test and therefore $Y \sim N(\mu_0(x) + \nu_s(x), 1/h_0)$.

[20] 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.

than written tests: $d_\eta > d_s$.

Hiring Rates With Written and Non-Written Exams Given the two signals previously described, η^* and s^*, and the perceived group productivity, $\mu_0(x)$, the hiring manager updates her assessment of expected productivity taking into account both exam signals:

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

where the overall screening precision is $h_T \equiv h_0 + h_\eta + h_s$. With two screening tools, evaluators place less weight on their group priors, which favors the hiring threshold of the minority group if $\mu_0(m) > \mu_0(f)$. Moreover, the overall bias now captures bias from both tools.

Hiring Rates With Blind Written Exam After the impartiality reform of 1988, federal employers using written exams as screening tools had to conceal candidates' identity. Within the model, blinding makes it impossible to assign individual candidates to a group, since hiring evaluators cannot observe whether a certain signal is generated by a male or female candidate. Let the blind written signal be defined as $b^* = y + \nu_s(x) + \varepsilon_s$, with $\varepsilon \sim N(0, 1/h_s)$. Note that the blind written exam has the same screening precision h_s and the same disparate impact $\nu_s(x)$ as the written test previously modeled.

When the screening technology includes blind written tests, the evaluator's objective becomes:

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

as discretion is entirely removed from the screening tool. Additionally, blinding the written test affects how the evaluator updates perceived candidate productivity, using the written signal, s, and the perceived *population* productivity, $\mu_0 = \frac{\mu_0(x)+\mu_0(y)}{2}$, since group membership is not identifiable:[21]

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

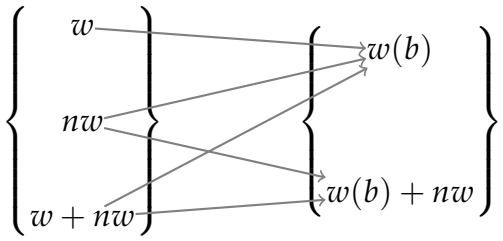
The hiring threshold for group x determined by b^* is similar to the expression obtained for written test screening, s^*, with an important distinction. While the signal given by the blind written test is just as informative as in the non-blind case, now evaluators update a group-neutral prior.

[21] 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.

Hiring Rates With Blind Written and Non-Written Exams Lastly, consider blinding a written exam when the screening process also includes a non-written test. This is similar to the previous case of combining screening signals from both exams, except for the blind written exam having no disparate treatment. However, evaluators still rely on group means and express bias in the overall posterior because of the non-written signal.

5.2 Empirical Predictions

Each of the previous five combinations of screening methods, pre and post 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 interested how each of the transitions above change hiring rates for men and women and the hiring gap, determined by hiring thresholds. In general terms, bias of any kind in a job process reduces selectivity of the favored group (lowers the hiring threshold). Removing or attenuating the expression of bias therefore raises the expected productivity of hired job applicants from that group.

- $w \longrightarrow w(b)$. Blinding a pre-existing screening practice removes disparate treatment by those who conduct the screening. By concealing candidates' identity, evaluator bias ($\pi_j(x)$) cannot be expressed and therefore taste-based discrimination, as well as other bias stemming from decision makers, is eliminated, if they existed. Statistical discrimination is also removed since now candidates' identity and thereby group membership are concealed. However, the bias of a written exam that is independent of the evaluator, $\nu_s(x)$, remains and can still impose disparate impact. If evaluators favor men, either through statistical discrimination or bias, blinding increases hiring rates for women. Alternatively, if an evaluator is biased in favor of women, blinding the exam curbs the eval-

uator's ability to balance women's penalty from statistical discrimination with personal bias, potentially decreasing the female hiring rate.[22]

- $w + nw \longrightarrow w(b) + nw$. Even though this change also blinds a pre-existing written exam, here a candidate's identity is still known during the non-written screening stages, so that group membership information is used in the employer's posterior. As a consequence, disparate treatment is removed only from the written exam, while some 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.

- $nw \longrightarrow w(b)$. This change induces most number of changes to the mix of screening tools. 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, $\nu_s = \nu_\eta$. It follows that the gender hiring gap decreases with the blind-written signal relative to the non-written signal as long as evaluators favor men or, alternatively, if bias toward women is sufficiently small. Relaxing the assumptions of identical disparate impact and screening precision between both tools lead to different predicted changes in hiring rates, depending on the relative values of both quantities. Once again, I discuss these in the appendix and when interpreting my empirical results.

- $nw \longrightarrow w(b) + nw$. This case maintains the use of non-written exams but, to comply with the impartiality requirement, the employer adds a blind-written test to the hiring process. This addition increases screening precision, which has a positive effect on female hiring since evaluators now place less weight on group means. With better screening precision the gender hiring gap narrows even if women on average perform worse on the written test. Additionally, adding a screening tool without introducing evaluator bias (since $d_b = 0$) reduces the weight of the discretion in non-written stages in determining hiring decisions. However, introducing an additional screening method can introduce disparate impact, $\Delta\nu_s$. This can have a negative effect on hiring rates of women, depending

[22]This implicitly assumes that an individual's measured performance (without bias) remains the same regardless of the conditions of the examination. While testing this assumption is difficult, the following exercise helps understand 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 blinding reduces the stereotype threat they face. 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.

on which group it favors and how it compares to the disparate impact that pre-existing non-written methods generate.

- $w + nw \longrightarrow w(b)$. Removing the non-written signal from a screening mix of written and non-written tests involves changes to all determinants of hiring rates. First, it decreases total screening precision. The loss in the number of productivity signals necessarily decreases the female hiring rate. Second, the removal of interviews also eliminates its disparate treatment in the job process, which increases women's hiring rates if evaluators favor men. Similarly, blinding the written exam removes evaluator's bias associated with the tool, which again decreases female selectivity. The third effect on hiring rates is determined by eliminating non-written exam bias. When written and non-written exams have the same disparate impact, removing the interview favors women. If the disparate impact between the two exams are different, then removing interviews increases female hiring if at least one of the exams favors women.

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, evaluator bias, and tool bias, and that depending on the compounded effect of changes in the practices mix, gender hiring gaps may either narrow or increase.

In this section, I take these multiple types of screening tool combinations 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, and the assumptions necessary for identification. I then estimate the effects of counterfactual changes in screening methods on final scores, hiring rates, and female participation in job processes.

6.1 Treatment Space

I begin by grouping 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 omit resume analysis stages. 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 previous section, I studied conceptually the effects on the gender hiring gap of five different changes to screening methods. To understand why these are the empirically relevant cases in the context of the impartiality reform, I trace out in Figure 5 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 difference-in-differences design (perfect compliance). Subgroups of these options would be subject to the monotonicity and exclusion restriction assumptions in the standard instrumental variables case (e.g., Kline and Walters (2016), Feller et al. (2016)). Of the six remaining transition cases, $w \longrightarrow w(b) + nw$ accounts for less than 1% of transitions in the data.[23]

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 section, 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 $w \longrightarrow w(b)$. It also 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, although the disparate impact in $w + nw \longrightarrow w(b) + nw$ combines the tool bias from both exam types.

[23]For this possible transition, on one hand, blinding the pre-existing written exam represents a reduction in partiality. On the other hand, the introduction of the non-written screening tool increases the overall level of discretion in the mix, thus, increasing partiality. The overall change in partiality depends on the respective weights of the written and non-written stages.

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 a setting with no discretion. Second, if both tools provide equally accurate productivity signals, screening precision does not change as a result of the transition, and therefore has no impact on the hiring gap. In contrast, if written tests have higher precision than interviews and female candidates have lower perceived productivity, the increase in screening precision helps women. Third, 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. Adding a blind exam does not introduce a 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 a 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 5 — induced by the policy's increase in hiring impartiality. Coupled with the reduced-form predictions 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 and Identification

Let any treatment D be defined over the support $\mathbb{D} = \mathbb{D}_+ \cup \{0\}$, where job process c receives treatment (dose) D_c, with potential outcomes in period $s = \{t-1, t\}$ given by $Y_{cs}(d)$. Assume further no anticipation, so that $Y_{ct-1} = Y_{ct-1}(0)$ and $Y_{ct} = Y_{ct}(D_c)$. By relaxing the binary treatment assumption ($D = \{0, 1\}$), I allow for any dose or treatment level $d \in \mathbb{D}_+$. Next,

consider the possible treatment types induced by the impartiality reform, given by:

$$d = \left\{ \begin{array}{c} w \longrightarrow w(b) \\ w + nw \longrightarrow w(b) \\ w + nw \longrightarrow w(b) + nw \\ nw \longrightarrow w(b) \\ nw \longrightarrow w(b) + nw \end{array} \right\}.$$

6.2.1 Multi-Valued Difference-in-Differences

For each d, I am interested in estimating the following versions of the baseline difference-in-differences binary treatment model,

$$\begin{aligned} y_{dit} = \delta_{o(d,i)} + \beta_d \left(\text{Fed}_{o(d,i)} \times \text{Post}_{d,o(i),t} \times \text{Female}_i \right) + \gamma_d \left(\text{Post}_{o(d,i),t} \times \text{Fed}_{o(d,i)} \right) \\ + \alpha_d \left(\text{Post}_{o(d,i),t} \times \text{Female}_i \right) + \theta_t + u_{dit}, \end{aligned} \tag{3}$$

which compares outcomes (y_{dit}) for female candidates relative to men (Female_i) participating in job processes for the same occupation ($\delta_{o(d,i)}$) that had screening practices changed (d) only in federal jobs ($\text{Fed}_{o(d,i)}$) and after 1988 $\left(\text{Post}_{o(d,i),t}\right)$. Using the results from Callaway et al. (2021), the standard parallel trends assumption is sufficient to identify the average effect of treatment d among job processes experiencing the treatment, $ATT(d|d) \equiv \mathbb{E}[Y_t(d) - Y_t(0)|D = d] = (\mathbb{E}[\Delta Y_t|D = d] - \mathbb{E}[\Delta Y_t|D = 0])$.

The following example illustrates the variation used to identify $\widehat{\beta}_{w \longrightarrow w(b)}$. The regression in (3) compares a job selection process for, say secretaries, in the federal government that used a written exam before the reform to screen candidates, to another process selecting secretaries to state governments also only using a written test. Under the previously stated difference-in-differences assumptions, the coefficient causally measures the effect of blinding a written exam in the selection of secretaries for federal jobs, using the fact that state-level job processes continued to use a non-blind written exam. Lastly, the interpretation of $\widehat{\beta}_{w \longrightarrow w(b)}$ is informed by the reduced-form predictions from Section 5 — the change in the outcome due to eliminating disparate treatment from w, holding constant screening precision and tool bias.

In my setting, the ability to counterfactually compare certain pairs of treatments is particularly useful. For example, an employer using written and non-written tests to screen employees would increase gender diversity more by blinding the written test and removing the interview ($w + nw \longrightarrow w(b)$) or by only blinding part of the process ($w + nw \longrightarrow w(b) + nw$)? To evaluate this hypothesis, which boils down to estimating the difference between the average treatment effects from each treatment type, $ATT(d|d) - ATT(d'|d') = (ATT(d|d) -$

$ATT(d'|d)) + (ATT(d'|d) - ATT(d'|d'))$, a modified version of the parallel trends assumption is necessary. Following Callaway et al. (2021), for a pair of treatments d, d' that involves the same pre-reform screening tool method, I assume the following set of assumptions, equivalent to a strong parallel trends (STP) assumption:

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0)\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0)|D = 0\right] \quad \text{(STP[1])}$$

$$\mathbb{E}\left[Y_t(d) - Y_{t-1}(0)\right] = \mathbb{E}\left[Y_t(d) - Y_{t-1}(0)|D = d\right] \quad \text{(STP[2])}$$

$$\mathbb{E}\left[Y_t(d') - Y_{t-1}(0)\right] = \mathbb{E}\left[Y_t(d') - Y_{t-1}(0)|D = d'\right] \quad \text{(STP[3])}$$

The set of conditions above replace the standard parallel trends assumption in the context of a multi-valued treatment difference-in-differences. Note that while STP[1] constrains the potential outcome of a group not experiencing treatment, STP[2] and STP[3] require that, on average, no selection occurs when experiencing d instead of d'. This local interpretation of the general result in Callaway et al. (2021) is convenient vis-à-vis the context of the impartiality reform.[24] Dimensionality reduction is a direct implication of the fact that most comparisons of estimated effects in d are not informative or even ill-determined. This, in turn, stems from my definition of a treatment as a *change* in the screening procedure used by an employer, which takes into the account the pre-intervention screening method.

6.2.2 Are the Identifying Assumptions Plausible?

Plausible violations of assumption STP[1] can be tested by conducting the standard pre-trends visual inspection between a treatment type d and its control group, which includes state-level job processes in the same occupation using the same pre-reform screening mix.[25] Figure 6 shows that across all five treatments, gender hiring gaps were not statistically different from zero prior to 1989.

[24] Generally, for all treatments d, the change in outcomes over time across all units if they had been assigned that treatment is the same on average as the change for all units that experienced that dose, $\mathbb{E}[Y_t(d) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d]$. Note that this assumption is still 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. Callaway et al. (2021) show that the assumption above and {STP[1], STP[2], STP[3]} are equivalent.

[25] 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

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

Assumptions STP[2] and STP[3] accommodate a more careful inspection. Jointly, these assumptions impose average null selection in the pairwise comparisons $nw \longrightarrow \{w(b) + nw, w(b)\}$ and $w + nw \longrightarrow \{w(b), w(b) + nw\}$. An empirical case that would violate both assumptions would be if all of the units using an interview pre-reform that wanted to decrease gender gaps switched to a blind test and the remaining units selected randomly between $w(b)$ and $w(b) + nw$. To support that this or other types of violations are unlikely in the context of the reform, I lay out next four main pieces of corroborating evidence.

First, it is important to identify at which level selection into a particular treatment to comply with the reform takes place. The use of a particular set of screening practices before 1989 was highly dependent on the occupation title, determined by customary practices and historical reliance. With the impartiality reform, decisions concerning which hiring practices should be adopted to comply with the Constitutional provision were taken at high organizational levels, still focusing on each occupation within the federal public sector. Bureaucrats, lawyers, and legal aides to different ministries, executive bodies, or regulatory agencies within the federal government put forth occupation-specific guidelines which shaped the changes to screening methods that, regardless of the employer, an occupation would follow. As a consequence, over 95% of occupations only followed one treatment type (i.e., received only one d).

Second, the impartiality provision in Brazil's new Constitution was not part of a broad diversity or anti-discrimination policy. Without bias reduction and improving diversity as clear goals in the impartiality reform as well as accompanying references to gender, racial, or other types of discrimination, the centralized decision-making process that decided which screening methods would be used for a given occupation focused on the legality of new procedures.

Third, even though each occupation overwhelmingly follows only one treatment type, each treatment contains a large number of occupations across a variety of dimensions, as shown in Table A.4. This is suggestive of the lack of interaction between underlying occupation features and potentially strategic selection into a treatment by employers, for example by having male-dominated occupations disproportionately keeping interviews. More formally, regressions in Table A.3 show almost no systematic relationship between selection of a pre-reform screening mix into a particular treatment and occupational skill level, degree of feminization, share of female applicants, and selection competitiveness.

Fourth, the pre-trends plots in Figure 6 are useful to analyze whether occupations that followed different treatments had similar outcomes for men and women. Gender hiring gaps both followed parallel trajectories and had indistinguishably estimated magnitudes between the pairs $nw \longrightarrow w(b)$ and $nw \longrightarrow w(b) + nw$, and $w + nw \longrightarrow w(b)$ and $w + nw \longrightarrow w(b) + nw$. This further suggests that, on average, selection of an occupation into a particular treatment shows no relationship with pre-existing differential gender disparities.

6.3 Estimation and Results

I now proceed to estimate model (3) in five separate regressions for gender final score gaps and gender hiring gaps. Table A.5 shows treatment effects of final scores. For conciseness, I center my discussion in Figure 7, which conducts the same analysis using the gender hiring gap as outcome. Since job offers are solely based on final scores and job openings, any improvement in women's hiring rates relative to men's implies a decrease in the gender final score gap.

Figure 7 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, or $w + nw$ — for 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 p.p., 17 p.p., and a slight gender hiring advantage in the $w + nw$ case of 0.5 p.p. (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 (relative to the baseline of 1.5 p.p. pre-policy) 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%.

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 informed than they are biased. Carefully weighing of these considerations is important to ensure a hiring process that is both more equitable and selects the most productive candidates.

Under the assumptions stated before, the two treatment types provide counterfactuals in a similar sense as Mountjoy (2022) in the IV case. When job processes switch from nw to a written-blind, the gender hiring gap decreases by almost 7 percentage points, starting from a baseline gap of almost 17 percentage points. When benchmarked against the initial gap level, the estimated magnitude implies a decrease in the 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. If evaluators favor men (which is the case in the first estimated effect), then 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 light of growing criticism over requiring standardized or written tests that could disadvantage women or minorities through a disparate impact channel. 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 predictions given by the theoretical framework in Section 5, this treatment improves screening precision without introducing additional disparate treatment by adding $w(b)$, which favors the minority group. With another productivity signal, the final score and therefore hiring threshold relies less on the unconditional group mean, reducing the level of statistical discrimination. In addition, less weight is given to evaluator bias still remaining in nw, further helping women. On the other hand, the introduction of $w(b)$ could have a negative effect on female hiring rates through the screening tool bias component. However, as long as the tool bias of written tests is *relatively* smaller than that of pre-existing non-written stages, it will also contribute towards decreasing the gender hiring gap.

The next two estimates compare alternative treatments of a screening process containing a mix of written and non-written tools, $w + nw$. First, removing the non-written while making the written blind 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 that provide more discretion to evaluators and thus, more prone to bias and seen as potential barriers to increasing diversity.

The lack of a statistically significant effect — despite being precisely estimated — indicates that potential gains from removing both disparate treatment and potential disparate impact from interviews are offset by loss in screening precision from dropping nw. Note that even if women score lower in interviews, eliminating nw has a negative effect on female hiring rates since the resulting noisier productivity signal makes evaluators rely more heavily on their group priors. How much does this precision loss matter? Assuming that evaluators favor men, either *i)* 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 *ii)*

the nw signal precision matters less because interviews favor women (on the disparate impact margin). Evidence in section 6.4 supports the first explanation.

Now consider a treatment that starts with the same screening mix $w + nw$, but only blinds the written stage. This allows to investigate whether partial blinding leads to spillovers in the non-blind stages of the hiring process. I estimate another null effect, although estimates are not precise. Note that the reduced form prediction for $w + nw \longrightarrow w(b) + nw$ is that the female hiring gap increases relative to men due to partial removal of evaluator bias. As the implementation of this type of job process implies using weights for different screening rounds, in Table 10, I further investigate how the estimated effect varies when written tests make up a larger share of the final score.

Columns (2), (4), and (8) show that when enough weight is exogenously placed on the written test (at least 50% of the final score), blinding has a positive and significant effect on women's final evaluation scores and hiring probability, and of similar magnitude as the first treatment $w \rightarrow w(b)$. Additionally, columns (6) and (7) test whether blinding the written test had an impact on how candidates were evaluated in non-blind hiring stages, depending on the blind stage weight. I do not find evidence of strategic or intentional offsetting of the blind scores in the non-blind hiring stages by evaluators.

To conclude this section, Table 9 compares female participation shares in applicant pools for each treatment type. Consistent with the idea from section 4.7 that perceived discrimination or unfair treatment during hiring may discourage minorities from applying in the first place, columns (1), (4), and (5) all show that by blinding a pre-existing written test the participation of women in the applicant pool increases by 2%, 5%, and 4% respectively. Column (2) shows that the switch from a non-written exam to a written stage did not increase women's 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, evidence of which I do not find. In line with a cost of application explanation, column (3) shows that there is also no differential response by gender when employers introduced an additional screening requirement. These results further align with the fact that the reform did not have diversity goals or was promoted as "favoring women", so that applicant responses would depend on the specific change to screening methods.

6.4 Additional Results

A potential issue with the previous estimated effects is whether what I attribute to changes in screening methods actually reflects a factor common to the occupations in each treatment type. For example, even though each treatment has a variety of occupational skill levels, perhaps the decrease in gender gaps from job processes in every treatment is actually picking up a disproportionate effect in low-skill occupations. In Tables A.8 and A.9, I replicate the multi-valued difference-in-differences approach by controlling for skill level and degree of feminization, which leaves estimated effects across the board unchanged.

An additional question is whether the forces determining gender hiring gaps interact with occupation characteristics. Table A.7 splits each treatment type into low and high-skill positions and again estimates treatment effects on the gender hiring gap. Overall, results are similar independent of skill level, except in the treatment $w + nw \longrightarrow w(b)$, where the gender hiring gap decreases in low-skill occupations. This result provides two insights. First, since the disparate impact of interviews is unlikely to vary depending on occupation skill, the precision loss from removing interviews should be the one changing with occupation skill, generally offsetting bias reduction except in low-skill jobs. Second, that interviews provide more information and therefore higher signal precision in high-skill jobs, where productivity could be more easily observed. Therefore, the trade-off between bias and information reduction depends on the precision of the signal provided by the removed screening tool, which can interact with the skill level of the occupation.

7 Impact of Who Hires

In the previous sections, I have focused on the redesign of screening tools to improve gender equity in labor markets. This approach recognizes that certain hiring practices may leave firms more susceptible to biased actions of decision makers. I now turn my analysis to another determinant of potential unequal treatment: decision makers themselves. Directly intervening in the mix of screening tools may be impractical depending on the context or difficult to implement when employers have limited information on relative precision and bias from the screening methods being used.

The changes in screening tools I studied that affect the degree of screening discretion take evaluators' potentially biased behavior as given. Blinding exams removes disparate treatment, which includes both evaluator bias and statistical discrimination. Separating these two forces is important because they can have different implications for efficiency. In a context without evaluator bias, blinding alone could decrease efficiency if statistical discrimination is accurate.

Focusing on decision makers in this section allows me to analyze how evaluator bias varies depending on the cost of bias expression, while leaving statistical discrimination unchanged.

7.1 The Role of Diverse Committees

A common strategy to improve hiring equity is to make the pool of evaluators more diverse. A diverse hiring committee may bring various viewpoints into the search process, thereby providing more nuanced evaluations of applicants with 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)).

Setting I expand on the public sector hiring data around the Impartiality Reform used previously and now consider all job processes from 1999-2019 with complete records on hiring committees. The data combine Brazil's federal and state governments and all written tests are necessarily blind. Committee members evaluate and report scores to candidates individually, but are free to engage in discussions with each other, particularly during non-written stages. To ensure that candidates do not time and form application decisions based on the committee composition of a particular job process, evaluators are only announced after the application deadline, sometimes with the exception of the committee's main evaluator (*presidente da banca*), whose name can be informed in the job announcement.

To ensure a quasi-double-blind mechanism of evaluator-candidate matching, committee members are selected before the job announcement is published and therefore the applicant pool known. Evaluators are generally selected by voting from other peers from a list of potential evaluators, both internally (i.e., same employer) or externally (e.g., different government branch). Employee eligibility to participate in the committee selection process follows a fixed set of alternating rules, as to distribute screening responsibilities somewhat uniformly across employees. That is because screening duties usually carry extra pay and public sector employers cannot give priority to certain individual or group of workers.

This is an important institutional feature since a threat to identification to causally link the composition of hiring committees to gender gaps in hiring could arise if employers who want to hire more women would be more likely to form committees with a higher representation of women. Supporting that this possibility seems unlikely, Table A.12 shows no evidence of any systematic association between the gender composition of committees and any proxies of candidate quality, as captured by blind written and CV scores, and job process characteristics, such as applicant pool size and number of female applicants. I also include employer fixed

effects in the analysis below to address the concern that more female-friendly employers may drive both committee female composition and propensity to hire women.

Table A.10 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, although neither of these differences is 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. As a consequence, women's final scores are 2 percentage points lower than men's.

While these raw differences do not necessarily reflect evaluator bias in non-written exams, Table A.11 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 rates of both groups begin to align, until female candidates become slightly favored when the committee is female-majority.

These patterns indicate that, either due to lower skill or actual performance in non-written exams, or due to some factor related to the higher degree of discretion in these stages, women's scores are lower in interviews, practical exams, and oral presentations. This penalizes their cumulative score, despite doing at least as well on blind exams as men. Moreover, when committees are more female-dominated, men's hiring rates decline, and women's increase. I begin my analysis by first addressing potential skill differences in exam types between genders.

7.2 Gender Gaps and Committee Composition

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

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

where the difference between a candidate's blind written and non-written exam scores is regressed on a female indicator interacted with the share of female committee members. The regression also controls for candidate and employer time-invariant characteristics, and the average gap in each hiring committee, so that β is identified off of candidates who were evaluated by committees with different gender compositions (i.e., individuals who applied to more than one job).

The coefficient of interest measures how the gender gap in subjective exams compared to blind exams varies depending on the committee's gender composition the same candidate faces. Note that in light of my conceptual framework, the within-individual comparison also accounts for differences in disparate impact between different screening tools. The two key identification assumptions to interpret β as loading gender bias are: *i)* no gender differences in how a candidate's abilities in written and non-written tests change across hiring processes and *ii)* no within-employer time-varying omitted factors driving the share of female committee members and propensity of hiring female candidates. Under these assumptions, differences between ΔS_{icj} can be interpreted as levels of evaluator bias a candidate faces in different committees.

Table 11 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 than men, consistent with the raw summary statistics previously discussed. 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 (8) show that the non-written premium for female candidates grows with more women on the committee. However, this effect is non-monotonic: when women represent more than 50% of the committee, increasing the number of female evaluators has the opposite effect on female candidates, decreasing their final score and non-blind premium relative to male candidates. Thus, the non-written premium, the final score received, and the probability that a woman receives an offer all rise relative to men as hiring committees' composition becomes gender-balanced.

7.3 Explaining the Change in Men's Behavior

What could explain a lower gender non-written penalty when screening committees are more feminized? I answer this question next by considering three hypotheses — stereotype threat, gender differences in candidate attribute screening, and gender group norms. Taken together, the evidence is only consistent with changing group norms as the gender mix of hiring committees becomes more diverse.

Stereotype threat. The first mechanism I investigate is the possibility that women's per-

formance changes relative to men's depending on the composition of hiring committees, in line with Steele (1997)'s stereotype threat hypothesis. In section 4.6, I discussed that the underlying context in which a screening method is implemented may generate differential behavioral responses between genders. For example, female candidates may perform better in non-written exams when facing committees with more women.

It is reasonable to assume that if the performance in non-written stages of female candidates improves when they face less stereotype threat, this change in observed performance should not be perceived differently by female or male hiring members. That is, if male evaluators give higher scores to women when committees are more feminized because these candidates perform better, so should female evaluators.

Table 12 conducts this exercise by comparing how women's scores given by female and male committee members change relative to men's as the share of women evaluators varies. The specification also controls for evaluator fixed effects, so that each reported estimate captures how the same evaluator of a given gender scores candidates of both genders when there are different proportions of female colleagues on the hiring committee. Columns (1) through (4) show that female committee members do not change their scoring behavior when there are more female colleagues. Particularly, the estimated effect on the non-written penalty is small and borderline significant at 10%.

Columns (5) through (8) perform the same exercise, now looking at gender differences in scoring by male committee members. The same male evaluator scores female candidates 0.7 percentage points higher when there are more female colleagues on the committee. Scoring on blind tests remains the same regardless of committee composition, which suggests that the average candidate skill in written exams does not change systematically across committees. Taken together, these effects imply a 1.4 percentage-point decrease in the non-written penalty to women.

Female and male evaluators screen for different attributes. Another potential explanation is that women evaluators may screen for a different set of candidate characteristics when giving marks during interviews relative to male colleagues. If these characteristics disproportionately favor female candidates, more women in the hiring committee could change the group discussion dynamics by pointing out to those characteristics possibly neglected by male evaluators. Similarly, women may be better able to screen other women due to shared group identity (e.g., Cornell and Welch (1996)).

Column (9) of Table, however, shows that female evaluators are harsher on female applicants, awarding final scores that are 2 percentage points lower relative to male candidates than male colleagues. This seems inconsistent with the idea that the behavioral response from male

evaluators can be primarily attributed to female colleagues screening for non-written skills that favor women, since they give female candidates an even greater gender penalty than male evaluators. Indeed, this in line with homophilic competition, for example, in a similar sense to Beaman et al. (2012), where members of an ethnic network face a trade-off in the context of job referrals due to competition over employment in an occupational niche.[26]

Group gender norms. While the previous two mechanisms seem inconsistent with the empirical evidence in my setting, the presence of female evaluators may induce a behavioral response from male colleagues through a different margin — norms-based costs (Field et al. (2021), Bertrand et al. (2015)). A shift in the dominant gender norm within the group (Akerlof and Kranton (2000)) or increasing awareness over unconscious bias are consistent with female members of a group influencing how men behave independently of women's own behavior and are met with evidence from other settings.

Undergraduate male students evaluate female colleagues more favorably in female-majority than in female-minority teams (Stoddard et al. (2021)), male board member attendance improves when there are more women on the board (Adams and Ferreira (2009)), male judges are more likely to hand down favorable decisions to plaintiffs alleging gender discrimination when they serve on panels with a female judge (Boyd et al. (2010)), and military male recruits become more egalitarian when assigned to the same squad as women (Dahl et al. (2020)).[27] The idea that expressing bias against female candidates becomes more costly as the committee minority share increases can be accommodated by the model laid out in Section 5 by endogeneizing the cost bias expression — $c_\theta \left[\sum_j f_j / \left(\sum_j f_j + \sum_j (1 - f_j)\right)\right]$ — with $c'_\theta > 0$ and where f_j represents the number of female committee members.

Changes in group gender norms affecting men are also consistent with the non-monotonic relationship between the female share of committee members and women's scores. As committees become more gender-balanced, male evaluators score female candidates more favorably in non-written stages, equalizing hiring outcomes between male and female applicants. Remember that this effect arises in a context where hiring evaluators come up with their own scores for individual job applicants, and despite possible interaction among committee

[26] Empirical findings consistent with women producing less favorable results to other women when compared to men span several settings. For example, Broder (1993) finds that female authors applying for NSF grants have lower chances of success when evaluated by female reviewers than male reviewers. Miller and Sutherland (2022) show that women in Congressional hearings are more likely than men to be interrupted by other women. Bagues and Esteve-Volart (2010), who document that female applicants to the Spanish judiciary have lower chances of being hired when they are randomly assigned to an evaluation committee that includes women.

[27] The effect of female peers on male behavior may also worsen gender disparities depending on the context. In external evaluation boards to decide academic rank promotions in Spain and Italy, Bagues et al. (2017) document that the presence of women evaluators decreases the probability that women receive a positive decision from male colleagues.

members, submit individual scores. As long as hiring committees remain gender-imbalanced toward male members, additional women increase the behavioral response from male evaluators beyond simply adding *one* female evaluator to the group. However, once committees are female-dominated, this reduction in men's bias is eventually offset by lower scores from female evaluators to women applicants.

8 Conclusion

Hiring decisions shape firm outcomes and determine individuals' access to labor market opportunities. To ensure fair recruiting processes and increase employee diversity, organizations face several challenging questions. Does replacing interviews with objective or standardized tests help or hurt female candidates? Should an employer remove screening practices with high discretion even if they may provide important information about applicants? Do different choices of individuals conducting the screening lead to more diverse hires? Causally linking the design of hiring practices to gender equity in hiring requires overcoming lack of data on recruiters' decision making process and generating appropriate variation in the choice and implementation of screening methods.

This paper studies how the design and implementation of firm hiring practices affect gender equity. I open the black-box of hiring decisions by developing a natural language processing algorithm that distills high-dimensional, unstructured text records into a uniquely-detailed dataset on the universe of hiring processes in Brazil's public sector. The data contain complete information on candidate performance, including job offers and individual scores, screening methods, and committee hiring members' evaluations to each job applicant in all hiring stages. This setting provides valuable lessons for private sector firms and professionalized bureaucracies around the world, as public employee selection in Brazil uses screening practices widely adopted in competitive job processes.

Several implications emerge from my analysis. First, gender disparities in hiring come both from screening practices – either by their differences in precision or the existence of bias toward a group – and decision makers. Second, hiring managers matter even in instances when the screening tool provides a relatively objective signal and already limits bias expression. Third, concealing candidate identity in an existing test benefits the less-favored group unambiguously, without leading to an equity-efficiency tradeoff. However, blinding alone may not be enough to improve gender diversity, as evaluator bias allowed by discretion in other hiring stages may be the dominant force.

Further, introducing an additional hiring step comprising blind tests helps female candidates the most. This indicates that improving the accuracy firms' assessment of applicant

productivity offsets evaluator bias in other stages of the job process. In fact, removing subjective tests and blinding exams fails to improve women's outcomes, suggesting that employers should carefully weigh precision loss and net gains from bias reduction. Finally, my results shed light on how to optimally design the composition of hiring committees to minimize gender disparities. Increasing the presence of female evaluators in hiring committees raises scores given by male colleagues to female applicants in subjective stages. More gender-balanced decision makers improve diversity even when the firm does not promote any changes to screening tools.

Remaining questions to be addressed in future research include the implications of different screening practices for the quality of candidates. In separate work, I refine the natural language processing algorithm developed in this paper to extract long-term career progression records of hired job applicants. This allows for empirically testing the theoretical predictions that most changes in screening practices that increase gender equity involve no efficiency losses. More generally, these output and performance measures provide the necessary information to study efficiency concerns in the design of more equitable recruiting practices.

References

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," *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.

Akerlof, George A and Rachel E Kranton (2000) "Economics and identity," *Quarterly Journal of Economics*, Vol. 115, No. 3, pp. 715–753.

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

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," *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," *Review of Economic Studies*, Vol. 77, No. 4, pp. 1301–1328.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva (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," *Quarterly Journal of Economics*, Vol. 131, No. 4, pp. 1593–1636.

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

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

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, 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," *Review of Economic Studies*, Vol. 86, No. 1, pp. 191–239.

Bertrand, Marianne, Emir Kamenica, and Jessica Pan (2015) "Gender identity and relative income within households," *Quarterly Journal of Economics*, Vol. 130, No. 2, pp. 571–614.

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.

Bohnet, Iris (2016) *What works*: Harvard university press.

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.

Bohren, J Aislinn, Peter Hull, and Alex Imas (2022) "Systemic Discrimination: Theory and Measurement," *Working Paper*.

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

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.

Broder, Ivy E. (1993) "Review of NSF Economics Proposals: Gender and Institutional Patterns," *American Economic Review*, Vol. 83, No. 4, pp. 964–970.

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," *NBER Working Paper*.

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

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri (2020) "Are Referees and Editors in Economics Gender Neutral?," *Quarterly Journal of Economics*, Vol. 135, No. 1, pp. 269–327.

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

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.

Cullen, Zoë B and Ricardo Perez-Truglia (2022) "The old boys' club: Schmoozing and the gender gap," *American Economic Review, Accepted*.

Dahl, Gordon B, Andreas Kotsadam, and Dan-Olof Rooth (2020) "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams," *Quarterly Journal of Economics*, Vol. 136, No. 2, pp. 987–1030.

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

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

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

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

Ductor, Lorenzo, Sanjeev Goyal, and Anja Prummer (2021) "Gender and collaboration," *Review of Economics and Statistics*, pp. 1–40.

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," *Annals of Applied Statistics*, Vol. 10, No. 3, pp. 1245 – 1285.

Field, Erica, Rohini Pande, Natalia Rigol, Simone Schaner, and Charity Troyer Moore (2021) "On Her Own Account: How Strengthening Women's Financial Control Impacts Labor Supply and Gender Norms," *American Economic Review*, Vol. 111, No. 7, pp. 2342–75.

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," *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.

Hansen, Fay (2003) "Diversity's business case doesn't add up," *Workforce*, Vol. 82, No. 4, pp. 28–33.

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

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

Hospido, Laura, Luc Laeven, and Ana Lamo (2019) "The gender promotion gap: evidence from central banking," *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.

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

Kline, Patrick and Christopher R. Walters (2016) "Evaluating Public Programs with Close Substitutes: The Case of Head Start," *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.

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

Miller, Michael G. and Joseph L. Sutherland (2022) "The Effect of Gender on Interruptions at Congressional Hearings," *American Political Science Review*, p. 119.

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

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

Mountjoy, Jack (2022) "Community colleges and upward mobility," *American Economic Review*, Vol. 112, No. 8, pp. 2580–2630.

Neumark, David (1996) "Sex discrimination in restaurant hiring: An audit study," *Quarterly Journal of Economics*, Vol. 111, No. 3, pp. 915–941.

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

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

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

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

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

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.

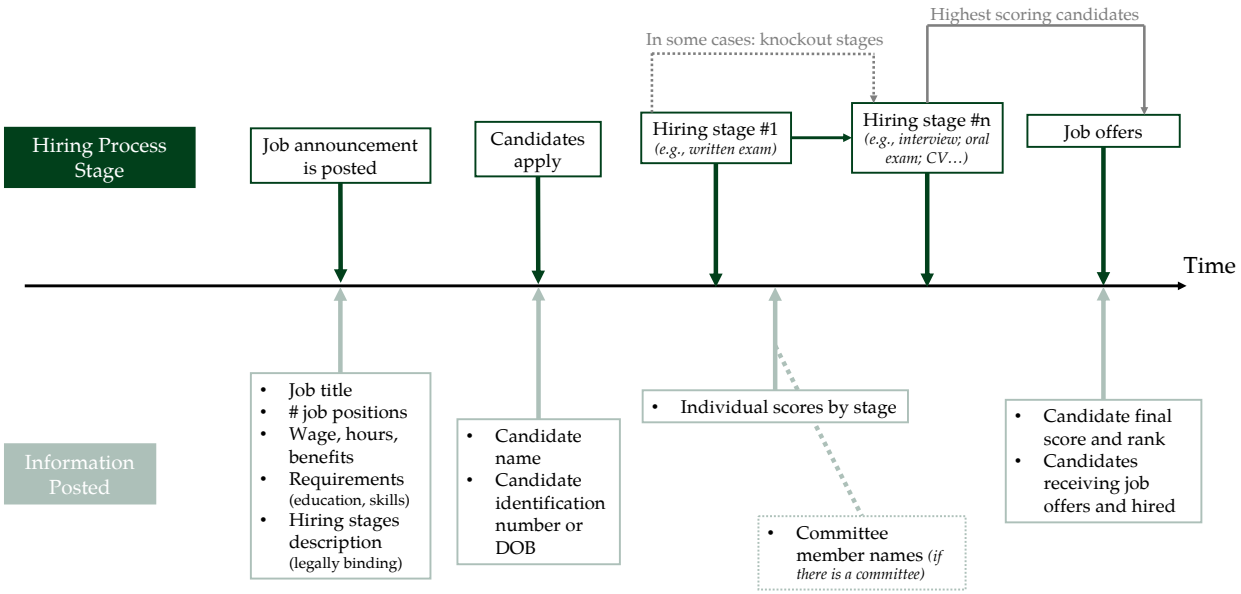
Steele, Claude M (1997) "A threat in the air: How stereotypes shape intellectual identity and performance.," *American Psychologist*, Vol. 52, No. 6, p. 613.

Stoddard, Olga, Christopher F. Karpowitz, and Jessica Preece (2021) "Strength in Numbers: A Field Experiment in Gender, Influence, and Group Dynamics," *Working Paper*.

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.

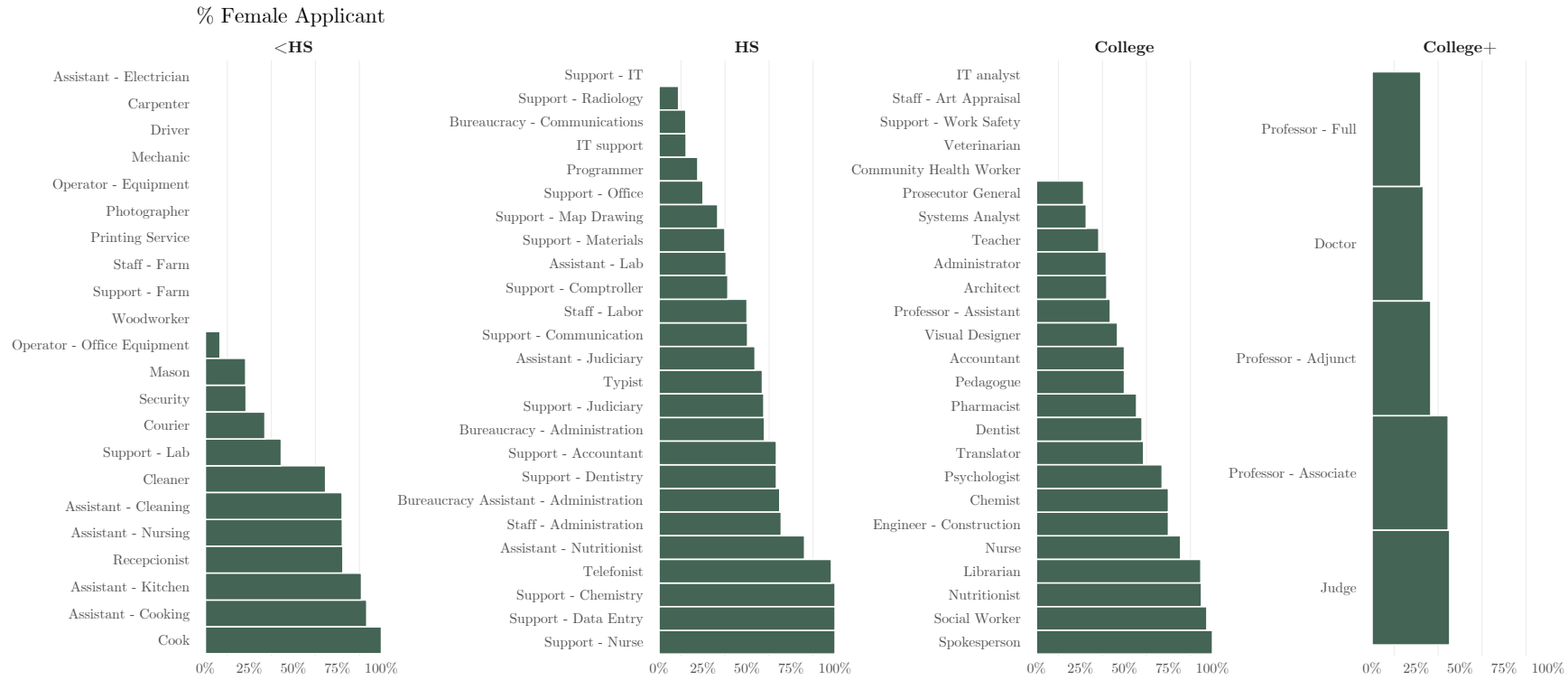
Xu, Guo (2018) "The costs of patronage: Evidence from the British empire," *American Economic Review*, Vol. 108, No. 11, pp. 3170–98.

Figures and Tables



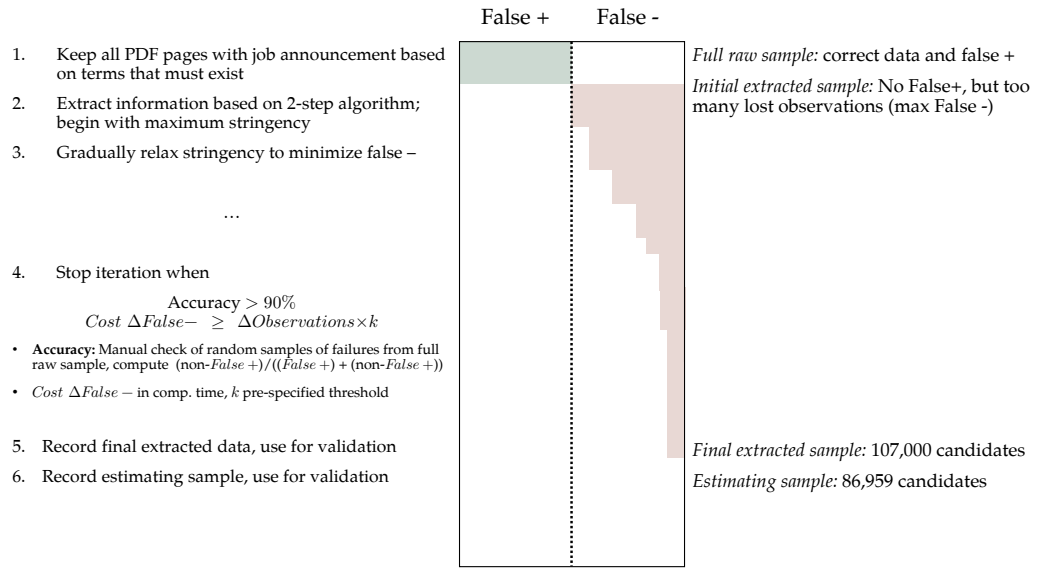
Notes: This shows the stylized structure of a hiring process in the Brazilian public sector posted in raw government publications (official gazettes). Information at the top (dark green) describes the screening dynamics from the moment a job is announced until job offers are sent out. The lower part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



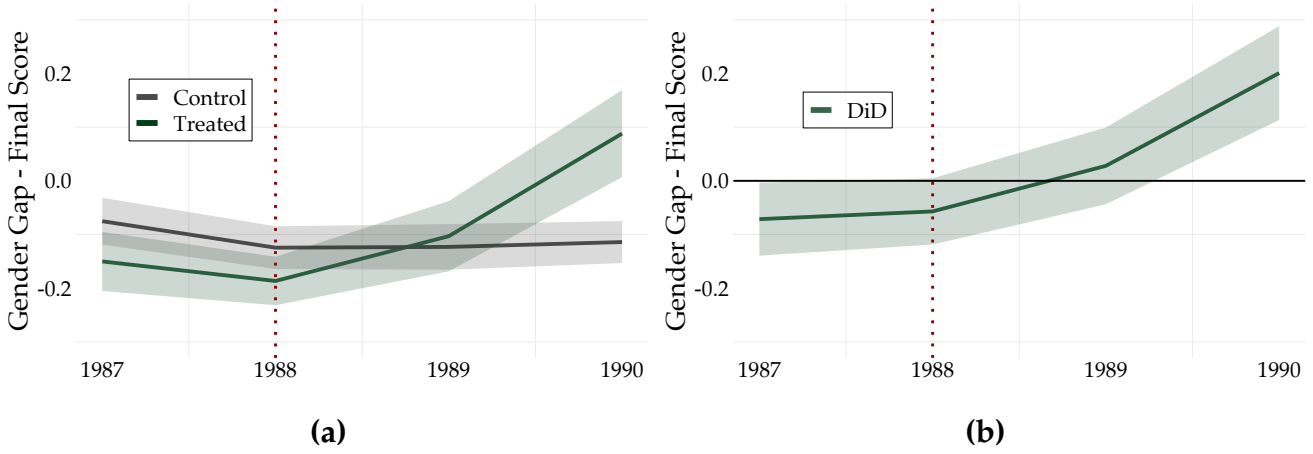
Notes: This figure shows the gender share distribution of job applicants to various occupations and skill levels in Brazil's public sector from 1986 to 1991. Occupation titles in the data follow employer-specific career titles given each organizational structure. These titles are translated from Portuguese 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: This figure shows the implementation pipeline of the second step of the natural language processing algorithm developed in the paper to transform unstructured text into ready-to-use data. Each implementation step (on the left) is associated with a level of generated false positives and negatives (center) and underlying sample size (right).

FIGURE 3: Pipeline Example: Retrieving Candidates Around Impartiality Reform



Notes: The figure on the left plots $\widehat{\gamma_0}$ estimates of the regression

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

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

$$y_{it} = \delta_{o(i)} + \beta \left(\text{Fed}_{o(i)} \times \text{Post}_{o(i),t} \times \text{Female}_i \right) + \gamma \left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)} \right) + \alpha \left(\text{Post}_{o(i),t} \times \text{Female}_i \right) + \theta_t + u_{it}$$

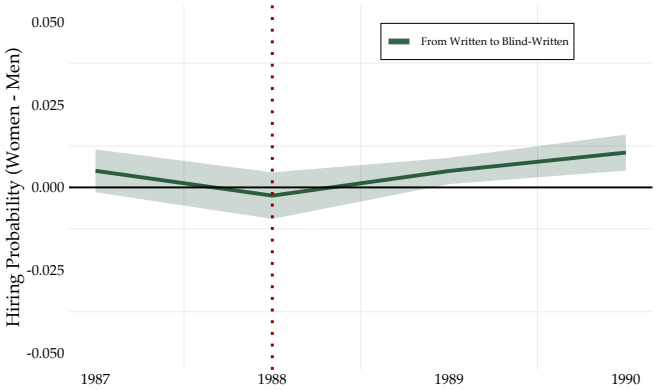
where $\text{Fed}_{o(i)}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level. Shaded areas are 95% confidence intervals. Pooled DiD estimates are shown in Table 4.

FIGURE 4: DiD Dynamic Effects of Impartial Hiring on Final Scores

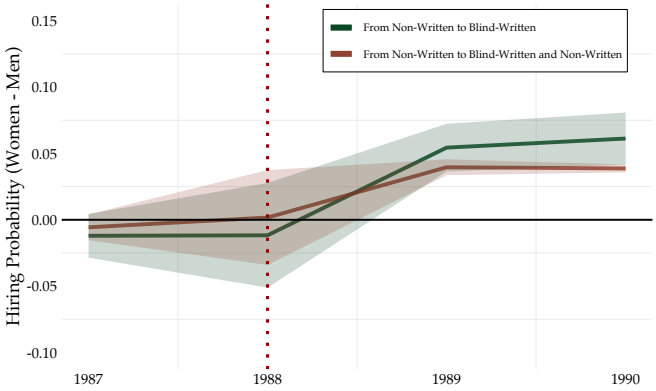
$z=1$ \ $z=0$	**Written**	**Non-written**	**Written & Non-written**	**Written Blind**	**Written Blind & Non-written**
Written	Always Written				
Non-written		Always Non-written			
Written & Non-written			Always Written & Non-written		
Written Blind	$w \to w(b)$ **[20%]**	$nw \to w(b)$ **[6.7%]**	$w+nw \to w(b)$ **[6.7%]**	Always Written Blind	
Written Blind & Non-written	$w \to w(b)+nw$	$nw \to w(b)+nw$ **[20%]**	$w+nw \to w(b)+nw$ **[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 are 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 to a blind-written ($nw \longrightarrow w(b)$), or only adding the blind-written test ($nw \longrightarrow w(b)+nw$), and a hiring process using a mix of written and non-written tools potentially dropping the non-written and blinding the written ($w+nw \longrightarrow w(b)$) or just blinding the written ($w+nw \longrightarrow w(b)+nw$). The potential treatment $w \longrightarrow w(b)+nw$ is accounts for less than 1% of transitions in the data. Written exams are shorthand for written or multiple-choice tests, and non-written indicate interviews, practical exams, or oral exams. Numbers in $[\cdot]$ give the frequency of each treatment type in the estimating sample.

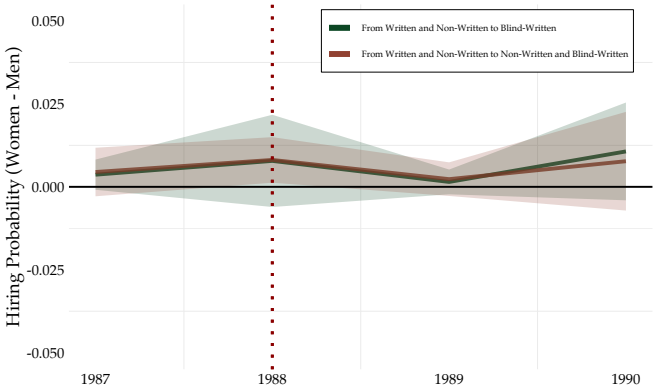
FIGURE 5: Potential Treatments Induced by Reform



Pre-Reform Mix: w



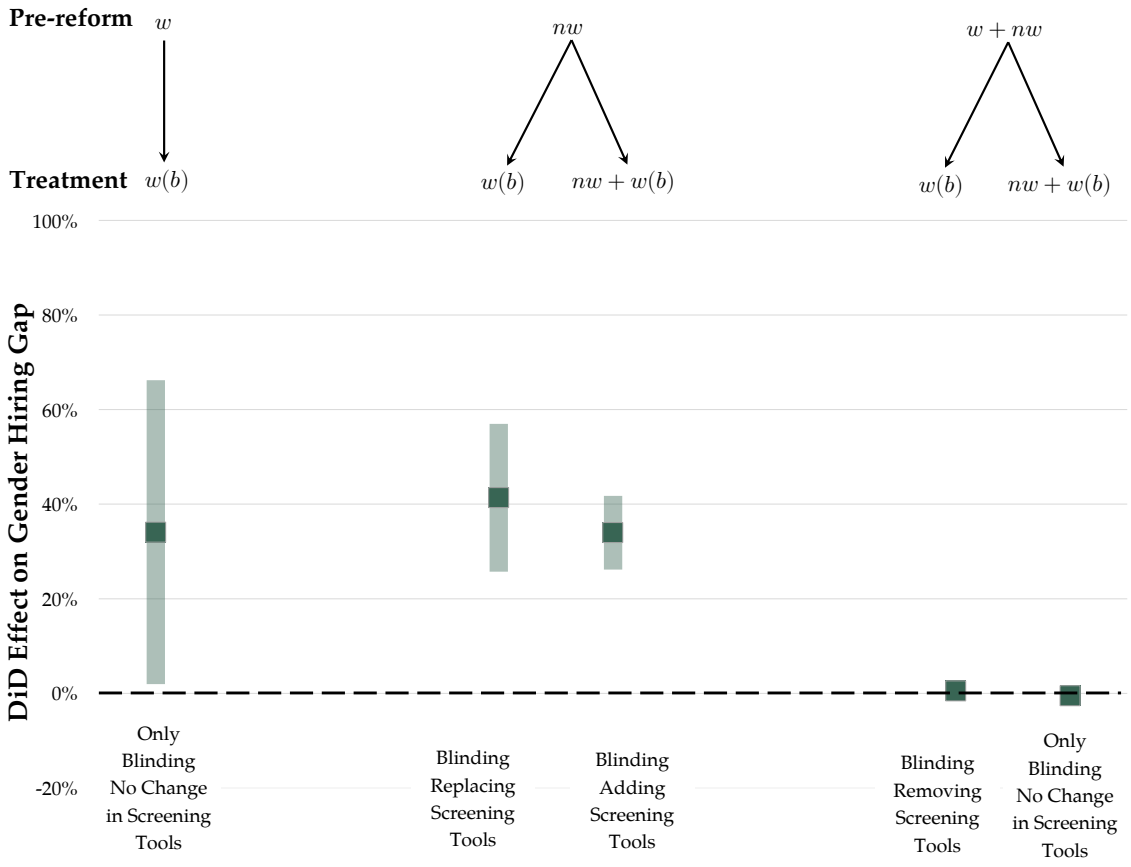
Pre-Reform Mix: nw



Pre-Reform Mix: $w + nw$

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

FIGURE 6: Outcome Paths in Each Treatment Type



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

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

where $\text{Fed}_{o(g,i)}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(g,i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Treatment type g represents 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 ($w + nw \longrightarrow w(b) + nw$). Standard errors are clustered at the job process level. Bars are 95% confidence intervals. Regression outputs are shown in Table A.6.

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

TABLE 1: Selected Characteristics of Job Processes

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

Notes: These are descriptive statistics of job processes used in the Impartiality Reform analysis. Treated job processes are those in Brazil's federal sector. The control group comprises processes in the states of Amazonas in the country's north region, Pernambuco in the northeast, Distrito Federal, Mato Grosso, and Mato Grosso do Sul in the central region, São Paulo — the largest and richest state — in the southeast, and Rio Grande do Sul in the south. The period is 1986 through 1991. Sample statistics for *Screening Steps* have occupation fixed effects to compare temporal changes within different hiring processes for the same job title.

TABLE 2: 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)
$\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}$	−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 $y_{ct} = \delta_{o(c)} + \gamma \left(\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}\right) + \theta_t + u_{ct}$, where outcomes in columns (1) through (6) at the job process level c are regressed on an interaction for post-1988 federal jobs, $\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}$, 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 3: Estimates of Screening Impartiality on Hiring and Application Rates

	$\mathbf{Pr}(Hired \mid \text{Female} = 1)$	$\mathbf{Pr}(Hired \mid \text{Male} = 1)$	$\mathbf{Pr}(Hired \mid \text{Applied} = 1)$	$\mathbf{Pr}(\text{Female} = 1)$
	(1)	(2)	(3)	(4)
$\text{Fed}_{o(i)} \times \text{Post}_{o(i),t}$	0.003**	−0.004***		0.010**
	(0.001)	(0.001)		(0.005)
$\text{Fed}_{o(i)} \times \text{Post}_{o(i),t}$			0.007***	
$\times \text{Female}_i$			(0.002)	
Pre-reform mean	0.09	0.11	−0.015	0.52
Occupation FE	X	X	X	X
Year FE	X	X	X	X
Obs.	54,892	32,067	86,959	86,959

Notes: The first column shows a regression coefficient capturing the probability that a given female job applicant receives a job offer: $\mathbf{Pr}(Hired = 1)_{it} = \delta_{o(i)} + \gamma \left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)}\right) + u_{it}$ which is ran only on female individuals, column (2) runs the same regression but in the male candidate subsample, column (3) runs $\mathbf{Pr}(Hired = 1)_{it} = \delta_{o(i)} + \beta \left(\text{Fed}_{o(i)} \times \text{Post}_{o(i),t} \times \text{Female}_i\right) + \gamma \left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)}\right) + \alpha \left(\text{Post}_{o(i),t} \times \text{Female}_i\right) + \theta_t + u_{it}$ which measures the probability a female job applicant receives an offer relative to male job applicants. Finally, column (4) regresses the specification $\mathbf{Pr}(\text{Female} = 1)_{it} = \delta_{o(i)} + \gamma \left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)}\right) + u_{it}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level.

TABLE 4: Estimates of Screening Impartiality on Candidate Scores

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

Notes: The table shows DiD estimates of the form $y_{it} = \delta_{o(i)} + \gamma \left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)} \right) + u_{it}$ only with female candidates in columns (1), (4), and (7), only with male candidates in columns (2), (5), and (8), and $y_{it} = \delta_{o(i)} + \beta \left(\text{Fed}_{o(i)} \times \text{Post}_{o(i),t} \times \text{Female}_i \right) + \gamma \left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)} \right) + \alpha \left(\text{Post}_{o(i),t} \times \text{Female}_i \right) + \theta_t + u_{it}$ 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). $\text{Fed}_{o(i)}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level.

TABLE 5: Estimates of Screening Impartiality on Job Process Outcomes

	% Women of	% Female	Log # Candidates		
	Hired	Candidates	All	Women	Men
	(1)	(2)	(3)	(4)	(5)
$\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}$	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 regressions $y_{ct} = \delta_{o(c)} + \gamma \left(\text{Fed}_{o(c)} \times \text{Post}_{o(c),t} \right) + \theta_t + u_{ct}$, where $\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}$ is an interaction 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 6: Estimates of Effects on Scores and Hiring Probability, Skill Level

	Final Score			$\mathbf{Pr}(Hired \mid \text{Applied})$		
	<High School (1)	High School (2)	College or Advanced Degree (3)	<High School (4)	High School (5)	College or Advanced Degree (6)
$\text{Fed}_{o(i)} \times \text{Post}_{o(i),t}$	−0.016	0.160	0.204***	0.002	−0.001	0.011*
$\times \text{Female}_i$	(0.084)	(0.111)	(0.080)	(0.003)	(0.002)	(0.005)
Occupation FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Obs.	35,475	22,071	29,413	35,475	22,071	29,413

Notes: This table reports, from columns (1) through (3), DiD estimates of the form $y_{it} = \delta_{o(i)} + \beta\left(\text{Fed}_{o(i)} \times \text{Post}_{o(i),t} \times \text{Female}_i\right) + \gamma\left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)}\right) + \alpha\left(\text{Post}_{o(i),t} \times \text{Female}_i\right) + \theta_t + u_{it}$ where outcome y represents a candidate's final score. $\text{Fed}_{o(i)}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Column (1) shows regression coefficients for a subsample of occupations with less than high-school education required. Column (2) runs the regression for a subsample of occupations with that require high-school education, and column (3) for a subsample of high-skill occupations that require a college degree or more. Columns (4) through (6) report regression coefficients for the respective subsamples capturing the probability that a given female job applicant receives a job offer relative to male applicants. Standard errors are clustered at the job process level.

TABLE 7: Estimates on Scores and Hiring Probability, Feminization Degree

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

Notes: This table displays regression coefficients of the model $y_{it} = \delta_{o(i)} + \beta\left(\text{Fed}_{o(i)} \times \text{Post}_{o(i),t} \times \text{Female}_i\right) + \gamma\left(\text{Post}_{o(i),t} \times \text{Fed}_{o(i)}\right) + \alpha\left(\text{Post}_{o(i),t} \times \text{Female}_i\right) + \theta_t + u_{it}$. The outcome represents either a female candidate's final score relative to a male candidate in columns (1)—(3), or the probability that a female candidate receives a job offer relative to a male candidate in columns (4)—(6). $\text{Fed}_{o(i)}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Columns (1) and (3) report estimates for a subsample of female-dominated occupations, defined as the proportion of women in that occupation $> 60\%$. Columns (2) and (5) run the regression for a subsample of occupations that are neutral or gender balanced, if the proportion of women in that occupation is between 40% and 60%. Columns (3) and (6) run the regression for a subsample of male-dominated occupations, defined as the share of women $< 40\%$. Standard errors are clustered at the job process level.

TABLE 8: Estimates of Effects on Female Share of Hires, Feminization Degree

Occupation Gender Identity	% Women of Hired		
	Female-dominated	Neutral	Male-dominated
	(1)	(2)	(3)
$\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}$	0.104	0.131**	0.259***
	(0.068)	(0.064)	(0.081)
Occupation FE	X	X	X
Year FE	X	X	X
Obs.	48,681	16,848	21,430

Notes: The table shows selection process regressions $y_{ct} = \delta_{o(c)} + \gamma \left(\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}\right) + \theta_t + u_{ct}$, where $\text{Fed}_{o(c)} \times \text{Post}_{o(c),t}$ is an interaction for whether the job process is for a federal-level position post-Impartiality reform ($t \geq 1989$). Each column runs the regression for subsamples of female-dominated (share of women $> 60\%$), neutral (share of women $\in (40\%, 60\%)$), or male-dominated occupations (share of women $< 40\%$). Standard errors are clustered at the job process level.

TABLE 9: 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)
$\text{Fed}_{o(g,c)} \times \text{Post}_{g,o(c),t}$	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

$$\%\ \text{Female Applicants}_{gct} = \delta_{o(g,c)} + \beta_g \left(\text{Fed}_{o(g,c)} \times \text{Post}_{g,o(c),t}\right) + \theta_t + u_{gct}$$

where $\text{Fed}_{o(g,c)} \times \text{Post}_{o(g,c),t}$ is an interaction 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 10: Decomposing Treatment Effects by Weight of Blind Stages

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

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

$$y_{git} = \delta_{o(g,i)} + \beta_g \left(\text{Fed}_{o(g,i)} \times \text{Post}_{g,o(i),t} \times \text{Female}_i \right) + \gamma_g \left(\text{Post}_{o(g,i),t} \times \text{Fed}_{o(g,i)} \right) + \alpha_g \left(\text{Post}_{o(g,i),t} \times \text{Female}_i \right) + \theta_t + u_{it}$$

where $\text{Fed}_{o(g,i)}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(g,i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Treatment type g represents job process transitioning from using a mix of written and non-written tools to blinding the written ($w + nw \longrightarrow w(b) + nw$). The outcome y represent either a candidate's final score, non-blind (non-written) score, or probability of being offered a job. Columns (2), (4), (6) and (8) condition on the weight on the blind written test in a job process to be $> 50\%$, and columns (3), (5), (7) and (9) for the weight on the blind written test to be less than 50%. Standard errors are clustered at the job process level.

TABLE 11: Effect of Committee Gender Composition on Gender Equity

	$\text{Score}^{nw} - \text{Score}^{w(b)}$	Final Score			$\mathbf{Pr}(Hired = 1)$
	Overall	Overall	$< 50\%$ Female	$> 50\%$ Female	Overall
	(1)	(2)	(3)	(4)	(5)
$\text{Female}_i \times$	0.407***	0.163***	0.587***	−0.396***	0.414***
$\%\text{Female Evaluator}_c$	(0.102)	(0.052)	(0.113)	(0.126)	(0.138)
Job Applicant FE	X	X	X	X	X
Job Process FE	X	X	X	X	X
Employer FE	X	X	X	X	X
Obs.	9,901	9,901	6,219	3,682	9,901

Notes: Column (1) shows estimates of the model model $\text{Score}^{nw}_{icj} - \text{Score}^{w(b)}_{icj} = \beta\left(\text{Female}_i \times \%\text{Female Evaluator}_c\right) + \gamma_c + \mu_i + \gamma_e + \varepsilon_{icj}$, where β captures evaluator bias in non-written exams. Columns (2) through (5) 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.

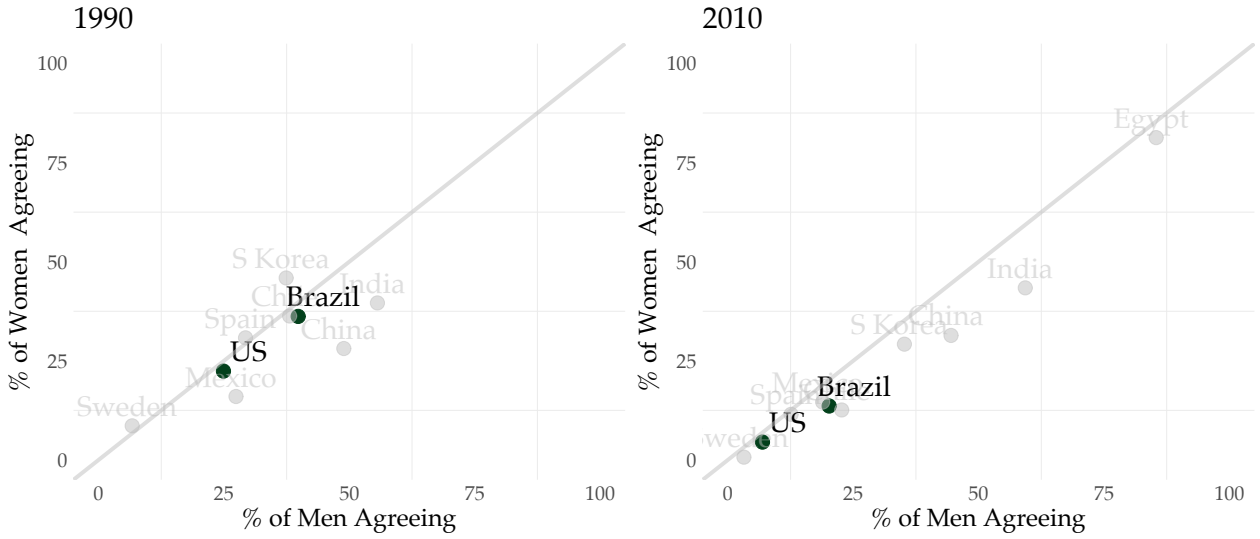
TABLE 12: Do Male Committee Members React to More Female Colleagues?

	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	Final Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\text{Female}_i \times$	0.026	−0.003	0.008*	−0.012	0.073*	−0.035	0.139***	−0.010	
$\%\text{Female Evaluator}_c$	(0.042)	(0.032)	(0.042)	(0.029)	(0.039)	(0.031)	(0.039)	(0.024)	
Female_i									0.007
									(0.006)
Female_j									0.010*
									(0.006)
$\text{Female}_i \times \text{Female}_j$									−0.018**
									(0.009)
Committee Member FE	X	X	X	X	X	X	X	X	X
Employer 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	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.

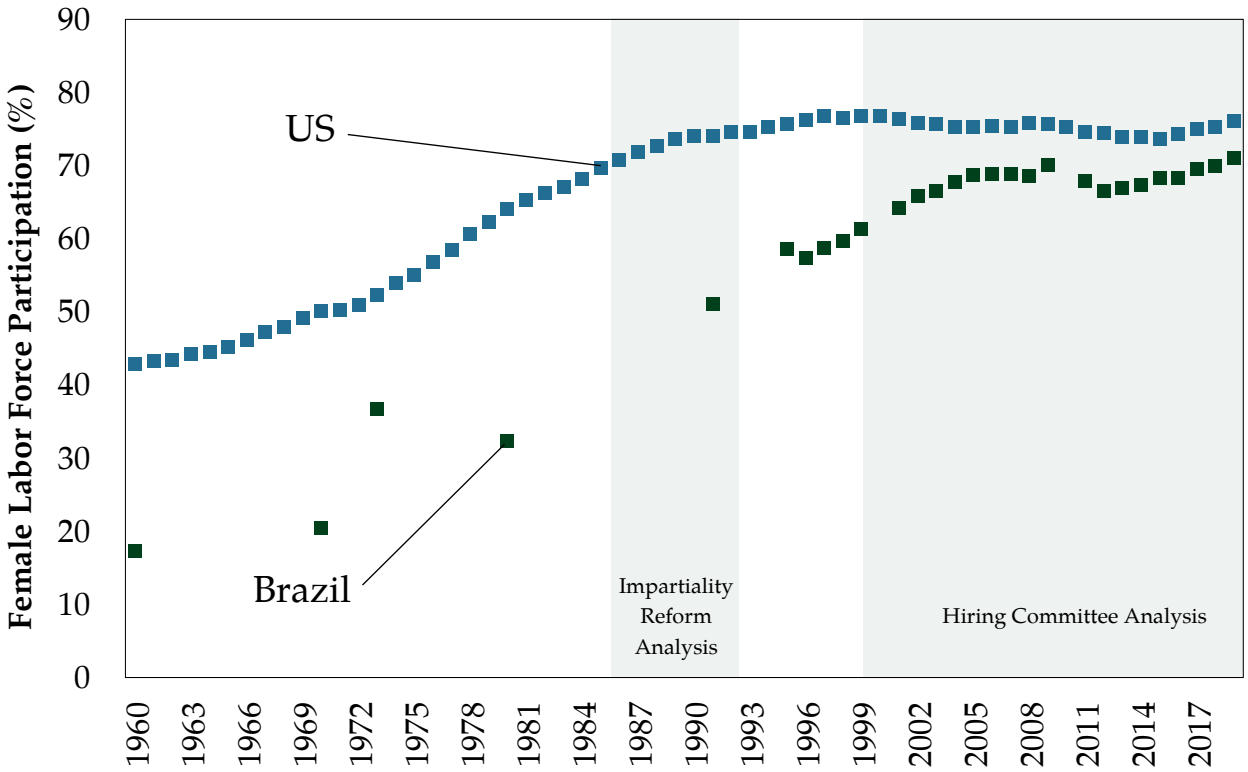
ONLINE APPENDIX

A Appendix Tables and Figures



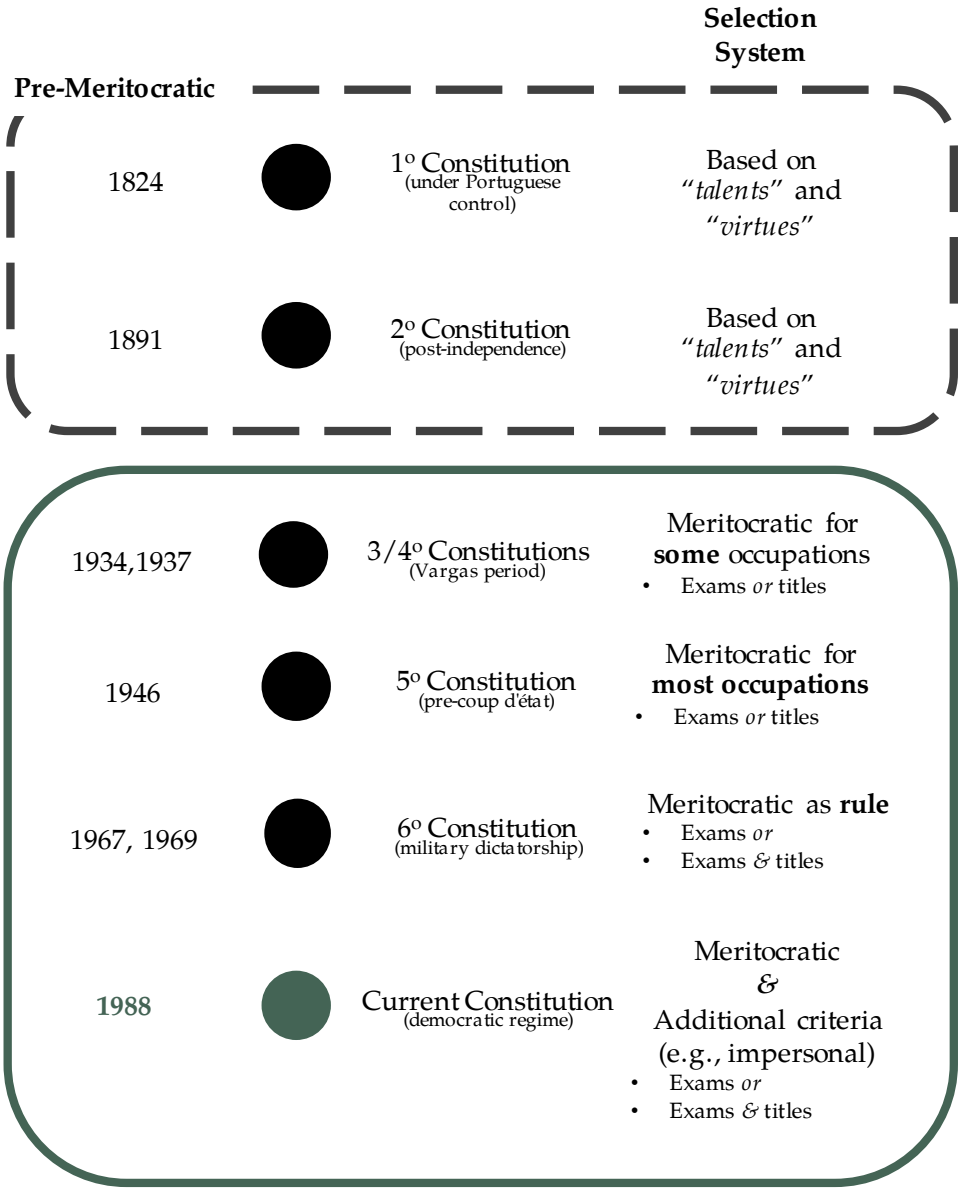
Notes: International Value Survey (IVS) answers for the 1990-1994 and 2010-2014 waves of women and men agreeing with the statement "*when jobs are scarce, men have more of a right to a job than women*". Countries plotted: South Korea, China, India, Brazil, US, Mexico, Spain, Sweden, and Egypt (for 2010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

FIGURE A.1: Gender Attitudes Across Countries



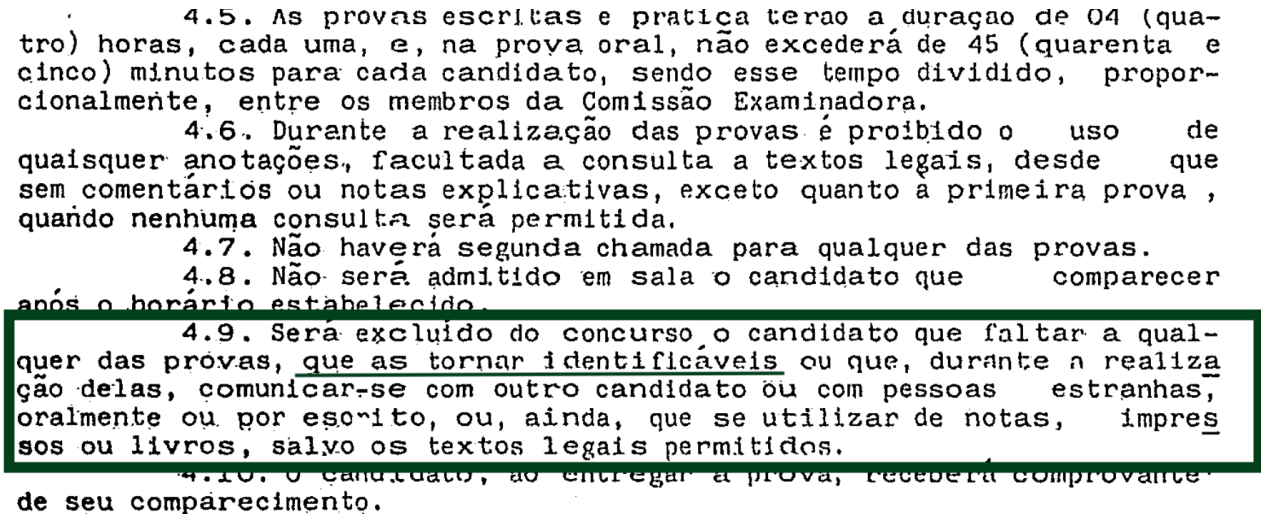
Notes: Female labor force participation rates (aged 25-54) for Brazil and the US. Shaded areas represent periods for different empirical analyses in the paper.

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



Notes: This 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.3: History of Changes in the Selection of Public Servants in Brazil



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

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

4.7. Não haverá segunda chamada para qualquer das provas.

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

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

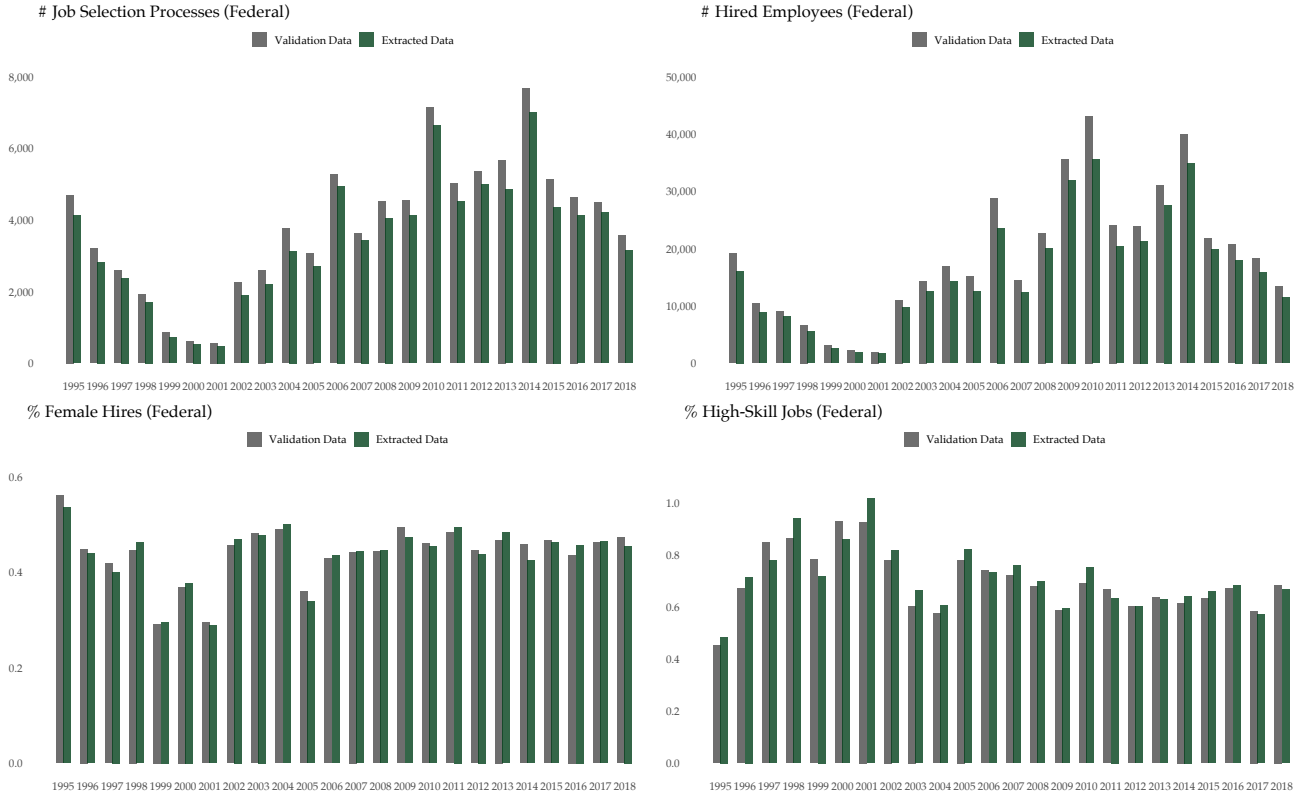
4.10. O candidato, ao entregar a prova, receberá comprovante de seu comparecimento.

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

FIGURE A.4: Enforcing Blind Exams After Reform

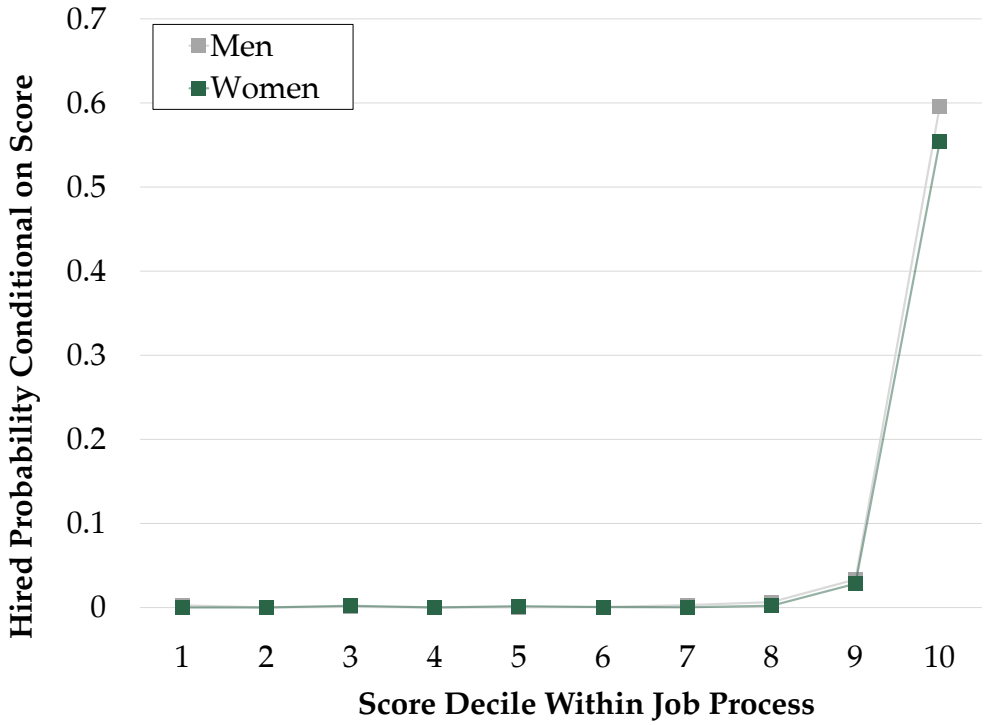
Notes: This is a sample of some layouts of job selection process postings in Brazilian government gazettes (similar to the Federal Register). Information on candidate scores, hiring committee members, exam types, selection process rules, results and other variables described in Figure 1 are contained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

FIGURE A.5: Examples of Raw Text Containing Hiring Information



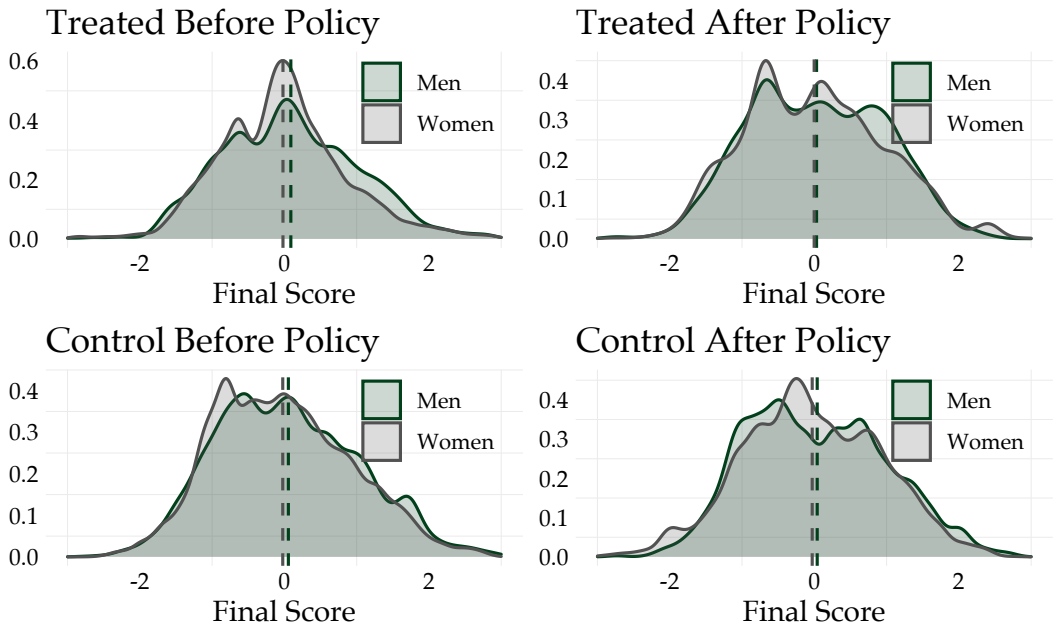
Notes: This figure compares aggregate statistics informed by Brazil's federal government on its public sector to calculated sample moments using data extracted from official government gazettes. For each statistic — annual number of job selection processes, number of hired employees, share of female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

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



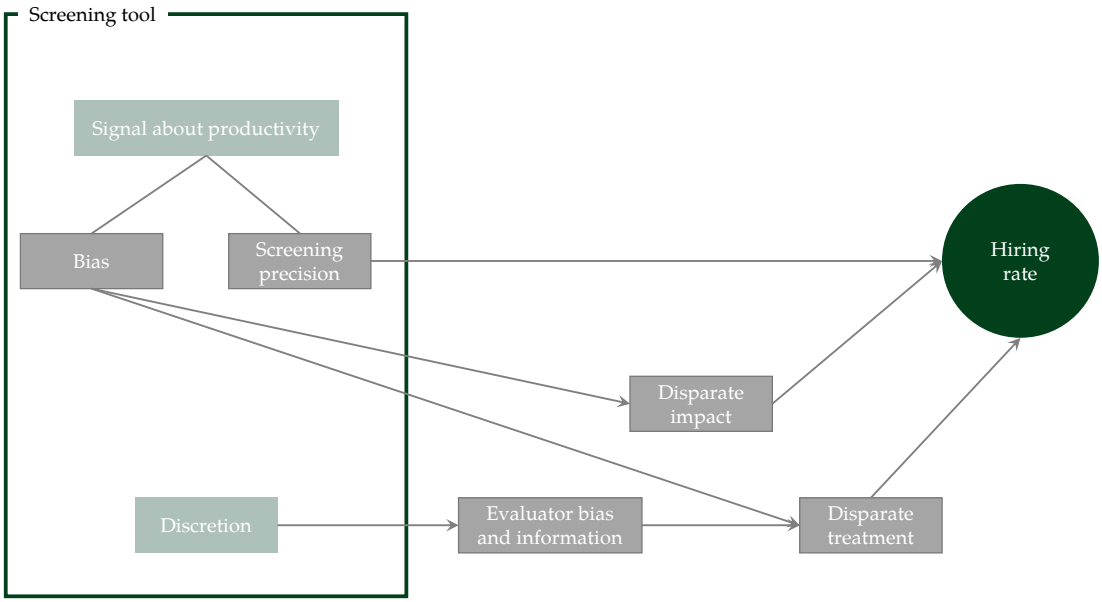
Notes: This figures illustrates how final scores completely determine candidates' probability of bein hired in public sector job processes. In accordance with the law requiring that the highest scoring candidates across all evaluation stages are offered jobs until all openings are fulfilled, candidates 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 A.7: Candidate Final Scores Determine Hiring Chances

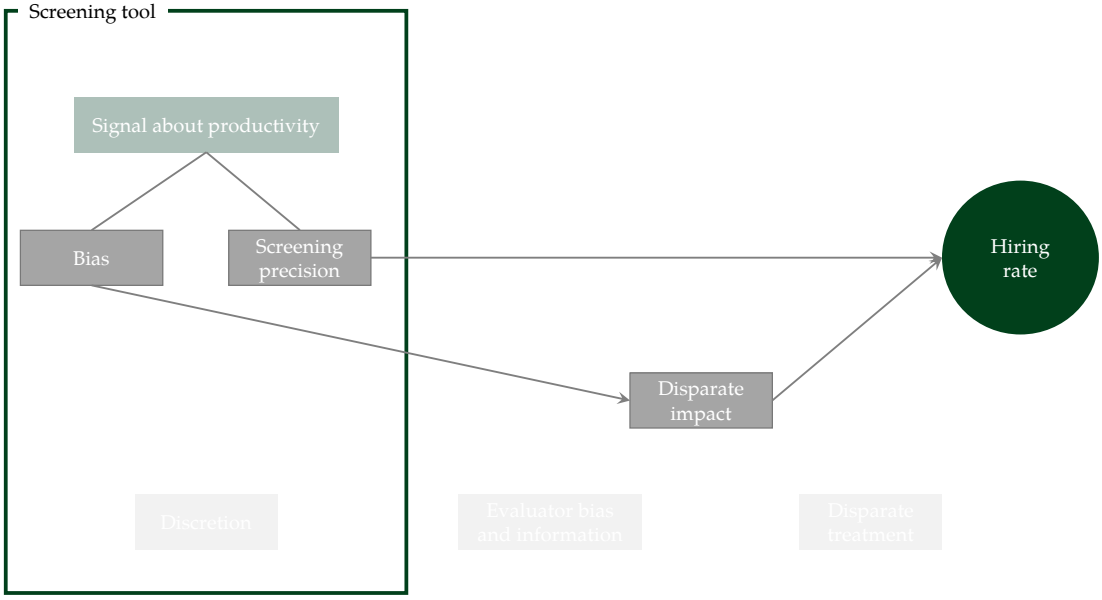


Notes: These panels show standardized final scores of male and female job applicants in federal and state job processes. Federal (treated) and states (control) before and after the impartiality reform are displayed in each panel. To compare magnitudes across densities, tails are censored between 2 standard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

FIGURE A.8: Final Scores Distributions



(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, such as 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 a 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 A.9: Hiring Rate Determinants: Conceptual Framework

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**

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: Estimates of Screening Impartiality on Candidate Score, Candidate FE

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

Notes: The table shows DiD estimates of the form

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

where $\text{Fed}_{o(i)}$ is an indicator for whether the job process is for a federal-level position, and $\text{Post}_{o(i),t}$ is a dummy for post-Impartiality reform ($t \geq 1989$). Standard errors are clustered at the job process level.

TABLE A.3: Correlation Between Occupation Characteristics and Selection into Treatment

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

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

TABLE A.4: Examples of Occupations in Each Treatment Type

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

Notes: These are examples of occupation titles in each treatment type induced by the impartiality reform in Brazil's new Constitution.

TABLE A.5: Treatment Effects of Changes in Screening Tools: Gender Score Gap

	Final Score				
	$w \longrightarrow w(b)$	$nw \longrightarrow w(b)$	$nw \longrightarrow nw + w(b)$	$w + nw \longrightarrow w(b)$	$w + nw \longrightarrow nw + w(b)$
	(1)	(2)	(3)	(4)	(5)
$\text{Fed}_{o(i)} \times \text{Post}_{o(i),t}$	0.040***	0.195***	0.141***	0.026*	0.002
$\times \text{Female}_i$	(0.016)	(0.0438)	(0.016)	(0.014)	(0.025)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Obs.	12,066	9,343	18,570	44,964	32,898

Notes: This table shows treatment effects for each treatment type d induced by the 1988 Impartiality Reform in Brazil's public sector. Each column estimates a version of the difference-in-differences regression

$$\text{Final Score}_{dit} = \delta_{o(d,i)} + \beta_d \left(\text{Fed}_{o(d,i)} \times \text{Post}_{d,o(i),t} \times \text{Female}_i \right) + \gamma_d \left(\text{Post}_{o(d,i),t} \times \text{Fed}_{o(d,i)} \right) + \alpha_d \left(\text{Post}_{o(d,i),t} \times \text{Female}_i \right) + \theta_t + u_{it}$$

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

TABLE A.6: Treatment Effects of Changes in Screening Tools: Gender Hiring Gap

	$\mathbf{Pr}(Hired \mid \text{Applied} = 1)$				
	$w \longrightarrow w(b)$	$nw \longrightarrow w(b)$	$nw \longrightarrow nw + w(b)$	$w + nw \longrightarrow w(b)$	$w + nw \longrightarrow nw + w(b)$
	(1)	(2)	(3)	(4)	(5)
$\text{Fed}_{o(i)} \times \text{Post}_{o(i),t}$	0.005**	0.070***	0.058***	0.005	−0.005
$\times \text{Female}_i$	(0.002)	(0.013)	(0.007)	(0.003)	(0.007)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Obs.	12,066	9,343	18,570	44,964	32,898

Notes: This table shows treatment effects for each treatment type d induced by the 1988 Impartiality Reform in Brazil's public sector. Each column estimates a version of the difference-in-differences regression

$$\mathbf{Pr}(Hired = 1)_{dit} = \delta_{o(d,i)} + \beta_d \left(\text{Fed}_{o(d,i)} \times \text{Post}_{d,o(i),t} \times \text{Female}_i \right) + \gamma_d \left(\text{Post}_{o(d,i),t} \times \text{Fed}_{o(d,i)} \right) + \alpha_d \left(\text{Post}_{o(d,i),t} \times \text{Female}_i \right) + \theta_t + u_{it}$$

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

TABLE A.7: Treatment Effects of Changes in Screening Tools by Occupation Skill

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

Notes: This table shows treatment effects for each treatment type d induced by the 1988 Impartiality Reform in Brazil's public sector, for subsamples of job processes in a given skill level. Each column estimates a version of the difference-in-differences regression

$$\mathbf{Pr}(Hired = 1)_{dit} = \delta_{o(d,i)} + \beta_d \left(\text{Fed}_{o(d,i)} \times \text{Post}_{d,o(i),t} \times \text{Female}_i \right) + \gamma_d \left(\text{Post}_{o(d,i),t} \times \text{Fed}_{o(d,i)} \right) + \alpha_d \left(\text{Post}_{o(d,i),t} \times \text{Female}_i \right) + \theta_t + u_{it}$$

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

TABLE A.8: Treatment Effects of Changes in Screening Tools: Skill Robustness

	$\mathbf{Pr}(Hired \mid \text{Applied} = 1)$				
	$w \longrightarrow w(b)$	$nw \longrightarrow w(b)$	$nw \longrightarrow nw + w(b)$	$w + nw \longrightarrow w(b)$	$w + nw \longrightarrow nw + w(b)$
	(1)	(2)	(3)	(4)	(5)
$\text{Fed}_{o(i)} \times \text{Post}_{o(i),t}$	0.005**	0.070***	0.058***	0.003	−0.004
$\times \text{Female}_i$	(0.002)	(0.013)	(0.007)	(0.004)	(0.008)
Occupation FE	X	X	X	X	X
Year FE	X	X	X	X	X
Occupation Skill FE	X	X	X	X	X
Obs.	12,066	9,343	18,570	44,964	32,898

Notes: This table shows treatment effects for each treatment type d induced by the 1988 Impartiality Reform in Brazil's public sector. Each column estimates a version of the difference-in-differences regression

$$\mathbf{Pr}(Hired = 1)_{dit} = \delta_{o(d,i)} + \omega_{skill(o,i)} + \beta_d \left(\text{Fed}_{o(d,i)} \times \text{Post}_{d,o(i),t} \times \text{Female}_i\right) + \gamma_d \left(\text{Post}_{o(d,i),t} \times \text{Fed}_{o(d,i)}\right) + \alpha_d \left(\text{Post}_{o(d,i),t} \times \text{Female}_i\right) + \theta_t + u_{it}$$

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

TABLE A.9: Treatment Effects of Changes in Screening Tools: Feminization Robustness

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

Notes: This table shows treatment effects for each treatment type d induced by the 1988 Impartiality Reform in Brazil's public sector. Each column estimates a version of the difference-in-differences regression

$$\mathbf{Pr}(Hired = 1)_{dit} = \delta_{o(d,i)} + \omega_{fem(o,i)} + \beta_d \left(\text{Fed}_{o(d,i)} \times \text{Post}_{d,o(i),t} \times \text{Female}_i\right) + \gamma_d \left(\text{Post}_{o(d,i),t} \times \text{Fed}_{o(d,i)}\right) + \alpha_d \left(\text{Post}_{o(d,i),t} \times \text{Female}_i\right) + \theta_t + u_{it}$$

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

TABLE A.10: 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 A.11: Raw Hiring Probabilities, 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 A.12: Predicting the Assignment of Female Committee Members

	Blind Score	CV Score	Candidate Pool Size	# Female Applicants
	(1)	(2)	(3)	(4)
% Committee Female	−0.049	0.018	−0.693	0.081
	(0.032)	(0.014)	(0.480)	(0.247)
R-squared	0.04	0.01	0.01	0.002

Notes: (1) and (2) regress job applicant quality proxies on the share women in the hiring committee, while (3) and (4) regress job-process pre-determined characteristics. Standard errors are clustered at the job-process level.

B Model Derivations

This appendix provides detailed derivations to the conceptual framework laid out in Section 5.

B.1 Hiring Rates for w

Start with the case of selecting candidates based on a written test, which is allowed to be biased. The distribution of written signals is given by:

$$s^* = y + \nu_s(x) + \varepsilon_s,\ \varepsilon_s \sim N(0, 1/h_s)$$
$$s^* \sim N(\mu_0(x) + \nu_s(x), 1/h_s)$$

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 the mean bias of the test, capturing the disparate impact of the screening tool, which favors men when $\nu_s(m) > \nu_s(f)$ and is women-favoring if $\nu_s(m) < \nu_s(f)$.

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

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

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

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

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

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

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

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

which, can finally be expressed as:

$$\frac{s-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_s}} > \underbrace{\frac{k_s-\nu_s(x)-\mu_0(x)-d_s\pi_j(x)}{\sigma_0\rho_s}}_{z_s^*(x)} \tag{4}$$

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 expected hiring rate of applicants from group x is $1-\Phi(z_s^*(x))$.

B.2 Hiring Rates for nw

The hiring rate for group x when candidates are screened using non-written tests is obtained following the same steps as in the previous case, observing the different distribution of non-written signals:

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

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, d_η. Since non-written screen tools, such as interviews or oral exams, are more subjective than written tests, it follows that the discretion given to managers is higher with non-written than written tests: $d_\eta > d_s$.

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

$$\frac{\eta-\mu_0(x)}{\frac{1}{h_0}+\frac{1}{h_\eta}} > \underbrace{\frac{k_\eta-\nu_\eta(x)-\mu_0(x)-d_\eta\pi_j(x)}{\sigma_0\rho_\eta}}_{z_\eta^*(x)} \tag{5}$$

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

B.3 Hiring Rates for $w+nw$

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

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

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

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

Thus, the hiring decision is given by:

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

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

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

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

Further manipulation gives the final hiring threshold:

$$\frac{\mu(x, \eta, s) - \mu_0(x)}{\sigma_0\rho_T} > \underbrace{\frac{k_T - \frac{h_s}{h_T}\nu_s(x) - \frac{h_0 + h_\eta}{h_T}\nu_\eta(x) - \pi_j(x)(d_\eta + d_s) - \mu_0(x)}{\sigma_0\rho_T}}_{z_T^*(x)}. \tag{6}$$

B.4 Hiring Rate for $w(b)$

A blind written exam provides the following signal:

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

with the same screening precision h_s and the same disparate impact $\nu_s(x)$ as the written test. Because discretion is entirely removed after blinding, evaluators rely on the *population* produc-

tivity for updating, $\mu_0 = \frac{\mu_0(x)+\mu_0(y)}{2}$, since group membership is not identifiable,

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

Manipulating the above gives the hiring threshold for group x:

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

and finally,

$$\frac{s - \mu_0(x)}{\frac{1}{h_0} + \frac{1}{h_s}} > \underbrace{\frac{k_b - \nu_s(x) - \mu_0(x)}{\sigma_0\rho_s} - \frac{h_0\rho_s}{2h_s\sigma_0}(\mu_0(y) - \mu_0(x))}_{z_b^*(x)}. \tag{7}$$

B.5 Hiring Rate for $w(b) + nw$

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:

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

and the updated mean

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

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

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

The hiring decision can then be rewritten as:

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

$$\frac{\mu(x, \eta, b) - \mu_0(x)}{\sigma_0\rho_T} > \underbrace{\frac{k_{\eta b} - \nu_s(x)\frac{h_s}{h_T} - \nu_\eta(x)\frac{h_\eta + h_0}{h_T} - d_\eta\pi_j(x) - \mu_0(x)}{\sigma_0\rho_T}}_{z_{\eta b}^*(x)}. \tag{8}$$

B.6 Change in Hiring Rate for $w \longrightarrow w(b)$

Without loss of generality, assume that female candidates are less productive on average than men: $\mu_0(f) < \mu_0(m)$, or in other words, that women are the minority group. By blinding the written exam, how do screening thresholds $z_s^*(x)$ and $z_b^*(x)$ compare and thus how are hiring rates affected? By inspecting expressions (4) and (7) and considering that written tests — whether blind or not — have the same screening precision and disparate impact, since they are otherwise identical save for hiding a candidate's identity, women face a lower hiring threshold in the blinded exam, $z_s^*(f) > z_b^*(m)$, if and only if

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

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

B.7 Change in Hiring Rate for $nw \longrightarrow w(b)$

I now analyze the potential change induced by the policy that most dramatically alters the mix of screening tools. To build intuition, consider an employer that solely relies on interviews to screen candidates. From the expression in (5), the disparate impact of interviews, their precision, and how much they enable evaluator bias to be expressed all determine an applicant's hiring odds. Only in terms of evaluator bias, under the assumption that interviews offer more discretion than written exams, this pre-policy state contains the highest expression of evaluator bias. In contrast, as discussed before, screening solely based on written exams is likely to provide a setting with low disparate treatment.

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

favor men $\pi_j(m) > 0$ or, alternatively, if the following is satisfied

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

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

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

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

Note that the left-hand side of the expression above is negative, so that if evaluators are men-favoring and interviews have a larger disparate impact than written exams, the inequality is satisfied and female hiring rates increase. In other words, if the principal substitutes a hiring tool for one that has a smaller disparate impact and eliminates discretion, the change will raise hiring rates of the minority group. More generally, if either evaluator bias favors men, or if the relative bias of non-written tests is lower than that of written tests, it can still increase female hiring as long as it satisfies the inequality above. Another way to interpret the inequality (9) is to rewrite it as

$$\frac{h_0 \rho}{h_s \sigma_0}\mu_0(f) + \frac{d_\eta \pi_j(f)}{\sigma_0 \rho} + \frac{(\nu_\eta(f) - \nu_s(f))}{\sigma_0 \rho} < \frac{h_0 \rho}{h_s \sigma_0}\mu_0(m) + \frac{d_\eta \pi_j(m)}{\sigma_0 \rho} + \frac{(\nu_\eta(m) - \nu_s(m))}{\sigma_0 \rho} \tag{10}$$

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

Finally, relax the assumption of identical screening precisions. If written tests are more precise, $h_s > h_\eta$, switching from non-written screening to written testing raises the hiring rate of the group with lower perceived productivity, that is, it raises the female hiring rate if (10) holds. However, if interviews have higher precision, $h_s < h_\eta$, the transition from interviews to written test decreases screening precision and leads to higher hiring rates of the favored group,

men. The net effect then depends on the losses from decreased screening precision relative to the gains from lower bias if (10) is satisfied.

B.8 Change in Hiring Rate for $nw \longrightarrow w(b) + nw$

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

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

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

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

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

$$\frac{\partial \left[1 - \Phi(z^*_{\eta b}(x))\right]}{\partial \rho_T} = \phi(z^*_{\eta b}(x)) \left[\frac{z^*_{\eta b}(x)}{\rho_T} - \frac{\partial k_{\eta b}/\partial \rho_T}{\sigma_0 \rho_T}\right] > 0 \tag{11}$$

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

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

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

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

B.9 Change in Hiring Rate for $w + nw \longrightarrow w(b)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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