Words Matter: Gender, Jobs and Applicant Behavior[*]

Sugat Chaturvedi[†]
Indian Statistical Institute

Kanika Mahajan[‡]
Ashoka University

Zahra Siddique[§]
University of Bristol

May 1, 2021

Abstract

We examine employer preferences for hiring men vs women using 0.16 million job ads posted on an online job portal in India, together with all applications made to these ads by 1.06 million active job seekers. We apply machine learning algorithms on text contained in the title and description of job ads to construct measures that indicate how predictive the job ad text is of an employer's explicit gender preference. We find that explicit gender preferences are more likely to be exhibited for low-skill jobs, that advertised wages are lowest for jobs with an explicit female preference, and, even in jobs without an explicit gender preference, advertised wages are substantially lower when the job text is predictive of an explicit female preference. An explicit female preference by an employer also reduces the total number of applications to a job ad whilst changing the gender mix of the applicant pool in favor of women. We systematically uncover words that lie beneath these associations by categorising words which are predictive of explicit gender preferences into those related to hard and soft skills, personality and flexibility. We find that decreased flexibility (indicated by increased travel requirements and unusual working hours) is associated with a *higher* advertised wage in job ads, with such ads attracting a smaller share of female applicants. At the same time job ads containing words indicating skills which are also predictive of an explicit female preference have a *lower* advertised wage and attract a larger share of female applicants. Our results highlight the important role played by explicit gender preferences and implicit gendered word associations within job ads in explaining gender disparities in the labor market.

Keywords: Gender, Job portal, Text analysis, India
JEL classification: J16, J63, J71

[*] We are grateful to Ashoka University for providing funding for this project through a departmental research grant. Rakesh Kumar and Sanchit Goel provided excellent research assistance. We thank Farzana Afridi, Sonia Bhalotra, Rosella Calvi, Rochana Chaturvedi, Sekyu Choi, Seema Jayachandran, Aprajit Mahajan, Nishith Prakash, Shekhar Tomar and Basit Zafar for useful discussions and feedback. We would also like to thank seminar participants at Samvaad, Association for Mentoring and Inclusion in Economics (AMIE), IWWAGE-ISI, India Data Foundation and the University of Verona for useful comments. All errors are our own.

[†] Address: Economics and Planning Unit, Indian Statistical Institute, 7, S.J.S. Sansanwal Marg, New Delhi, Delhi 110016, India. Email: sugatc15r@isid.ac.in

[‡] Address: Ashoka University, Rajiv Gandhi Education City, Sonepat, Rai, Haryana, India, 131029. Email: kanika.mahajan@ashoka.edu.in

[§] Address: University of Bristol, Senate House, Tyndall Avenue, Bristol BS8 1TH, United Kingdom. Email: zahra.siddique@bristol.ac.uk.

1 Introduction

Gender disparities in labor force participation and wages exist in both developed and developing countries, albeit to a varying degree. These disparities can arise due to differences in search behaviour and final matching with jobs across gender. Akerlof and Kranton (2000) show how association with a group identity matters for economic outcomes.[1] What remains under explored are the mechanisms through which such differences may be generated. In a recent study, Bordalo et al. (2019) use a cooperative game setup and find that gender associations or stereotypes contribute to gender gaps in self-confidence and consequently in behavior. In this paper, we look at how implicit gender associations can influence job search behaviour and thus could be important in shaping labor market outcomes.

Goldin (2014) shows that within-occupation wage differentials account for a larger proportion of the gender wage gap than between-occupation wage differentials for the U.S. Hence, a focus on attributes of jobs within occupations then becomes important to understand labor market disparities. We show how data from job vacancies, text of the job ad and applications can be used to test the role played by job attributes within occupations in affecting the search behaviour of candidates. We use propriety data from an online job search portal in India for our analyses. This data allow us to surmount the difficulty arising from lack of large scale and high quality data on labor markets in developing country contexts to investigate the causes and mechanisms behind gender disparities in job search. In fact our setting enables us to exploit presence of explicit gender preferences exhibited by firms in the job ads they post online, to obtain implicit gender associations contained in these ads, and examine their consequent effect on gender differences in job search. This helps us understand the role of employer demand in influencing labor market outcomes. Notably, such perceived associations can be held by employers and candidates. Therefore, we further see how candidates respond to these employer gender associations. Lastly, we directly look at how text contained in job ads matters for gender mix of applicants.

A distinctive feature of the Indian labor market today is the presence of large gender gaps in labor force participation (Fletcher et al., 2018). Despite high economic growth, increases in

[1] Several studies evaluate gender differences along various dimensions such as risk, overconfidence, competitive behavior, undertaking negotiations, and ambiguity aversion and show how these relate to economic returns in the labor market. See Shurchkov and Eckel (2018) for a review of the literature.

educational attainment and a decline in fertility over time, female labor force participation rates have remained stagnant among urban households (Klasen and Pieters, 2015; Afridi et al., 2018). In 2017–2018 only 20.6% of working age Indian women (age 15–65) in urban households were part of the labor force, compared to 78.9% of working age Indian men.[2] In fact, India's female labor force participation rates rank among the lowest in the world today and it also has one of the largest gender wage gaps in the world.[3] Though there are provisions within the Indian legal framework that could prohibit employers from posting job ads that explicitly request a male or female, the implementation of labor laws is generally inadequate.[4] Therefore, it is not unusual for employers to express explicit gender preferences in the job ads they post online.

We examine such preferences by using data on 0.16 million job ads posted on an Indian job portal between July 2018 and February 2020 together with all applications made in response to these ads. We look at compliance by candidate applications to the gender request in job ads. Next, we use explicit gender preferences to derive an implicit employer gender preference (*femaleness* or *maleness*) signalled by the job text and examine their effect on wages and applications. We examine these effects both across occupations and within occupations in a location. We then unveil the words that contribute to the gendered beliefs held by employers i.e., towards *femaleness* or *maleness* scores derived using gender requests. We categorize these words into five categories —hard-skills, soft-skills, personality and flexibility —and examine which categories affect wages and the gender mix of applicants. Lastly, we examine the possibility that candidates can respond to words because they believe that it indicates an implicit gender preference by the employer or because they attach themselves to these attributes. To check the overlap in employer and candidate beliefs across categories, we look at the direct effect of words on the gender mix of applicants.

Around 7.7% of the job ads exhibit an explicit gender preference in the data, with slightly more

[2] Based on own calculations from the Periodic Labor Force Survey (PLFS) 2017–2018. Labor force status is defined using activity status over the previous year. A person is in the labor force if they are self-employed, an unpaid family worker, a regular salaried employee, a casual worker or unemployed.

[3] According to the World Bank, India's current female labor force participation rates are only better than those in Yemen, Iraq, Jordan, Syria, Algeria, Iran, and West Bank and Gaza and are comparable to Morocco, Afghanistan, Somalia, Pakistan, Egypt, Saudi Arabia, Lebanon and Tunisia. The ILO Global Wage Report 2018-19 documents the large gender wage gap in India.

[4] Article 16 of the Constitution of India prohibits discrimination on the basis of sex in public employment, while Article 39 guides the state to direct its policy towards ensuring "equal pay for equal work for both men and women". The Equal Remuneration Act, 1976 implements the provisions of Article 39 and prohibits sex based discrimination in payment of salary for same work (or work of similar nature) as well as in recruitment, promotion, training and transfer.

ads exhibiting an explicit female preference than an explicit male preference. We find that jobs with an explicit gender preference tend to be low-skill jobs (in terms of lower education requirements and advertised wages). Jobs with an explicit female preference have the lowest advertised wage, on average. In terms of applicant responses to explicit requests, we find that an employer's explicit female preferences dramatically reduce the number of applications to a job ad. At the same time, they increase the *share* of female applications by 15.4 percentage points (or 48%) while an explicit male preference reduces this share by 9.5 percentage points (or 30%). Hence, explicit gender preferences have a substantial impact on the gender mix of the applicant pool. We use the explicit gender requests to arrive at implicit gender association using machine learning methods and find that even among jobs without any explicit gender preference, job ads containing text predictive of an employer's female preference (*femaleness*) have a significantly lower advertised wage. We also find that a higher fraction of women apply to these low-wage jobs.

We find that gender segregation can occur not only due to the employer's explicit gender preferences, but also because the applicant pool for a job may be proportionately larger for a given gender if the text contained in a job ad displays an implicit gender association. In line with findings in Kuhn et al. (2020) for China, we also find that women's share in the applicant pool increases when an explicit request for women is made than when it is not made for the same implicit gender association, showing that women may be more "ambiguity-averse". However, unlike China, we find that the gap in the share of female applicants who apply to female targeted ads and gender non-targeted ads initially increases and then remains constant as the implicit *femaleness* associated with a job ad increases. These results show that implicit gender associations contained in the job text matter, and that they matter for Indian women even when the ad includes an explicit preference for women.

We then open the black box of these implicit employer preference measures by using methods in explainable artificial intelligence. We uncover words that contribute towards the *femaleness* and the *maleness* scores. We classify the words into five broad categories described earlier. Broadly, employer preference indicated through hard skills and job flexibility matter consistently for both the advertised wages and the share of female applicants. For jobs that do not express any explicit gender request, we find that an increase in net score for words in the category of hard-skills associated with women is associated with a lower advertised wages and a higher share of female applicants.

On the other hand an increase in net score for words in the category of flexibility associated with men (require travel or odd hours of work) is associated with a higher advertised wages and a lower share of female applicants. For jobs that request a female we find that compliance with the gender requests decreases as words associated with skills and flexibility associated with the opposite gender increase.

Our work is inspired by a series of papers which investigate explicit gender preferences in the Chinese labor market (Kuhn and Shen, 2013; Helleseter et al., 2020). Kuhn and Shen (2013) were the first to document explicit gender preferences in job ads. They found evidence of a persistent negative skill-targeting relationship, i.e. as a job's skill requirements increased (as measured by required education, required experience or advertised wages) the share of ads expressing explicit gender preferences declined. Using additional data (including from a job board in Mexico) Helleseter et al. (2020) found that firms' explicit gender requests shifted from female to male workers as they sought older (vs younger) workers, a feature also referred to as the 'age twist'. We find a similar pattern in the urban Indian labor market regarding the negative skill-targeting relationship but we find only limited evidence of the age twist. Ningrum et al. (2020) also examine employer's explicit gender preferences using data from a job portal in Indonesia. In contrast to these papers, we derive implicit gender associations from explicit gender requests and are able to observe applicants' behavior in our data; this allows us to address additional important research questions such as how explicit and implicit employer demand for a gender can influence search behavior. We contribute to the above literature by providing a first comprehensive analysis of gender targeting in job ads for India. In related work, Chowdhury et al. (2018) examine explicit employer's gender preferences from an Indian job portal called *Babajob* but do not use detailed occupation controls, data on applications or derive implicit associations.

We build on work by Kuhn et al. (2020) who use applications data from an online job portal in China. We additionally employ several techniques from the literature on machine learning. First, we use a short text topic model to classify job titles to detailed occupation categories. Second, we construct implicit associations (*femaleness* and *maleness*) attached to a job from the text that appears in the job title and *job description* using a different machine learning algorithm, which we argue improves upon their method. This method also allows us to look at implicit associations within occupational categories. Third, we examine the relationship between implicit gender associations with advertised wages and not just with the gender mix of applicants to a job.

Lastly, our work is the first to derive job attributes associated with each gender using recent research in explainable artificial intelligence, hitherto not applied in the field of economics. We use these methods to arrive at a word list that may be used to identify gender associations or stereotypes. This word list is likely to be useful to researchers who are interested in uncovering gender associations in a similar labor market setting but who do not have information on explicit gender preferences. We undertake this from both employer and applicant side. Such word lists can be a useful resource for detecting biases in textual data. For instance, Burn et al. (2019) use word vectors to calculate cosine similarity of words from the existing literature in the field of industrial psychology with phrases in jobs ads to detect bias against older workers in the US.

Further, we use the word list indicative of employer gender preference to examine how these affect the advertised wages and applicant behavior. Thus, this paper also extends the recent literature on job attributes (like work hours, part time work, ability to schedule work according one's own preference), wage penalty and the gender wage gap (Goldin and Katz, 2011; Goldin, 2014; Mas and Pallais, 2017; Bustelo et al., 2020; He et al., 2019). Flory et al. (2015) look at application rates, by varying the job content from a stereotypical male (sportsperson) to stereotypical female (administration) job. Related to this, studies in psychology show that the wording of job ads might affect students' inclination to apply for a job, however none of these studies consider actual applications.[5] Moreover, in the existing literature, the characteristics that attract women and men are drawn from small and non-representative surveys. For instance, Taris and Bok (1998) compile 20 characteristics based on 512 job ads judged by 40 students as being typically male or female while Gaucher et al. (2011) use lists of words denoted as feminine and masculine (based on gender differences in linguistic style) from existing studies.[6]

More generally, we add to a growing literature on various aspects of labor markets using high frequency data from online job portals. Hershbein and Kahn (2018) use data from job vacancies to investigate how skills demand changed over the Great Recession in the US. A number of recent

[5] For instance, Born and Taris (2010) find that females respond more to feminine characteristics than men respond to masculine characteristics among 78 applicants. The study used the characteristics "solid business sense" and "decisiveness" (both masculine), and "communication skills" and "creativity" (both feminine) to describe desired candidate profile. In a sample of 96 participants, Gaucher et al. (2011) find women were more likely to find job ads appealing where a greater proportion feminine words were used and candidates were also more likely to anticipate gender diversity in such job ads.

[6] Abele and Wojciszke (2014) further divide these associations with words into largely communal and agentic types.

papers examine high frequency changes in employment and skill demand with the onset of the Covid-19 pandemic by making use of job portal data.[7] While we use a similar data source, the research questions we study are quite distinctive from these papers.

Our work is also relevant for a growing empirical literature motivated by directed search models which investigates where job seekers send their applications (Moen, 1997). Marinescu and Wolthoff (2020) use data from the US job portal *Careerbuilder* to find that job titles and posted wages affect the applicant pool that a firm attracts. Banfi and Villena-Roldan (2019) use data from a Chilean job portal to find that job ads with higher wages attract more applicants while Banfi et al. (2019) use the same dataset to document novel facts related to job search behavior of employed and unemployed job seekers. Several studies have also used field experiments to examine similar research questions.[8]

Our paper provides a first comprehensive examination of the nature and consequences of explicit gender requests in job ads within the distinctive Indian context, where female labor force participation rates are low and gender disparities in the labor market are larger in comparison to China, Indonesia or Mexico. We construct implicit gender associations from textual analysis to further investigate how gender wage disparities and applicant behavior are shaped by the interaction of explicit gender requests *and* implicit gender associations. Our work is the first to identify words associated with explicit female and male gender preferences in job ads, using machine learning methods hitherto not applied in the field of economics. These words indicate underlying gender associations or stereotypes and are informative for researchers studying similar labor market contexts. Importantly and more generally, our results highlight how the text contained in a job ad matters for search behaviors.

In the next section we provide a detailed description of our data set, as well as different variables that we construct from the underlying job ad text. In Section 3 we discuss our empirical methodology while section 3.2 presents our estimation results. In Section 4 we describe the gender word associations we identify by carrying out further textual analysis and their effect on applicant

[7] These include Forsythe et al. (2020) for the US, Chiplunkar et al. (2020) for India, Hayashi and Matsuda (2020) for Bangladesh and Sri Lanka, and Campos-Vazquez et al. (2020) for Mexico.

[8] Belot et al. (2017) set up a field experiment to find that experimentally manipulated high wage jobs receive significantly more applications. Ibanez and Reiner (2018) examine the effect of affirmative action statements on application decisions using three field experiments in Colombia. Flory et al. (2015) use a field experiment to examine how workers' application decisions respond to competitive work environments while Mas and Pallais (2017) use a field experiment to examine how these decisions respond to non-wage job attributes.

behavior. Section **??** concludes.

2 Data

We analyse data from a leading job portal in India which primarily caters to young job seekers. Job seekers can create a profile for free and start applying to posted ads while employers need to pay a fee to post ads and view applicants (≈ USD 20). We use data on the population of jobs advertised on the portal with a last date of application between 24^{th} July 2018 and 25^{th} February 2020 together with data on all applications made to these ads. In our analysis we use data on 'active' job ads and job seekers; so we use job ads to which at least one male or female job seeker applied, and job seekers who applied to least one ad during this time.

Job seekers can view all jobs advertised on the portal and sort these by date of posting or popularity. They can also filter jobs based on job role, sector, location, education and type of job (govt/private). Job seekers who additionally register for a premium service are provided with specific job recommendations and alerts on new jobs by e-mail. The proportion of job seekers who registered for this service in our data was ≈ 0.5%; hence, the chances that applications are driven by matching algorithms used by the portal are negligible.

2.1 Job ads

There were a total of 196,821 job ads posted on the portal during this time. We exclude ads which had a location outside India and which had an application window of less than a day or more than four months (120 days), leaving us with 1,88,857 job ads. Next we drop duplicate job ads, where a duplicate ad was posted within a month of the original ad; this leaves us with 1,75,126 unique job ads.[9] When examining applicant behaviors we aggregate applications across duplicated ads to ensure we use data on *all* job seekers who apply to a job ad. We further drop job ads which had no male or female applicants which leaves us with 1,71,960 ads. We also restrict the sample to job ads that explicitly mention an education and experience requirement (which reduces the sample to 1,71,940 ads) and job ads that specify cities within a single Indian state as the location of the

[9] Approximately 70% of duplicate job ads were posted within a month of the original ad. We keep duplicate job ads posted more than a month after the original job ad as separate job ads since these are likely to reference new vacancies.

job (which reduces the sample to 1,58,249 ads).[10] We further restrict the sample to those jobs for which we obtain an occupational classification based on the method described in Section 2.2, leaving us with a final sample of 1,57,888 job ads.

We construct variables indicating an employer's gender and other requirements by carrying out a text search using the job title and description for each job ad in our sample.[11] In constructing a variable for gender requirement we make use of this text since words such as 'female only' or 'female preferred' (conversely 'male only' or 'male preferred') tend to appear here. We search the text for the following words which indicate an explicit female preference: 'female', 'females', 'woman', 'women', 'girl', 'girls', 'lady' or 'ladies'. Similarly, we undertake a search for the following words which indicate an explicit male preference: 'male', 'males', 'man', 'men', 'guy', 'guys', 'boy', 'boys', 'gent' and 'gents'. Some job ads include words related to both genders in the job title or description. We categorise such job ads as having no explicit gender preference, together with ads which did not include words related to either gender. About 4.2% of the job ads in our sample have an explicit female preference (F jobs), 3.5% have an explicit male preference (M jobs) and the rest have no explicit gender preference (N jobs).[12]

To construct a variable indicating whether an employer has an age requirement, we split a job description if the following words appear: 'years of age', 'years old', 'years to', 'age', 'age limit'. We examine the 25 characters before and after the split. We search for numerals starting from 18 to 45 (since 45 is the maximum numeral found across all ads) among these characters and create variables for each number. If an ad has two numbers, the minimum of these is taken to specify the minimum age requirement and the maximum is taken to specify the maximum age requirement. In jobs where only one numeral appears, we combine it with words such as 'above', 'below', 'more than' and 'not above', 'not below', 'not less' to determine whether the age specified is a minimum or maximum required age.

Finally, we create dummy variables indicating the presence of a beauty requirement in an ad

[10] We find that restricting the sample of jobs to those that specify cities in a single state as the location of a job does not change the distribution of observable characteristics of the sample of job ads; this comparison is available on request.

[11] The portal does not have a separate field that allows employers to directly state the preferred gender for an advertised job to job seekers.

[12] The fraction of F and M jobs we find are smaller than those reported by Chowdhury et al. (2018) using data from *Babajob*. This could be because, unlike the job portal we use, *Babajob* had a separate field where employers could directly state the preferred gender to job seekers. Chowdhury et al. (2018) found that a third of all employers used this field. Of the total job ads on *Babajob*, 21% preferred men and 14% preferred women.

and indicating whether a job requires working a night shift since these features may be correlated with gender preferences in a job. To identify whether a job has a preference for beauty, we undertake a word search for: ‘height’, ‘weight’, ‘beautiful’, ‘charming’, ‘delightful’, ‘pretty’, ‘attractive’ (ignoring a combination of words that specify an attractive salary or package), ‘good looking’, ‘nice looking’, ‘complexion’, ‘pleasing’, ‘appearance’ and ‘handsome’ within the job description of an ad.[13] Similarly, we create a dummy variable for the presence of a night shift requirement in a job ad by carrying out a word search for: ‘night-shift’, ‘night shift’ or ‘night’. We exclude ads if the job description specifically mentions ‘no night shift’ or that transportation will be made available like ‘night shift fully secured’, ‘cab drop’, ‘drop facility’, ‘cab facility’, ‘dropping available’, ‘pick’ and ‘drop’.

A very small fraction of jobs advertised on the portal either did not specify an education requirement or specified it as none (or illiterate). We keep these ads in our estimation sample, and group these together with ads requiring a secondary education or less as the base category in our empirical analysis. In general, N tend to have higher education requirements than either F or M jobs. For instance, N jobs are *less* likely to require only a secondary and senior secondary education (as opposed to higher education categories of graduate and postgraduate) than F or M jobs (Appendix Table A.1). A lower fraction of M jobs require a graduate or postgraduate degree compared to either N or F jobs. F jobs, in turn, are far more likely than M jobs to require a graduate degree in a non-STEM subject. Consistent with the portal catering primarily to young job seekers, we find that most job ads (at $\approx 67\%$) require less than one year of experience. We also find that N jobs are more likely to list two or more years of experience compared to other jobs. Ads specifying a gender preference are also more likely to specify other preferences, such as those related to age or beauty. We find that M jobs are more likely to also specify an age preference and these jobs tend to specify a higher minimum and maximum required age than F or N jobs. F jobs are most likely to also specify a beauty requirement while M jobs are most likely to require working night shifts.

[13] To find words related to beauty, we started out with an initial list of beauty related words such as ‘beautiful’ and ‘handsome’. We then appended to this list by considering the cosine similarity of vector representation of these words with other words using the unsupervised GloVe algorithm (Pennington et al., 2014). The 300 dimensional pre-trained word vectors have been obtained by training the algorithm on web data from common crawl, and comprise 2.2 million unique words. Cosine similarity between any two vectors is a score $\in [0, 1]$, which in this case indicates the relatedness of any two words in terms of the context in which they appear on the internet, and can essentially help identify synonyms.

Employers include a wage range in 88% of the job ads advertised on the portal that are in our sample. Wages are more likely to be missing for jobs requiring higher education and experience; thus, the sample of job ads with wage information is a somewhat selected sample of lower skill jobs. Nevertheless, we are able to observe wages for a far higher fraction of job ads in our sample than existing studies.[14] The mean of the mid-point of the wage range is 221 thousand rupees per year. This is higher than the national average of salaries earned by urban Indian workers with an age distribution similar to candidates on the portal, indicating that our estimation sample consists of relatively high skill urban jobs. N jobs have the highest mean wage while M jobs have a higher mean wage than F jobs. The wage distribution for N jobs is also shifted to the right of the wage distribution for M jobs which, in turn, is shifted to the right of the wage distribution for F jobs (Figure 2(a)).

The share of female applicants to N jobs is 32%. This is because there are fewer female applicants on the portal compared to male applicants (Appendix Table A.2). For F jobs this share rises to 52% while for M jobs it falls to 13%. This indicates that there is some compliance with explicit gender requirements in job ads but this compliance is far from perfect. Overall compliance with gender in F and M jobs i.e. percent applications that are of the requested gender is 68%. In order to account for compliance that can occur by chance (expected compliance) due to the distribution of job and candidate characteristics on the portal, we use Cohen's kappa.[15] Cohen's kappa κ for compliance with gender requirements is 35%. Compliance with education and experience requirements i.e. the percentage of applications that have at least as much education or experience as requested, across jobs ads, is 98% ($\kappa = 97\%$) and 32% ($\kappa = 25\%$).[16] Thus, compliance with gender requirements is lower than with education requirements but higher than with experience requirements.

There are about 41 applications per job ad, on average. The average number of applications to F jobs is less than half of this, at about 17, while the average number of applications to M jobs is about 31. This indicates that explicit gender preferences lead to a substantial reduction in the

[14] In comparison wages are advertised in just 16.4% of job ads in Kuhn and Shen (2013) using a Chinese job portal and 20% of job ads in Marinescu and Wolthoff (2020) using *Careerbuilder*.

[15] Cohen's kappa is defined as $\kappa \equiv \frac{Compliance_{observed} - Compliance_{expected}}{1 - Compliance_{expected}}$. The component of compliance on gender that is expected to occur by chance is 53%.

[16] These compliance figures are calculated after dropping jobs which have no education requirement and a minimum experience requirement of 0 years.

number of applications, particularly by job seekers of the opposite gender to the preferred one.

2.2 Job titles and occupations

Job ads also include information on which role a particular job belongs to, out of 33 job roles pre-specified by the portal. However, these job roles are too coarse to characterize occupation for a job ad. Marinescu and Wolthoff (2020) use data from *Careerbuilder* in the US to show that job titles can provide a much finer classification of occupations since titles not only capture the job role, but also the hierarchy and specialization within a role; they also find that words contained in job titles are predictive of wages as well as applications. Figure 1 shows word clouds of job titles in F, M and N jobs in our sample. As may be seen, job titles such as 'telecaller' and 'office executive' occur with high frequency among F jobs while titles such as 'delivery boy' and 'sales executive' occur with high frequency among M jobs. This indicates that explicit gender preferences operate to maintain existing occupational gender stereotypes.

However, job titles may also vary due to noise in the word choice of an ad without any meaningful difference in content. Therefore, we use an unsupervised machine learning technique to classify semantically similar job titles into dis-aggregate occupation categories. Specifically, we use the collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM) proposed by Yin and Wang (2014), and apply it to text contained in job titles. GSDMM is very effective for short text topic modeling, outperforming Latent Dirichlet Allocation (LDA) and several other methods at this task (Qiang et al., 2020). GSDMM makes the assumption that each document (or in our case, job title) comprises a single topic, an assumption suitable for short texts. The algorithm probabilistically combines documents into groups such that documents in the same group contain a similar set of words, whereas documents in different groups contain a different set of words. We provide details of the data pre-processing steps, algorithm, and our hyperparameter choice in Technical Appendix Section B.1. The final number of topics (or dis-aggregate occupation categories) discovered by GSDMM for our sample of job ads is 483.

Our empirical results are largely robust to an alternative manual clustering of job ads based on existence of word unigrams, bigrams and trigrams in job titles which has been used in the existing literature (Banfi and Villena-Roldan, 2019; Marinescu and Wolthoff, 2020).[17] To implement this

[17] We discuss estimation results using the alternative categorization in Section **??**.

alternative manual clustering we calculated n-gram counts after removing duplicate job ads.[18] We then classified jobs on the basis of the most frequently occurring trigrams in job titles, subject to the trigram existing in at least 50 job titles. The remaining jobs are classified based on the most frequently occurring bigrams, and then unigrams in the job title, with the restriction that the bigrams and unigrams occur in at least 100 job titles. The precedence given to higher order n-gram followed by their frequency of occurrence ensures that each job ad is classified into no more than one cluster or occupation category. This way we obtain a total of 747 occupation categories.[19]

We prefer GSDMM to a manual classification of job titles since it provides dimension reduction based on co-occurrence of words in the corpus of job titles. This is accomplished by probabilistically clustering together documents which do not share any common word between them, but are linked together through sharing common word(s) with some other documents that act as a bridge between the two. For instance, the jobs titled 'english transcriber' and 'japanese translator' are assigned the same cluster as they are linked through 'transcriber translator'. These jobs cannot be assigned the same cluster using the manual classification as they do not share any common word. This also ensures that most of the job ads in the topic model get assigned to meaningful clusters. In contrast, over 5,800 jobs could not be assigned to any cluster using the manual classification because the word n-grams contained in them occur with a low frequency across the corpus.

2.3 Job seekers

We also use data on 1.06 million job seekers who applied to at least one job ad using the portal; descriptive statistics for job seekers by gender are given in Appendix Table A.2. There are 0.37 million female and 0.69 million male job seekers. The smaller number of female job seekers is consistent with lower female labor force participation rates in urban India compared to males (Appendix Table A.3). Notably, while the labor force participation rate of men is about three times that of women, there are only slightly less than twice as many male job seekers on the portal as female job seekers. Also, once female job seekers start searching for jobs using the portal, they make a similar number of job applications, on average, as male job seekers. Most job seekers on the portal (at 86.5%) have a graduate or post-graduate degree with female job seekers being more

[18] For the purpose of this classification job ads made by the same employer, with the same job title and job description are considered duplicates.

[19] We could not classify around 3% of job ads to any occupation using this method.

likely to have a postgraduate degree. Job seekers are also relatively young, with an average age of 24 years; female job seekers are slightly younger than male job seekers and have less experience. About 76% of job seekers have less than a year of experience, again indicating that the portal caters primarily to young job seekers. Lastly, female job seekers (unconditionally) apply to job ads with slightly higher posted wages than male job seekers. However, female job seekers tend to have more education than male job seekers; in fact, conditional on candidate characteristics, women apply to job ads with 3% lower posted annual wages than do men.[20]

We compare job seekers on the portal with the urban working age population in India using the Periodic Labor Force survey (PLFS) from 2017–18, which is a nationally representative survey of employment in India. Appendix Table A.3 Panel A uses the PLFS to give the average annual earnings for those in casual or salaried employment among working age adults (age 16-60) in urban Indian districts (with $\geq 70\%$ urban population).[21] Advertised wages on the portal are higher than this nationally representative sample by $\approx$ Rs. 20,000 per annum. However, wages in PLFS could also be high because it has older, more experienced workers. To make the PLFS sample comparable to the age group catered to by the online job portal we only keep adults who are 18-32 years old in Appendix Table A.3 Panel B since around 95% of the job seekers on the portal are in this age group. The gap in annual earnings increases to Rs 45,000 per annum, or the average advertised wage is now 25% higher on the job portal. Thus, the job portal caters to younger, and inexperienced but more educated and skilled workers.

These patterns can also be observed in Figure 2. The wage distribution for employed men and women using the nationally representative PLFS data is centered at a lower log wage and more spread out in comparison to the distribution of posted wages on the job portal. This is particularly so for female wage distributions, indicating that gender wage disparities among employed Indian workers exceed gender disparities in posted wages across F and N jobs that we observe on the portal. The larger wage disparities among employed workers could also arise due to differential matching of workers across jobs after the application stage, wherein females get matched to relatively lower

[20] We also estimate an alternative specification at the application rather than candidate level to estimate the gender wage gap in applications in which we include occupation controls. We regress the log of the posted wage for the applied job on job ad and candidate characteristics, giving each candidate equal weight, and continue to find that women, on average, apply to jobs with 1.8% lower wages than men.

[21] Annual earnings are obtained by multiplying monthly earnings by 12 for salaried workers and weekly earnings by 52 for daily wage workers.

wage jobs. To look at this, we restrict the PLFS sample to employed workers with more than school education and those between 18-32 years old. We then regress the log of wage on a gender indicator, worker education, age and occupation and find that women earn 8% lower wages than men in this sample. On the portal, female applicants apply to jobs that on an average offer 2% lower annual wages after including controls for candidate education, age and occupation times location. This indicates that around 25% of the gender wage gap among educated individuals in the Indian labor market is driven by applications of female job seekers to lower wage jobs.

2.4 Implicit *femaleness* and *maleness*

The text contained in a job ad may also convey an implicit signal to a candidate about whether the employer posting the ad prefers a female or a male candidate for the job even in the absence of an explicit gender preference. We define the implicit "*femaleness*" (F_p) and "*maleness*" (M_p) of a job as:

$$F_p \equiv \text{Prob(explicit female request} \mid \text{job text)}$$

$$M_p \equiv \text{Prob(explicit male request} \mid \text{job text)}$$

We use supervised machine learning to infer F_p and M_p associated with each job ad based on the job text. We implement a Multinomial Logistic Regression (LR) classifier with balanced class weights. We concatenate the job title and description to reflect the complete job text.[22] We follow standard pre-processing steps as in the natural language processing literature which we outline in Technical Appendix Section B.2. We then convert our corpus of processed documents to their bag-of-n-grams representation using term frequency-inverse document frequency (TF-IDF) vectors—which we use as inputs to the model. The output class in the model can take three values depending on the employer making an explicit request for men, women, or no gender request. We perform stratified 10-folds cross-validation wherein we split the data in 10 parts and preserve the percentage of the sample that belongs to each class. F_p and M_p are then the estimated probabilities of a document belonging to the female or male class when it belongs to the test set. We discuss implementation details for TF-IDF and k-fold cross-validation in Technical Appendix Sections B.3 and B.4.

[22] We use a total 196,857 jobs which include an additional set of jobs provided to us by the portal to increase data points for the classification model. We use balanced class weights since the classes are highly imbalanced and only a small fraction of total jobs explicitly request a female or a male.

Conditional on the employer making an explicit gender request, the model correctly predicts requests for females and males in 74.43% and 72.18% of job ads when they are part of the test set. The corresponding figure when employers do not explicitly request a gender is 79.50%. Furthermore, F_p and M_p capture employer requests very well, with correlations of 0.38 and 0.44 with binary variables indicating explicit female and male requests.[23] F_p takes high values for jobs with titles such as 'beautician', 'personal secretary' and 'school teacher' while M_p takes high values for jobs with titles such as 'cargo loader', 'delivery executive' and 'network engineer'. Even for the same job title, F_p and M_p can greatly vary based on the job description. For example, for the job titled 'business development executive' F_p is high when the job description mentions working from home or restarting career, while M_p is high when the job involves travel or working night shifts. Similarly, for 'sales executive', high F_p is associated with jobs emphasizing appearance or communication skills, whereas M_p is high for jobs involving field work. Figure 3 shows, as expected, that F_P on average is high for F jobs. In contrast, M_p on average is higher for M jobs. For N jobs, both F_P and M_p have similar distributions.

We differ from Kuhn et al. (2020) in calculating F_p and M_p in multiple ways. First, as opposed to using only the job title, we make use of both the job description as well as the title in predicting implicit gender association of a job posting. This is reasonable as M_p and F_p often vary greatly even for the same job title based on the precise description of the job and we expect that the candidates will make use of both to infer their hiring prospects and suitability for a job. This is validated when we include occupation fixed effects based on job titles in the specification while inferring the impact of F_p and M_p on applicant behavior. Including occupation fixed effects based on job titles reduces the responsiveness of job seekers to implicit gender associations but does not eliminate it. This shows that candidates use information contained in the job description in conjunction with job titles to infer their suitability for a job. Second, we improve measures F_p and M_p using an LR classifier instead of the Bernoulli Naive Bayes (NB) classifier. For comparison we also implemented the NB classifier on our data using the methodology in Kuhn et al. (2020). We found that the NB classifier does not perform well in our context. It gives a much worse measure of F_p and M_p in our

[23] The correlations using balanced class weights, i.e. weighted by inverse frequency of observations belonging to the two classes, are 0.71 and 0.74 for F_p and M_p with binary variables indicating explicit female and male requests respectively.

data with correlations of 0.23 and 0.22 with explicit employer requests for women and men.[24]

3 Gender preferences of employers

3.1 Empirical methodology

We first examine characteristics of jobs where employers exhibit explicit gender preferences. The regressions we estimate are variations of the following specification:

$$Y_{ijst}^{k} = \alpha^{k} + \beta^{k} X_{ijst} + \gamma_{j\times s} + \phi_{t} + \epsilon_{ijst}^{k} \tag{3.1}$$

where the superscript $k \in \{FM, M\}$ indicates two different dependent variables capturing the *presence* and *direction* of explicit gender preferences. The first dependent variable Y_{ijct}^{FM} is a binary outcome which takes the value one if there is either an explicit male or female preference exhibited in job ad i which advertises for a job of occupation j in state s at time (or month and year) t. The second dependent variable Y_{ijct}^{M} takes on three values: minus one if there is an explicit female preference, zero if there is no gender preference and one if there is an explicit male preference exhibited in a job ad.[25] X_{ijst} is a set of job ad specific variables including a set of dummy variables for education requirements, a set of dummy variables for experience requirements, dummy variables for the presence of age and beauty requirements, a dummy variable for the presence of a night shift requirement and a quadratic in log advertised wage. In our preferred specification we include occupation and state fixed effects ($\gamma_{j\times s}$) as well as time (or month and year) fixed effects (ϕ_t). We use a detailed categorisation of jobs to occupations, as described in Section 2.2, with 483 distinct occupation categories derived from job titles. The use of fixed effects ensures we use *within* occupation and state variation only to identify the effect of different variables on whether a job ad exhibits a gender (or male) preference. We cluster standard errors by occupation and state.

The wage difference across ads that explicitly request men and women is available from es-

[24] The requests for females and males conditional on an employers' explicit gender request are correctly predicted in 76.58% and 75.71% of job ads. For jobs that make no explicit gender request, correct predictions are made in 69.29% and 70.31% of job ads in the model for F_p and M_p. The correlations are much lower than those for the LR model discussed previously. This demonstrates that even though NB is a reasonable classifier, it does a poor job of estimating probabilities associated with the classes.

[25] While we estimate and report linear regressions in the paper, we also carried out non-linear estimations (probit and ordered probit) on these dependent variables using coarser job role and state fixed effects. We find that our results are largely unchanged; results are available on request.

timatation of equation 3.1 when Y^{M}_{ijct} is the dependent variable. In a separate set of regressions we also examine whether wage differences exist when the text of a job ad is predictive of explicit gender preferences by the employer (Section 2.4), separately for F, N and M jobs. The regressions we estimate are variations of the following specification:

$$\ln W_{ijst} = \alpha^{W} + \lambda^{W} F_{p,ijst} + \nu^{W} M_{p,ijst} + \beta^{W} X_{ijst} + \gamma_{j\times s} + \phi_t + \varepsilon_{ijst} \tag{3.2}$$

where $\ln W_{ijst}$ is the log wage advertised in a job ad.[26] $F_{p,ijst}$ is a measure of implicit *femaleness* and $M_{p,ijst}$ is a measure of implicit *maleness*. The coefficients on these variables (λ^{W} and ν^{W}) tell us how the advertised log wage changes as predicted *femaleness* (*maleness*) of a job ad increases from zero to one i.e. the probability of a job being a stereotypical female (male) increases from zero to one, everything else equal. X_{ijst} is a set of job ad specific variables (or dummy variables for education and experience requirements).[27] In our preferred specification we include occupation and state fixed effects ($\gamma_{j\times s}$) as well as time fixed effects (ϕ_t). As before, we cluster standard errors by occupation and state.

We also examine how explicit gender preferences affect job seeker's responses to an ad by estimating variations of the following specification:

$$Y^{TA}_{ijst} = \alpha^{TA} + \pi^{TA} F_{ijst} + \theta^{TA} M_{ijst} + \beta^{TA} X_{ijst} + \gamma_{j\times s} + \phi_t + \mu_{ijst} \tag{3.3}$$

where Y^{TA}_{ijst} is the total number of applications to a job ad. F_{ijst} is a binary variable taking the value one if ad i has an explicit female preference and zero otherwise. Similarly, M_{ijst} is a binary variable taking the value one if ad i has an explicit male preference, and zero otherwise. The coefficients on these binary variables (π^{TA} and θ^{TA}) give the difference in total applications sent to ads that exhibit an explicit female or male preference in comparison to ads that exhibit no such preference (the base category), everything else equal. X_{ijst} is a set of job ad specific variables which are the same as those specified in equation (3.1). In our preferred specification we include occupation and state fixed effects ($\gamma_{j\times s}$) as well as time fixed effects (ϕ_t); as before, we also cluster standard errors

[26] Since job ads generally include a wage range we take the mid-point of this range and take the log of this mid-point.

[27] We do not include dummies for the presence of age, beauty or a night shift requirement since the words used to construct these variables are also highly predictive of an explicit gender preference.

by occupation and state.

In order to examine job seekers compliance with the gender requirement set by the employer, we estimate variations of the following specification:

$$Y^S_{ijst} = \alpha^S + \pi^S F_{ijst} + \theta^S M_{ijst} + \beta^S X_{ijst} + \gamma_{j\times s} + \phi_t + \mu_{ijst} \tag{3.4}$$

where Y^S_{ijst} is the share of female applicants to a job ad. Apart from the difference in the dependent variable, the specification in equation (3.4) is similar to (3.3); coefficients on the binary variables (π^S and θ^S) give the difference in the share of female applicants across ads that exhibit an explicit female or male preference and those that exhibit no such preference (the base category), everything else equal. These regressions are also weighted by the total number of male and female applications made to a job ad.

Finally, we examine how F_p (M_p) derived from the text of a job ad affect applicant behaviors; specifically we regress the share of female (male) applicants to a job on explicit gender requests as well as quartics in F_p and M_p, following the strategy adopted by Kuhn et al. (2020). We include the set of controls in equation (3.4), focusing on specifications that either exclude or include occupation and state fixed effects.[28] Further, we interact the quartics in F_p and M_p with explicit gender requests and use these as additional explanatory variables. Using these regression estimates, we predict the share of female (male) applicants as a function of F_p (M_p) for each type of job (F, M and N jobs).

3.2 Results

Columns (I)-(III) of Table 1 give results from estimation of equation (3.1) when the dependent variable is Y^{FM}_{ijct}. Column (I) includes all controls apart from the advertised wage as well as time (or month and year) fixed effects. Column (II) adds occupation and state fixed effects while column (III) additionally controls for a quadratic in log advertised wage.[29] The results support a negative skill-targeting relationship i.e. jobs with a higher skill requirement (a higher education requirement or log advertised wage) are *less* likely to have an explicit gender preference; however, we find mixed

[28] We do not include wage controls in these regressions to ensure we use the full sample of job ads. We also do not include dummies for age, beauty or night shift requirements since the words used to construct these measures are also highly predictive of explicit gender preferences.

[29] Since wages are not posted for all jobs, some observations are lost when moving to from column (II) to column (III).

results for experience.[30] We also find that the presence of an age or beauty requirement leads to an increased probability that a vacancy has an explicit gender preference (columns (II) and (III), Table 1).

Columns (IV)-(VI) of Table 1 give results from estimation of equation (3.1) when the outcome of interest is male preference in a job ad. Here we find that jobs having an explicit male preference are less likely to belong to higher education categories compared to the base category. This effect becomes slightly smaller with the addition of occupation and state fixed effects but remains highly statistically significant. We find that a higher advertised wage is associated with an increased preference for men, with this effect declining at higher wages. We also find that the presence of age requirements and working night shifts leads to increased preference for men while the presence of a beauty requirement leads to a reduced preference for men (columns (V) and (VI), Table 1).[31]

Jobs with an explicit preference for men offer higher wages than jobs with an explicit preference for women. We next examine estimation results of equation (3.2) (estimated separately for F, N and M jobs) to evaluate the effect of implicit *femaleness* (*maleness*) on the advertised wage; the results are reported in Table 2. As expected, higher education and experience requirements lead to an increase in the advertised wage for all kinds of jobs. For N jobs we find that an increase in *femaleness* from 0 to 1 leads to a reduction in the offered wage by 39%, without occupation and state controls (column (III), Table 2). Once occupation and state controls are included, the effect of *femaleness* on offered wages drops to 27% but remains highly statistically significant (column (IV), Table 2). This coefficient estimate translates to a decrease in the advertised wage of 5.4% for a one standard deviation increase in the *femaleness* measure ($SD = 0.2$). On the other hand, an increase in *maleness* is associated with a smaller decline in the log wage; the p-value from a test of difference in coefficients on *femaleness* and *maleness* is very close to zero. This provides evidence that jobs with higher female association (or jobs where applicants are likely to infer that

[30] When occupation and state fixed effects are not included, jobs that specify a higher experience category ($>$ 2 years relative to $0-1$ years) are associated with a lower probability of exhibiting a gender preference. When occupation and state fixed effects as well as wage controls are included, higher experience is associated with an *increased* probability of exhibiting an explicit gender preference. This reversal occurs due to inclusion of controls for advertised wages; experience is positively correlated with advertised wage, and wages have a strong negative correlation with the probability of a job ad exhibiting a gender preference.

[31] To further investigate whether a male preference in a job ad is associated with a higher maximum age requirement (or to check for evidence of the 'age twist' in explicit gender preferences) we also estimated regressions on the sub-set of ads which specify a maximum required age and used maximum required age instead of any age requirement as the explanatory variable of interest. While we found a positive effect of maximum required age on male preference, we did not find that this effect to be statistically significant. These results are available on request.

the employer would prefer a female from reading the job text) offer systematically lower wages even when the job ad does not exhibit any explicit gender preference. We find a similar pattern for F and M jobs, but the negative effect of the *femaleness* measure on log wage is smaller in these jobs than for N jobs, although it is still statistically significant.[32] The negative effect of *maleness* on log wage in F jobs is indistinguishable from zero but this negative effect becomes larger (and almost as large as the negative effect of $femaleness$) in M jobs.

To examine the effect of explicit gender preferences on applicant behaviors we estimate and report regressions specified by equation (3.3) where the outcome variable is the total number of applications to a job ad; the results are reported in columns (I)-(III) of Table 3. We find that the number of applications are reduced dramatically (≈ 20) if the job ad exhibits an explicit female preference in the absence of occupation and state fixed effects. Once we include these fixed effects the decline is smaller but continues to be statistically significant ($\approx 5-8$). On the other hand, the change in the number of applications to job ads that exhibit an explicit male preference (vs no gender preference) is not statistically significantly different from zero. We also find that the number of applications to jobs with higher education requirements tend to increase and then fall. Notably, job postings which specify a graduate degree in a STEM subject see the largest increase in the number of applications compared to the base category. The number of applications to job ads with higher experience requirements is reduced but there isn't a statistically significant effect of advertised wages on the number of applicants. This could possibly reflect a slack youth labor market in India where over the last decade the unemployment rate, especially among the educated youth, has been on a rise.[33]

Next, we estimate and report the regressions specified by equation (3.4) where the outcome is the fraction of female applicants to a job ad; the results are reported in columns (IV)-(VI) of Table 3. We find that the fraction of female applicants to a job ad increases by $15.4-15.6$ percentage points when the ad exhibits an explicit female preference and reduces by $9.5-9.9$ percentage points when the ad exhibits an explicit male preference (columns (V)-(VI), Table 3). These translate to an increase of 48% and decrease of 30% in the share of female applicants to a job ad, which are substantially large effects. In addition we find that a higher fraction of women apply to job ads

[32] In later results we also find that a higher fraction of women apply to these low-wage jobs (column (VI), Table 3).

[33] The unemployment rate for urban young men reached 18% in 2017-18. See: Mint Report.

which have higher education and lower experience requirements. This is likely to be driven by more educated and younger women on the portal (Appendix Table A.2). A smaller fraction of women apply to vacancies which specify working night shifts as part of the job description. We also find that a somewhat higher fraction of women apply to vacancies with a lower advertised wage.

Lastly, we examine the response of share of female and male applicants to the implicit association of the job text with a gender preference across F, N and M jobs. Figure 4(a) gives the predicted share of female (male) applicants as F_p (M_p) changes while keeping M_p (F_p) constant at its mean level using a specification that excludes occupation and state fixed effects. Figure 4(b) gives the same variation but uses a specification with occupation and state fixed effects. Strikingly, Figure 4(a) shows that the predicted share of female applicants increases as F_p rises (or as we switch from 'mechanical engineer' to 'receptionist' jobs) not only for N jobs but also in F and M jobs. The increase is almost linear for N jobs and as F_p increases from zero to one, the share of female applicants increase from 35 percentage point to 45 percentage points, a 31% increase. On the other hand, the rise for F and M jobs is not consistent; it is more rapid at low F_p for F jobs (by 46% as F_p increases from zero to half) and at high F_p for M jobs (by 200% as F_p increases from half to one), though the effects are imprecise. The predicted share of male applicants also increases as M_p increases (as we switch from 'telecaller' to 'electrician/IT hardware engineer' jobs) for M, N and F jobs; however, there is a decline in this share at high M_p for F jobs. The effect is highest and most consistent for N jobs, where an increase in M_p from zero to one increase the share of male applicants by 24%. For M jobs as well the male applicant share by 22% as M_p increases from zero to one. Notably, we find that the difference in predicted share of male applicants across M and N jobs is generally smaller and further declines as M_p increases in comparison with the difference in predicted share of female applicants across F and N jobs as F_p increases.

Figure 4(b) uses within occupation and state variation only and shows that as F_p associated with a job increases (or as we switch to jobs with an increasingly female job description *within* the same occupation and state) from zero to one, the predicted share of female applicants increases from 34 percentage point to 39 percentage point (by 16%) for N jobs. The share is more responsive to changes in F_p for F and M jobs as well. At low F_p (< 0.3) the predicted share of female applicants increases rapidly as F_p increases in F jobs (by 35%) while at high F_p the predicted share of female applicants is relatively constant in these jobs. In M jobs this share decreases as F_p increases from low to mid F_p but then rapidly increases as F_p increases at high F_p values; the turning point occurs at around $F_p = 0.7$ with an increase of almost 180% again. On the other hand, as M_p associated with a job increases (or as we move along jobs with an increasingly male job description *within* the same occupation and state) the predicted share of male applicants increases consistently for N jobs by 14% as M_p increases from zero to one. For M jobs there is only a slight increase as M_p increases from zero to one by 12%, though imprecise; there is no increase in male applicants share with a rise in M_p for F jobs. As before, the difference in predicted share of male applicants across M and N jobs is generally smaller as M_p increases than the difference in the predicted share of female applicants across F and N jobs as F_p increases.[34]

We also re-construct our measures of F_p and M_p using the Bernoulli NB classifier (Appendix Figure A.1(d)). We estimate similar regressions as before to find the predicted share of female (male) applicants using state fixed effects rather than occupation and state fixed effects since F_p and M_p are now constructed using text in job titles only and these job titles are also used to assign jobs to different occupations. We continue to find that the predicted female (male) applicant shares increase, as F_p (M_p) increases, across F, N and M jobs.

Our results bear similarities and differences from those reported by Kuhn et al. (2020). We also find that explicit female requests matter more for female applicant shares than explicit male requests matter for male applicant shares indicating that women are more "ambiguity averse." However, our findings show that gender associations seem to play a role in changing the gender mix of the applicant pool even in F and M jobs. For instance, in F jobs with text highly predictive of an explicit *female* preference we find that predicted female applicant shares increase as F_p increases further (although confidence intervals are wide since there are few such jobs). Importantly, these findings persist even within a given occupation in a location, though the magnitudes decline. For instance, when a female request is made, an increase in implicit femaleness from zero to one increase female applicant share by 50% across occupation and location while within a given occupation and location the share increases by 35%.

[34] Predictions at very high and very low values of F_p and M_p are not estimated very precisely since there are few observations at the endpoints.

4 Deconstructing gender preferences of employers

Our analysis show that *femaleness* and *maleness* in a job ad matter for the advertised wages as well as the share of women who apply for the job. We find these associations even when exploiting variation *within* an occupation and state, though they are stronger across occupations and states. A natural question that follows is: which words contribute to these gender associations? In this section, we deconstruct gender word associations using two methodologies. First, we open the black box of *femaleness* and *maleness* which are predictive of explicit gender requests by employers by looking at which words contribute to these. We then combine these gendered words into meaningful categories and see which ones drive changes in advertised wages and affect the female applicant share in each job type. Second, we directly uncover the gender associations held by applicants by using job descriptions for N ads, and detect which words lead to differential application rates by women.

4.1 Empirical methodology

We uncover words associated with gender requests in job ads by explaining the classification decisions of the above machine learning (ML) model.[35] We then aggregate the gendered words by categories (such as hard skills, soft skills, personality and flexibility) to examine which of these categories matters for advertised wages and female applicant shares. Our analysis allows us to understand which job attributes matter for advertised wages and whether variation in the female applicant share along these attributes reflects different willingness to pay for these by gender.

We use the Local Interpretable Model-agnostic Explanations (LIME) algorithm proposed by Ribeiro et al. (2016) to explain which words in job texts correspond to explicit gender preferences of employers. LIME can explain the predictions of any classifier and overcomes the *black box* nature of complex ML models. It estimates the extent to which each input x contributes towards making a specific classification decision by perturbing x (in our case randomly removing words from a job ad) and then obtaining predictions $f(x)$ returned by the ML model f. This gives a new data set of inputs with predictions on which an interpretable *surrogate* model is trained.[36]

[35] In a recent paper, Arceo-Gómez et al. (2020) use job ads to classify the most common words based on their relative frequency of occurrence in female or male requesting ads.

[36] To approximate a black-box model locally (instead of globally), the weights are assigned based on the similarity of the perturbed input to the original job ad.

LIME has been used to explain predictions made by ML models such as deep neural networks (DNN) in many applications ranging from biomedical domain, music content analysis, and computer vision to natural language processing (NLP). We introduce LIME to the domain of economics and demonstrate how labeled text data based on explicit gender requests in job ads can be used to explain decisions of the underlying ML model and systematically extract words that reflect gender associations. Explainability in itself might be desirable to assess the validity and the generalizability of the model, and hence to gain trust in its predictions.[37] We use LIME to answer what change in words will make a job ad more or less female or male targeted.

We outline the steps to explain the predictions of our Multinomial Logistic Regression model and to assign the contribution of individual words to the female, male, and neutral class below.

Word scores by gender: We map the classification scores returned by the Multinomial Logistic Regression model into the input space using the LIME algorithm over the test set documents. This allows us to assign a relevance score, $R^G_{w,i}$, to every word w in each job ad i to indicate the importance of that word to each class $G \in \{F, N, M\}$.[38] Figure 5 shows a heat map visualization of words in distinctive job ads with explicit female and male preferences. Panel (a) refers to correctly classified F jobs, while panel (b) refers to correctly classified M jobs. Job ads (i), (ii) and (iii) in both the panels refer to jobs titled 'software trainee', 'business development manager', and 'sales market executive' respectively. We find that words representing personality, appearance, communication skills and basic computer proficiency are associated with request for women. On the other hand, working in rotational shifts, field work and travel requirements are associated with requests for men.

Now we explain how the relevance scores are used to arrive at overall gender associations for each word. We take the median relevance score for every word for both the female and the male classes MR^G_w. A positive median score on the female (male) class ($MR^G_w > 0$) here indicates that the word w is associated with requesting a female (male). However, a word that is associated with

[37] A model can spuriously achieve a high accuracy on the test data without learning anything meaningful. For instance, Lapuschkin et al. (2019) show that a Fisher Vectors based model which achieved high accuracy on an image classification task "misused" source tags that distinctively occurred in lower left corners of horse images instead of classifying on the basis of an actual horse image.

[38] Assigning relevance score to each word (unigram) using LIME instead of assigning scores to each unigram, bigram and trigram helps us simplify the generated explanations. It also allows the score of each word to vary depending on the context. We use the implementation of LIME available as TextExplainer (See: Link). We restrict our analysis to top 200 most relevant words for each class in a given job ad for our analysis.

a female as well as a male request may not contribute *differentially* towards either towards the female or the male class; in other words, it may merely indicate the presence of a gender request. Therefore, to obtain the net contribution of every word towards the female class, we calculate the difference in the median score for that word across the female and the male class. A positive (negative) net score for the word reflects that it contributes more towards female (male) requests in job ads.

Category scores by gender: We restrict our analyses to words that occur at least ten times in the 13,735 M and F jobs for the categorization exercise. There are 3,927 words that meet this criteria. These words constitute 95% of all the word occurrences in N jobs as well. We first classify these words manually into four categories (C): hard-skills (280 words), soft-skills (63), personality/appearance (95), and flexibility (12) associated with a job. We classify words under the category hard-skills if they are related to knowledge about a particular software, hardware or specific skills such as driving or typing. The category of soft-skills includes words that refer to communication or interpersonal skills. The third category personality/appearance refers to other personal attributes of a prospective candidate that a job requires. Lastly, job flexibility captures words related to job timings and travel requirements. The remaining words could not be classified into any of these categories (most words are generic or reflect occupation or other job and candidate specific attributes) or fall under multiple categories; we classify these words as "others".

We construct net scores for each category $C \in$ {hard-skills, soft-skills, personality, flexibility, others} by gender for every job ad i. To do this, we again use the above 3,927 words and their median relevance scores MR_w^G. We sum the median relevance scores for all words within job ad i which are also classified to a given category C in explaining F jobs, or $S_{i,F}^C = \sum_{(w \in i) \wedge (w \in C)} MR_w^F$; similarly we sum the median relevance scores for words in explaining M jobs, or $S_{i,M}^C = \sum_{(w \in i) \wedge (w \in C)} MR_w^M$. For each category, we then take the difference between the two sums to arrive at a net score towards the female class ($NS_i^C = S_{i,F}^C - S_{i,M}^C$) in each category. A positive (negative) net score for a category indicates that the job contains words that contribute towards a gender request for a female (male) more than a gender request for a male (female).

Estimation Strategy: We next examine which category of words used by the employer matter for the earlier observed relationship between implicit gender association and the advertised wage, as we as between implicit gender association and the female applicant share. Instead of using individual words here, we use the aggregate category scores to derive meaningful interpretations. We use the net scores obtained above for each category for this exercise. Additionally, we take into account the possibility that positive net scores (net positive contribution towards F jobs) and negative net scores (net positive contribution towards M jobs) in a category can have a different impact on both advertised wages and the applicant mix. Therefore, we construct gender-category variables as below:

$$FW_i^C = \mathbb{1}[NS_i^C > 0] \times NS_i^C$$

$$MW_i^C = \mathbb{1}[NS_i^C < 0] \times NS_i^C$$

For instance, for the category hard-skills two separate variables are generated—'Female (hard-skills)' and 'Male (hard-skills)'. Here, 'Female (hard-skills)' takes on the value of the net score when the net score is positive and zero otherwise. The net score for hard-skills for a given job ad will be positive if words classified under the category of hard skills contribute more towards F jobs relative to M jobs. Similarly, 'Male (hard-skills)' takes on the absolute value of net score when the net score is negative and zero otherwise. Net score for hard-skills will be negative if words classified under the category of hard skills for a given job ad contribute more towards M jobs relative to F jobs. If a job ad does not have any word in a given gender-category then it gets a zero net score for it in both 'Female (hard-skills)' and 'Male (hard-skills)'. This procedure is used to construct ten gender-category variables, two for each of the five categories (including "others"). The obtained score in each gender-category are then standardized for ease of interpretation. We report the summary statistics for the non-standardized gender-category variables in Table A.8 separately for jobs with male, female and no gender preference. The word categories reflecting implicit preference for women score the highest in F jobs. For instance, female hard-skills get an average score of 0.17, 0.11 and 0.07 in F, N and M jobs respectively. However, for words categories reflecting a preference for a male candidate do not always score the highest in M jobs. In fact, words related to hard skills used to indicate a male preference have the highest scores for N jobs at 0.16 and then for M jobs at 0.12. Notably, the scores for M jobs are higher than that for F jobs consistently on male category scores.

We use the standardized scores for each category to estimate the following regression specifications:

$$Y_{ijst} = \alpha + \sum_{C} \delta^{FW,C} FW^{C}_{ijst} + \sum_{C} \delta^{MW,C} MW^{C}_{ijst} + \beta X_{ijst} + \gamma_{j\times s} + \phi_t + \varepsilon_{ijst} \tag{4.1}$$

where $Y \in \{\ln W_{ijst}, Y^{S}_{ijst}\}$. Here, $\ln W_{ijst}$ is the log wage advertised in job ad i at time t (defined as in equation 3.2) and Y^{S}_{ijst} is the share of female applicants to job ad i. The explanatory variables include standardized Female-Category (FW^{C}_{i}) and Male-Category scores (MW^{C}_{i}) in each of the five categories. X_{ijst} is a set of job ad specific variables (or dummy variables for education and experience requirements). The coefficients of interest are $\delta^{FW,C}$ and $\delta^{MW,C}$ which give the change in the outcome variable (percent change for wage and percentage point change for female applicant share) for a one standard deviation increase in the Female-Category scores and Male-Category scores respectively, everything else equal. We control for occupation × state fixed effects as well as time fixed effects. We cluster standard errors by occupation and state. As before, regressions with female applicant share as the outcome variable are also weighted by the total number of male and female applications made to a job ad.

4.2 Results

To obtain the most relevant words for each gender under each category, we sort the words on the basis of their net scores, obtained as outlined in Section 6, within each category (C). The top words on this ordered list contribute relatively more towards female requests and the bottom words contribute relatively more towards male requests. We list at most 20 words that are mostly highly associated with requests for females and males within each category in Table 4. The results are striking and show that many words that are typically associated with male and female job roles indeed show up on the list.

Within hard skills (columns (I) and (II), Panel A), skills associated with a beautician (*nailcare, pedicure, manicure, facial, makeup*), accounting tasks and software (*ledger, expense statements, tally*), and knowledge of tools used for communication, word processing and designing (*computer, ms (office), word, ppt, zoho, coral, autocad*) and keyword analyses appear for women. For men, skills related to jobs in IT/hardware/engineering (*rcm, mysql, rf, qc, machine learning, troubleshoot*), finance (*demat, audit, receivable*) and manual repair tend to dominate. Next we look at soft skills (columns (III) and (IV), Panel A) and again find a stark distinction within required soft skills across gender. While jobs requesting women focus on communication skills, interpersonal skills and coordination to maintain customer relations (crm), those requesting men include skills requiring assertiveness or leadership such as pitching to a client, liaison, negotiating, dealing, persuading, supervising, motivating.[39]

The gender contrast is also evident in different personality traits across jobs that request women and men (columns (I) and (II), Panel B). Jobs requesting women require the candidate to be pleasing, presentable, confident, mature, careful, physical traits like height, and other characteristics such as politeness, patience, adaptability, punctuality and sincerity. However, some contrasting words like being pro-active and entrepreneurial are also present. On the other hand, personality traits such as energetic, enthusiastic, ability to handle pressure, passionate, resourceful, prompt, creative, good first impressions, ethical/honest, methodical and physical traits like chest measurement (cm) and no scars/tattoos are used when requesting a male candidate to apply for a position. Lastly, words indicating job flexibility such as work involving undertaking *skype* calls and possibility of work from *home* or *home* based work are usually associated with jobs requesting a female (column (III), Panel B).[40] On the other hand, night/evening/rotational shifts, working on weekends, possible relocation and travel (petrol/fuel) are associated with male requests (column (IV), Panel B). Overall, we find that fairly distinct skills and personality traits are associated with jobs that request men and women. Next, we use the generated net scores in the above four categories along with words in the "other" category to see how the advertised wage and the gender mix of the applicant pool is affected by these measures.

4.2.1 Gendered words and the advertised wage

Table 5 reports estimation results when the log of the advertised wage is the dependent variable in equation 4.1. Our estimates for N jobs show that an increase of one standard deviation in net scores for 'Female (hard-skills)' decreases the advertised wage significantly by 3% while an increase in net scores for 'Female (soft-skills)' and 'Female (personality)' increases the advertised wage by 0.6% and

[39] In certain words the distinction between soft skills and personality is difficult to ascertain, e.g. accommodate.

[40] The word 'home' is mostly used in the context of work from home but can also be used for home of the clients (home tutor/demo/care) and pick/drop from home facility.

3.3%) (column (III)). While, a one standard deviation increase in the net scores for 'Male (hard-skills)', 'Male (soft-skills)', 'Male (personality)' and 'Male (flexibility)' all increase the advertised wage by 1.9%, 2.1%, 1.3% and 3.4% (column (III), Table 5). Once we include occupation and state fixed effects in column (IV) of Table 5, we find that the negative effect of 'Female (hard-skills)' on the advertised wage and the positive effects of 'Female (soft-skills)', 'Female (personality)', and all male related words persists, albeit the magnitudes decline. An increase in one standard deviation of the net score for words related to 'Male (flexibility)' results in the highest increase in wages when using within occupation and state variation only, by 2.2%, while a similar increase in the net score for words related to 'Female (skills)' results in the largest decline in the advertised wage by 1.4%.

The results for F jobs in Table 5, column (II), show that an increase in one standard deviation in net score for words related to 'Female (hard-skills)' decreases the wage significantly by 2.1% within F jobs. However, if the job ad contains words that indicate 'Male (flexibility)' it is associated with a higher advertised wage. An increase in one standard deviation in net score for words related to 'Male (flexibility)' increases the wage significantly by 4.6% within F jobs. Thus, if employers want a female for a position, they are willing to pay an even higher wage premium if the job requires longer working hours, travel, relocation or night shifts. Lastly, the results for M jobs in Table 5, column (VI), show that none of the words related to women significantly matter to advertised wages within a given occupation and state. Only the presence of words related to 'Male (flexibility)' increases the advertised wage within an occupation for M jobs.

These results show that 'Female (hard-skills)' and 'Male (flexibility)' matter the most for advertised wages when we use within occupation and state variation only. While words related to 'Female (hard-skills)' decrease advertised wages, those related to 'Male (flexibility)' increase advertised wages. There is also an observed wage penalty for 'Female (soft-skills)' and a wage premium for 'Female (personality)', 'Male (hard-skills)', 'Male (personality)', but it is significant only for N jobs. These results align with female skills getting penalized in the labor market and also indicate the trade-off between job flexibility and wages.

4.2.2 Gendered words and the female applicant share

Table 6 reports estimation results when the outcome is the proportion of female applicants in equation 4.1. We estimate regressions separately for each type of job (N, F and M type) since the effect of gendered words can be different across jobs that request a particular gender versus those that do not. The results for N jobs show that an increase of one standard deviation in 'Female (hard-skills)' and 'Male (hard-skills)' increases the fraction of female applicants by 1 pp or 3% (column (III)). An increase in words related to 'Male (personality)' also do not deter women from applying. On the other hand, increased gendered words for 'Male (soft-skills)' and 'Male (flexibility)' reduce the female applicant share by 0.5 pp (1.6 %) and 0.3 pp (1%). Once occupation × state fixed effects are included, only the positive effect of 'Female (hard-skills)' and the negative effect of 'Male (flexibility)' on the female applicant share persists, and is almost equal at 0.4 pp or 1.25% (column (IV)).

On the other hand, in jobs that explicitly request a female (F jobs) none of the female gendered words in any category matter significantly (column (I) and (II)). A one standard deviation increase in net scores in the categories of 'Male (hard-skills)', 'Male (soft-skills)' and 'Male (flexibility)' reduce the female applicant share by 6 pp (11%), 1.9 pp (3.6%) and 2.6 pp (5%), thus reducing compliance with the employer's gender requirement (column (I)). However, after including occupation and state fixed effects, only 'Male (flexibility)' reduces the female applicant share significantly by 2.1 pp (4%) in column (II). For the M jobs none of the gendered words significantly matter for female applicant share (column (VI)). An increase in net scores for words related to 'Male (flexibility)' leads to a decline in female applicant share by a similar magnitude as for N jobs, but this effect is imprecisely estimated.

These results show that for N jobs, both female oriented hard skills and other female attributes matter the most in increasing the share of women applicants while male oriented job flexibility and other male attributes decrease the proportion of female applicants to a job, even when we use *within* occupation and state variation only. Compliance with a female gender request also falls if words related to the opposite gender are included in a job ad. In combination with the results in Table 5, these findings show that the proportion of female applicants increase with female oriented words especially those related to skills, but that these words are associated with a negative wage premium. On the other hand, the female share of applicants decreases with words related to male flexibility (that largely reflect greater travel requirements, working on weekends or night shifts), which are associated with a positive wage premium in the labour market. Thus, we find that the wording of a job ad matters for the gender mix of the applicant pool and the wage premium, especially for words related to hard skills and job flexibility even when using within occupation and state variation only. This provides evidence that job attributes within an occupation, partcularly those related to flexibility, are one of the main drivers of gender wage gaps (Goldin, 2014). Additionally, our results provide evidence for this mechanism using data on search behaviour of candidates. These findings show that women are willing to pay for flexible hours, lower travel requirements and other factors associated with higher male flexibility or perceived job inflexibility.

5 Robustness checks

We further examine the robustness of our results to different modifications:

Manual classification of occupations: We also carry out all estimations using a more disaggregate manual occupational classification (with 747 occupation categories) derived from the job title of an ad as described in Section 2.2; we find that our results are largely robust. We continue to find that explicit gender preferences are less likely in high skill jobs with a higher education requirement (column (I), Appendix Table A.4). Our results on male preferences when using the alternative occupation classification are very similar in sign and significance, with some differences in the size of the coefficients (column (IV), Appendix Table A.4). In wage regressions that use the sample of N jobs we find that the decrease in advertised wage associated with an increase in F_p continues to be far higher than the decrease associated with the same increase in M_p (column (I), Appendix Table A.5). We also find a similar pattern of effects when we examine either the total number of applications or the share of female applicants as our dependent variables of interest upon using the alternative occupation classification (columns (I) and (IV), Appendix Table A.6). We continue to find a similar responsiveness of changes in F_p (M_p) on predicted female (male) applicant shares in F, N and M jobs (Appendix Figure A.1(a)). Lastly, our results on employer word use and its consequences also continue to hold; we continue to find a decrease in advertised wage and an increase in female applicant share with 'Female (hard-skills)' and an increase in advertised wage and a decrease in female applicant share with 'Male (flexibility)' persist (Appendix Table A.9).

Firm fixed effects: We also carry out estimations with firm × state fixed effects rather than occupation × state fixed effects, and our most restrictive specification uses firm × occupation ×

state fixed effects.[41] A caveat is that we observe a few firms posting a large number of jobs across different sectors. Since we only observe a company ID and not firm names, we cannot rule out that some firms are actually HR recruiters. Nevertheless, we continue to find that our results are largely robust. We still find that higher education requirements result in a higher probability that a job ad has an explicit gender preference (columns (II) and (III), Appendix Table A.4) and that higher F_p has a larger negative effect on the advertised log wage than higher M_p among N jobs, although the p-value testing the difference in coefficients on F_p and M_p rises to 0.137 with firm × occupation × state fixed effects (columns (II) and (III), Appendix Table A.5). We also continue to find that an explicit female preference leads to a large reduction in the number of applications while there is a substantial shift in the gender mix of the applicant pool in favor of women if there is an explicit female requirement in a job ad (columns (II)-(III) and (V)-(VI), Appendix Table A.6). We also continue to find a similar responsiveness of changes in F_p (M_p) on predicted female (male) applicant shares in F, N and M jobs when using firm × state fixed effects (Appendix Figure A.1(b)); however, when using firm × occupation × state fixed effects the confidence intervals on the predicted shares become quite wide (Appendix Figure A.1(c)). Lastly, our results on employer's indication for gender preference through use of words relating to job flexibility and its positive effect on wages and negative effect on female applicant share persist (Appendix Table A.10). In fact, we now see an a significantly increase in the female applicant share when words indicating greater flexibility occur in job ads posted by the same firm for the same occupation. However, the results on words related to hard skills are now insignificant.

Candidate characteristics: We also estimate an alternative specification where the dependent variable is the share of female applicants to control for candidate characteristics. Here the regressions are at the applicant rather than job ad level, and the dependent variable takes the value one if an applicant to a job ad is female and zero if it is a male. We then estimate regressions including controls for job characteristics, occupation times state and month fixed effects along with candidate characteristics such as the applicant's highest education level, age and its square and indicator variables for applicant's years of experience mentioned in Table A.2. We find that the

[41] In Appendix Tables A.4-A.6 we report the number of observations as job ads for which the gender requirement or dependent variable varies within firms in a given state or within a firm and occupation in a given state (depending on the fixed effects used) since we are effectively only using these job ads in our estimations.

effect of employer's explicit gender preference on the probability that a female applies continues to be significantly negative when an explicit request for a male is made and positive when an explicit request for female is made (Appendix Table A.7). Similarly, we estimate the responsiveness of whether a female applicant applies to a job ad based on text predictive of a gender preference. Our previous results continue to hold. We also estimate the effect of category-gender scores on the probability of female applying for a job and find that our previous results on hard skills and job flexibility persist. These results are available on request.

Contextual gender-category scores: We also checked the robustness of our results on the effect of word use by employers to an alternative way of constructing the gender-category scores. Rather than taking the median score for each word, we took the score associated with the word in that job ad, given the context in which the word comes in the job text. Here we again find that our previous results on 'Female (hard-skills)' and 'Male (flexibility)', if anything, become stronger. On average, female related word categories lead to a decline in the advertised wage and an increase in female applicant shares while male related word categories, especially flexibility, increase posted wages and decrease the female applicant share (Appendix Table A.11).

6 Word list using candidate association

To the extent that words used by the employers while stating a female or a male request relate to typically masculine and feminine associations, our analyses show whether and what type of gendered wording used by employers matters for wages and applicant response. However, there may not be complete overlap between the gendered notions of an employer and that of a candidate. Since applicants can harbor specific gendered notions themselves, the gender mix of applicant share can vary either in the same or the opposite direction to that of an employer's notions. For instance, in our data employers are likely to specify beauty traits when requesting a woman, but female applicants may or may not respond when such specific requests regarding looks are made. This will be partially captured in the estimates provided by equation 4.1 on female applicant share. However, it may be of direct interest to researchers (when formulating experiments) and practitioners, as to which words increase the share of female applicants and by what magnitude.

Therefore, in this section we lay out the methodology to look at which words contained in job ads can matter directly for the gender mix of applicants. First, we only use jobs ads in which no explicit request was made (N jobs). Second, we estimate the part of applicant share variation which is not due to job characteristics and within a given occupation-location. To do this, we regress the female applicant share on job characteristics (education and experience requirement, month-year of posting) and occupation × state fixed effects. The regression is weighted by total applicants to a job. We predict the residual applicant share after the above estimation. The obtained residuals are then used to estimate a Ridge regression model using word unigrams (with TF-IDF scores) as features.[42] The model gives a coefficient for each word which can be interpreted as the marginal effect of the presence of that word on the female applicant share, after controlling for job characteristics and within a given occupation-location.

Candidates can respond to words or content in a job ad if they think that the words send an implicit message about the employers' preference for a particular gender or if they themselves have a particular attachment towards a stereotype. The previous analyses shows how candidates respond to words that indicate an implicit employer gender preference in job ads. We now discuss which words directly contribute to changes in the gender mix of applicants. These are likely to reflect both the candidate's response to the association that these words may have with employer's gender preference and the candidates own direct association with these words.

We first classify the words into female and male categories. To do this, we use the marginal effects of each word on the female applicant share, calculated from the RIDGE regression. Table 7 displays the list of top 20 words in each of the four categories with the marginal effect for the word in parentheses.[43] If the marginal effect is positive (negative) the word is classified in the female (male) columns for a given category. A positive (negative) marginal effect shows that a word increases (decreases) female applicant share. We discuss the nature of these words and whether these differ from those in Table 4. Within the category of hard skills (Table 7, panel

[42] Ridge regression prevents overfitting that happens using OLS in the presence of a large number of collinear features by imposing a penalty on the size of coefficients. Therefore, it reduces the sensitivity of estimates to random errors in the dependent variable. We prefer Ridge regression over Lasso as we are interested in the marginal effect of all the words instead of a sparse number of features. Secondly, Ridge regression gives a better out-of-sample fit than Lasso or random forest regressor in our case. We use 10-folds cross-validation and use the regularization parameter $\alpha = 23$, which gives the highest R^2 on the cross-validation set. For each word, we use the mean coefficient across the 10 folds.

[43] We keep words which have a marginal effect exceeding one percentage point in the table.

A, column (I)), words related to beauty, accounting, architecture skills still appear in the list of words which lead to a greater share of female applicants but we also get additional words related to legal professions, software and database management, automation and content creation that now also appear on this list. For the category of hard skills where male candidates apply the most we continue to find a dominance of engineering related, analytics and quantitative skills like python, machine learning, robotics, plc, server, desktop, configuration, network management, es, ui, seo (panel A, column (II)). In the category of soft skills, female applicant share increases when words related to communication skills related to coordination, counselling, managing customer relations increase ((panel A, column (III))). On the other hand, words related to team work and collaboration, negotiation, supervision still tend to dominate when it comes to decreasing female applicant share or increasing male applicant share (panel A, column (IV)).

The category of personality in Panel B of Table 7, does show quite a few deviations from the employers use of words in this category. Female applicant share increases when a range of personality related words appear that reflect determination, being pro active, willing to go to the last mile, ethical, creative, thinker, taking initiative and motivated are used. However, in Table 4 the reported results show that employers tend to use more appearance related words and words relating to patience, and being careful and punctual. Clearly, we did not find much effect of gendered words in personality when looking at the female applicant share response in Table 6. Similarly, we find that there is little overlap with personality related words which the employers use when requesting men, and the actual response of proportion of male candidates (panel B, column (II)). Last, we look at job flexibility related words in panel B, column (III) and (IV). In general, very few words were classified in this category. For women we can see that again the most important words are those related to being able to take skype calls and weekday working which tend to increase the female share of applicants by approximately 2.5 percentage points. While, words that reflect job characteristics involving night shift and travel decrease female applicant share by 10 to 4 percentage points. These are very large marginal effects. In fact, these results, in line with those in Section 4.2.2, show that the largest marginal effects on the gender mix of the applicant pool come from words relating to hard skills and job flexibility.

7 Discussion and Conclusion

The above results have broader implications for the literature on gender wage gaps and labor market structure. In general, the gender wage gap reflects human capital differences between men's and women's productivity as well as differential treatment of men and women in the labor market. In most developed and developing countries the proportion of gender wage gap due to differences in human capital investments has fallen over time (Kunze, 2018; Deshpande et al., 2018). Commensurately, the unexplained or the residual wage gap has increased over time. This residual wage gap can be due to taste or statistical discrimination, occupational segregation, degree to which women negotiate, compete, accumulation of human capital. Goldin (2014) argues that variation in gender wage gap with age and children profile of women does not board well with the innate differences between men and women. She finds that in the U.S., within-occupation wage differentials account for a larger proportion of the gender wage gap than between-occupation wage differentials. Thus, a focus on attributes of jobs within occupations then becomes important. She argues that attributes related to 'job flexibility' can matter to within occupation differences.[44]

Goldin (2014) uses survey data on employed individuals from the U.S. and demonstrates this by showing that wages vary non-linearly with hours of work for employed individuals. A flexible schedule is hence less likely to be rewarded in the labor market. Recent studies by Mas and Pallais (2017) and Bustelo et al. (2020) using discrete choice experiments show that women workers are more likely to pay for flexible schedules and working from home in the U.S., Colombian and Chinese labor markets, albeit across different category of workers. The willingness to pay ranges from 8% to 20% depending on the nature of flexibility offered and the context. These estimates are higher for women and further higher for women with children (Bustelo et al., 2020). We use data from real job ads posted on an online platform and applications to these job ads to test the applicant behavior as job attributes change. Our findings show that job ads which offer less flexibility post higher advertised wages and the proportion of female applicants were higher for such jobs. These results are striking since they hold even within the same occupation and location and lend support to the hypothesis that the residual gender wage gap within an occupation can reflect the earnings that women are willing to pay for job attributes, most notably for greater job flexibility. Thus, the

[44] Workplace flexibility can incorporate many features of the workplace like total hours, precise timings, flexibility to schedule one's day, possibility of work from home, extensive client meetings, travel etc.

residual wage gap may not just reflect employer discrimination but sorting of workers across jobs based on the job attributes and worker preferences. Albeit, these preferences can be shaped by societal roles with women being the primary caregivers

We examine explicit gender requests in job ads posted on an online job portal in India. We find that jobs with lower skill requirements are more likely to place a request and that ads requesting women offer lower wages. Applicant responses show high, but imperfect compliance, to these requests with women (men) applying proportionately more to jobs with an explicit female (male) request. We use detailed occupation level controls in our analyses to ensure our estimates are not capturing explicit requests being made in occupations regarded as traditionally male or female. Further, we use explicit gender preferences to derive implicit gender associations, and find that a job ad containing text predictive of an explicit female preference, i.e. *femaleness*, offers lower wages. On the applications side, an increase in *femaleness* associated with a job ad, leads to a higher female applicant share, even if an explicit female request is made. Lastly, uncovering the job ad wording, we find that ads requesting women and men differ along pre-existing gender stereotypes regarding skills, job flexibility and personality traits. These words have implications for compliance with gender requests in a job ad and affect applicant behavior for ads that do not make such requests.

Our results bear significant relevance, given the low female labor force participation rates in India, and in the absence of effective legal bans on gender requests in job ads (unlike the US or China). Our results indicate that placing restrictions on gender targeting in job ads can increase the share of job applications by women. We find that implicit gender associations matter for wage gaps and application rates, and a concerted policy to address these is important (Dhar et al., 2018). Lastly, our results using data from primarily entry level job ads are striking. We show that stereotypes matter at a stage when young people are entering the labor market, and can have important cumulative consequences for future labor market returns.

References

Abele, A. E. and B. Wojciszke (2014): "Communal and agentic content in social cognition: A dual perspective model," in *Advances in experimental social psychology*, Elsevier, vol. 50, 195–255.

Afridi, F., T. Dinkelman, and K. Mahajan (2018): "Why are fewer married women joining the work force in rural India? A decomposition analysis over two decades," *Journal of Population Economics*, 31, 783–818.

Akerlof, G. A. and R. E. Kranton (2000): "Economics and identity," *The quarterly journal of economics*, 115, 715–753.

Arceo-Gómez, E. O., R. M. Campos-Vázquez, R. Y. B. Salas, and S. López-Araiza (2020): "Gender Stereotypes in Job Advertisements: What Do They Imply for the Gender Salary Gap?" Mexico. Retrieved from http://conference. iza. org/conference_files ....

Banfi, S., S. Choi, and B. Villena-Roldan (2019): "Deconstructing Job Search Behavior," Unpublished manuscript.

Banfi, S. and B. Villena-Roldan (2019): "Do high-wage jobs attract more applicants? Directed search evidence from the online labor market," *Journal of Labor Economics*, 37, 715–746.

Belot, M., P. Kircher, and P. Muller (2017): "How Wage Announcements Affect Job Search Behaviour - A Field Experimental Investigation," Unpublished manuscript.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019): "Beliefs about gender," *American Economic Review*, 109, 739–73.

Born, M. P. and T. W. Taris (2010): "The impact of the wording of employment advertisements on students' inclination to apply for a job," *The Journal of social psychology*, 150, 485–502.

Burn, I., P. Button, L. F. M. Corella, and D. Neumark (2019): "Older Workers Need Not Apply? Ageist Language in Job Ads and Age Discrimination in Hiring," Tech. rep., National Bureau of Economic Research.

Bustelo, M., A. M. Díaz Escobar, J. Lafortune, C. Piras, L. M. Salas Bahamón, J. Tessada, et al. (2020): "What is The Price of Freedom?: Estimating Women's Willingness to Pay for Job Schedule Flexibility," Tech. rep., Inter-American Development Bank.

Campos-Vazquez, R., G. Esquivel, and R. Badillo (2020): "How has labor demand been affected by the COVID-19 pandemic? Evidence from job ads in Mexico," CEPR Press.

Chiplunkar, G., E. Kelley, and G. Lane (2020): "Which jobs are lost during a lockdown? Evidence from vacancy posting in India," Unpublished Manuscript.

Chowdhury, A. R., A. C. Areias, S. Imaizumi, S. Nomura, and F. Yamauchi (2018): *Reflections of employers' gender preferences in job ads in India: an analysis of online job portal data*, The World Bank.

Deshpande, A., D. Goel, and S. Khanna (2018): "Bad karma or discrimination? Male–female wage gaps among salaried workers in India," *World Development*, 102, 331–344.

Dhar, D., T. Jain, and S. Jayachandran (2018): "Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in India," Tech. rep., National Bureau of Economic Research.

Fletcher, E., C. Moore, and R. Pande (2018): "Women and Work in India: Descriptive Evidence and a Review of Potential Policies," Unpublished Manuscript.

Flory, J., A. Leibbrandt, and J. List (2015): "Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions," *Review of Economic Studies*, 82(1).

Forsythe, E., L. Kahn, F. Lange, and D. Wiczer (2020): "Labor demand in the time of COVID-19: Evidence from vacancy postings and UI claims," *Journal of Public Economics*, 189, 104238.

Gaucher, D., J. Friesen, and A. C. Kay (2011): "Evidence that gendered wording in job advertisements exists and sustains gender inequality." *Journal of personality and social psychology*, 101, 109.

Goldin, C. (2014): "A grand gender convergence: Its last chapter," *American Economic Review*, 104, 1091–1119.

Goldin, C. and L. F. Katz (2011): "The cost of workplace flexibility for high-powered professionals," *The Annals of the American Academy of Political and Social Science*, 638, 45–67.

HAYASHI, R. AND N. MATSUDA (2020): "COVID-19 impact on job postings: Real time assessment using Bangladesh and Sri Lanka online job portals," Asian Development Bank, ADB Briefs.

HE, H., D. NEUMARK, AND Q. WENG (2019): "Do Workers Value Flexible Jobs? A Field Experiment," Tech. rep., National Bureau of Economic Research.

HELLESETER, M. D., P. KUHN, AND K. SHEN (2020): "The Age Twist in Employers' Gender Requests Evidence from Four Job Boards," *Journal of Human Resources*, 55, 428–469.

HERSHBEIN, B. AND L. KAHN (2018): "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings," *American Economic Review*, 108, 1737–1772.

IBANEZ, M. AND G. REINER (2018): "Sorting through affirmative action: three field experiments in Colombia," *Journal of Labor Economics*, 36(2).

KLASEN, S. AND J. PIETERS (2015): "What Explains the Stagnation of Female Labor Force Participation in Urban India?" *The World Bank Economic Review*, 29, 449–478.

KUHN, P. AND K. SHEN (2013): "Gender discrimination in job ads: Evidence from china," *The Quarterly Journal of Economics*, 128, 287–336.

KUHN, P., K. SHEN, AND S. ZHANG (2020): "Gender-targeted job ads in the recruitment process: Facts from a Chinese job board," *Journal of Development Economics*, 102531.

KUNZE, A. (2018): "The gender wage gap in developed countries," *The Oxford Handbook of Women and the Economy*, 369.

LAPUSCHKIN, S., S. WÄLDCHEN, A. BINDER, G. MONTAVON, W. SAMEK, AND K.-R. MÜLLER (2019): "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, 10, 1–8.

MARINESCU, I. AND R. WOLTHOFF (2020): "Opening the black box of the matching function: The power of words," *Journal of Labor Economics*, 38, 535–568.

MAS, A. AND A. PALLAIS (2017): "Valuing alternative work arrangements," *American Economic Review*, 107(12).

Moen, E. (1997): “Competitive Search Equilibrium,” *Journal of Political Economy*, 105(2).

Ningrum, P., T. Pansombut, and A. Ueranantasun (2020): “Text mining of online job advertisements to identify direct discrimination during job hunting process: A case study in Indonesia,” *Plos One*, 15(6), e0233746.

Pennington, J., R. Socher, and C. D. Manning (2014): “GloVe: Global Vectors for Word Representation,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Qiang, J., Z. Qian, Y. Li, Y. Yuan, and X. Wu (2020): “Short text topic modeling techniques, applications, and performance: a survey,” *IEEE Transactions on Knowledge and Data Engineering*.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016): “” Why should i trust you?” Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Shurchkov, O. and C. C. Eckel (2018): *Gender differences in behavioral traits and labor market outcomes*, Oxford, UK: Oxford University Press.

Taris, T. W. and I. A. Bok (1998): “On gender specificity of person characteristics in personnel advertisements: A study among future applicants,” *The Journal of psychology*, 132, 593–610.

Yin, J. and J. Wang (2014): “A dirichlet multinomial mixture model-based approach for short text clustering,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 233–242.

Tables & Figures

Table 1: Explicit gender preferences

Dependent variable:	any gender preference			male preference		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Education requirements:						
Senior secondary	–0.0642***	–0.0272***	–0.0249***	–0.0710***	–0.0360***	–0.0376***
	(0.0104)	(0.0077)	(0.0078)	(0.0118)	(0.0080)	(0.0082)
Diploma	–0.0796***	–0.0299***	–0.0274***	–0.0564***	–0.0377***	–0.0405***
	(0.0129)	(0.0076)	(0.0077)	(0.0151)	(0.0079)	(0.0080)
Graduate degree, STEM	–0.1014***	–0.0370***	–0.0263***	–0.0475***	–0.0333***	–0.0319***
	(0.0130)	(0.0074)	(0.0075)	(0.0153)	(0.0079)	(0.0080)
Graduate degree, non-STEM	–0.0811***	–0.0325***	–0.0254***	–0.0737***	–0.0392***	–0.0410***
	(0.0127)	(0.0073)	(0.0075)	(0.0148)	(0.0081)	(0.0083)
Postgraduate degree, STEM	–0.1149***	–0.0547***	–0.0452***	–0.0823***	–0.0331***	–0.0301*
	(0.0147)	(0.0093)	(0.0131)	(0.0169)	(0.0101)	(0.0150)
Postgraduate degree, non-STEM	–0.0902***	–0.0401***	–0.0047	–0.0872***	–0.0360***	–0.0434*
	(0.0147)	(0.0107)	(0.0175)	(0.0170)	(0.0118)	(0.0194)
Experience requirements:						
$1-2$ years	0.0191***	0.0129***	0.0220***	–0.0006	–0.0018	–0.0029
	(0.0039)	(0.0025)	(0.0029)	(0.0041)	(0.0023)	(0.0028)
>2 years	–0.0112***	–0.0035	0.0122***	0.0092***	0.0043	0.0026
	(0.0025)	(0.0022)	(0.0030)	(0.0025)	(0.0023)	(0.0031)
Other job requirements:						
Age requirement present	0.0232	0.0501***	0.0680***	0.0586***	0.0383***	0.0444***
	(0.0122)	(0.0091)	(0.0106)	(0.0156)	(0.0074)	(0.0086)
Beauty requirement present	0.0296***	0.0284***	0.0277**	–0.0592***	–0.0559***	–0.0585***
	(0.0109)	(0.0106)	(0.0111)	(0.0073)	(0.0081)	(0.0084)
Working night shifts specified	–0.0050	0.0106	0.0121	0.0576***	0.0593***	0.0614***
	(0.0093)	(0.0080)	(0.0083)	(0.0095)	(0.0086)	(0.0088)
Advertised wage:						
ln(wage)			–0.2002***			0.1073***
			(0.0374)			(0.0326)
$\ln(\text{wage})^2$			0.0067***			–0.0041***
			(0.0014)			(0.0013)
Fixed Effects	month	month, occ × state	month, occ × state	month	month, occ × state	month, occ × state
N	157888	156221	136698	157888	156221	136698

Notes: The dependent variable in columns (I)-(III) takes the value 1 if a job ad shows a male or female preference and 0 otherwise. The dependent variable in columns (IV)-(VI) takes the value -1 if a job ad shows a female preference, 0 if it does not show a gender preference and 1 if it shows a male preference. The omitted category among education requirement categories includes other, illiterate, and secondary education; among experience requirement categories it is 0 to <1 year of experience. Standard errors are clustered at the state and occupation level, and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.
Source: Data from the population of all job ads on the portal, subject to the restrictions described in Section 2. Columns (II)-(III) and (V)-(VI) report the effective number of observations after incorporating occ × state fixed effects which exclude job ads for which there is no variation in the dependent variable within an occ × state cell.

Table 2: Advertised wages

Sample:	F jobs		N jobs		M jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
Femaleness	–0.181***	–0.202***	–0.390***	–0.272***	–0.319***	–0.179**
	(0.053)	(0.039)	(0.023)	(0.017)	(0.069)	(0.070)
Maleness	–0.101	–0.087	–0.127***	–0.138***	–0.112*	–0.139***
	(0.064)	(0.062)	(0.019)	(0.013)	(0.053)	(0.046)
Education requirements:						
Senior secondary	0.058*	0.045**	0.067***	0.041***	0.093***	0.008
	(0.028)	(0.020)	(0.009)	(0.007)	(0.036)	(0.024)
Diploma	0.118***	0.101***	–0.018	0.026***	0.066	0.098**
	(0.035)	(0.029)	(0.014)	(0.008)	(0.054)	(0.043)
Graduate degree, STEM	0.113	0.129	0.164***	0.177***	0.128	0.113**
	(0.068)	(0.077)	(0.018)	(0.012)	(0.079)	(0.049)
Graduate degree, non-STEM	0.095***	0.104***	0.052***	0.066***	0.131***	0.087***
	(0.027)	(0.023)	(0.012)	(0.007)	(0.039)	(0.026)
Postgraduate degree, STEM	1.168	0.158	0.445***	0.343***	0.926	1.111**
	(0.661)	(0.270)	(0.056)	(0.049)	(0.510)	(0.458)
Postgraduate degree, non-STEM	–0.047	0.089	0.247***	0.280***	–0.041	–0.095
	(0.115)	(0.076)	(0.037)	(0.034)	(0.111)	(0.057)
Experience requirements:						
$1-2$ years	0.099***	0.114***	0.064***	0.074***	0.110***	0.083***
	(0.020)	(0.015)	(0.009)	(0.007)	(0.027)	(0.022)
>2 years	0.246***	0.253***	0.318***	0.307***	0.291***	0.264***
	(0.024)	(0.021)	(0.013)	(0.011)	(0.038)	(0.032)
Fixed Effects	month	month, occ × state	month	month, occ × state	month	month, occ × state
Femaleness = Maleness, p-value	0.218	0.037	0.000	0.000	0.001	0.486
N	5729	5729	124892	124892	4801	4801

Notes: The dependent variable is the log of the mid-point of the wage range advertised in a job ad. The omitted category among education requirement categories includes other, illiterate, and secondary education; among experience requirement categories it is 0 to <1 year of experience. Standard errors are clustered at the state and occupation level, and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.

Source: Data from the population of all job ads on the portal which advertise a wage range, subject to the restrictions described in Section 2. Columns (II), (IV) and (VI) report the effective number of observations after incorporating occ × state fixed effects which exclude job ads for which there is no variation in the dependent variable within an occ × state cell.

Table 3: Applications

Dependent variable:	total applications			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female preference	–20.271***	–7.657***	–4.900***	0.206***	0.155***	0.154***
	(2.607)	(0.846)	(0.853)	(0.014)	(0.006)	(0.007)
Male preference	–2.845	–1.221	–2.272	–0.131***	–0.099***	–0.095***
	(4.539)	(4.664)	(2.921)	(0.009)	(0.005)	(0.005)
Education requirements:						
Senior secondary	–0.081	2.464***	1.743**	0.047***	0.027***	0.028***
	(0.774)	(0.716)	(0.768)	(0.004)	(0.003)	(0.003)
Diploma	23.450***	3.705*	2.081	–0.006	0.021***	0.023***
	(1.983)	(1.729)	(1.668)	(0.009)	(0.004)	(0.004)
Graduate degree, STEM	107.462***	55.408***	50.829***	0.068***	0.046***	0.047***
	(14.501)	(7.442)	(6.987)	(0.012)	(0.004)	(0.004)
Graduate degree, non-STEM	23.527***	11.164***	8.139***	0.117***	0.054***	0.055***
	(4.140)	(1.808)	(1.438)	(0.006)	(0.004)	(0.004)
Postgraduate degree, STEM	7.769	1.694	–1.518	0.171***	0.112***	0.121***
	(5.100)	(7.276)	(15.687)	(0.011)	(0.013)	(0.015)
Postgraduate degree, non-STEM	–3.538***	1.285	–10.019***	0.150***	0.078***	0.085***
	(1.292)	(2.408)	(2.504)	(0.020)	(0.011)	(0.014)
Experience requirements:						
$1 - 2$ years	–26.258***	–24.635***	–18.596***	–0.026***	–0.015***	–0.017***
	(4.253)	(3.636)	(2.475)	(0.004)	(0.002)	(0.003)
> 2 years	–40.952***	–46.829***	–37.045***	–0.065***	–0.037***	–0.036***
	(5.879)	(6.834)	(4.432)	(0.005)	(0.003)	(0.003)
Other job requirements:						
Age requirement present	–12.754***	–3.829***	–2.930**	–0.055***	–0.004	–0.002
	(2.336)	(1.053)	(1.208)	(0.009)	(0.003)	(0.003)
Beauty requirement present	–7.699***	–2.764	–2.694	–0.011	0.008	0.008
	(2.543)	(2.375)	(2.386)	(0.006)	(0.009)	(0.008)
Working night shifts specified	12.388	19.469*	15.324	–0.019**	–0.031***	–0.028***
	(8.453)	(8.775)	(8.328)	(0.008)	(0.007)	(0.008)
Advertised wage:						
ln(wage)			–21.582			–0.043*
			(52.342)			(0.021)
ln(wage)^2			1.742			0.002*
			(2.191)			(0.001)
Fixed Effects	month	month, occ × state	month, occ × state	month	month, occ × state	month, occ × state
N	157888	156221	136698	157888	156221	136698

Notes: The dependent variable in columns (I)-(III) is the number of applicants to a job ad and in columns (IV)-(VI) is the fraction of female applicants. The omitted category among education requirement categories includes other, illiterate, and secondary education; among experience requirement categories it is 0 to < 1 year of experience. Regressions in columns (IV)-(VI) are weighted by the total number of applications made to a job ad. Standard errors are clustered at the state and occupation level and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2. Columns (II)-(III) and (V)-(VI) report the effective number of observations after incorporating occ × state fixed effects which exclude job ads for which there is no variation in the dependent variable within an occ × state cell.

Table 4: Employer gendered word representations

(I)	(II)	(III)	(IV)
Panel A			
Hard skills		**Soft skills**	
Female	Male	Female	Male
autocad	hardware	fluency	fluently
facial	wpm	telugu	arabic
pedicure	rcm	fluent	supervise
manicure	regulation	malayalam	liaison
ppt	qc	talk	pitch
tally	manual	counsel	negotiation
computer	mysql	communicator	verbally
cake	scan	speak	marathi
auto	machine	gujarati	persuade
coral	sql	edit	punctuation
hashtag	audit	verbal	write
zoho	troubleshoot	bengali	french
word	receivable	hindi	motivate
ms	rf	crm	communicate
ledger	trouble	accommodate	read
expense	visual	oral	negotiate
manuscript	demat	convince	liaise
makeup	instagram	english	advise
keyword	outward	coordinate	ar
architectural	campaign	etiquette	grammar
Panel B			
Personality/Appearance		**Job Flexibility**	
Female	Male	Female	Male
personality	honest	home	petrol
punctual	energetic	skype	night
presentable	pressure		relocate
patiently	cm		shift
smile	empathy		fuel
confidence	calm		weekend
mature	impression		outstation
keen	passionate		weekday
getter	honesty		travel
height	prompt		rotational
pleasant	ethical		
polite	complexion		
flair	problem		
adaptability	methodical		
proactive	enthusiastic		
rejection	chest		
entrepreneurial	listener		
positive	scar		
careful	resourceful		
tone	creatively		

Notes: The table shows the top 20 words in each of the four categories - Hard skills/Skills, Soft skills, Personality/Appearance, Job Flexibility/Benefits - for females (Column I and III) and males (column II and IV). Words are sorted in decreasing order of importance within each gender-category combination. Abbreviations - wpm (words per minute), rcm (reliability centered maintenance), qc (quality control), rf (radio frequency), crm (customer relationship management)
Source: Data from the population of all job ads.

Table 5: Employer gender word representations and the advertised wage

Sample:	*F* Jobs		*N* Jobs		*M* Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.036***	–0.021***	–0.031***	–0.014***	–0.033***	–0.017
	(0.006)	(0.006)	(0.003)	(0.002)	(0.010)	(0.010)
Female (soft-skills)	–0.006	–0.004	0.006**	0.004*	0.009	–0.002
	(0.006)	(0.004)	(0.002)	(0.002)	(0.009)	(0.009)
Female (personality)	0.014**	0.007	0.033***	0.010***	0.033***	0.005
	(0.005)	(0.005)	(0.003)	(0.002)	(0.007)	(0.005)
Female (flexibility)	0.008	–0.001	0.001	–0.000	–0.002	0.001
	(0.007)	(0.007)	(0.002)	(0.002)	(0.011)	(0.011)
Female (others)	–0.023***	–0.018***	–0.043***	–0.021***	0.046**	0.029
	(0.006)	(0.006)	(0.004)	(0.003)	(0.019)	(0.021)
Male (hard-skills)	0.006	0.022	0.019***	0.017***	–0.014	0.025
	(0.022)	(0.020)	(0.004)	(0.003)	(0.014)	(0.014)
Male (soft skills)	0.022	0.015	0.021***	0.015***	0.022***	0.024***
	(0.013)	(0.012)	(0.003)	(0.002)	(0.008)	(0.007)
Male (personality)	0.002	0.004	0.013***	0.010***	0.004	0.006
	(0.010)	(0.008)	(0.003)	(0.003)	(0.007)	(0.006)
Male (flexibility)	0.059***	0.046***	0.034***	0.022***	0.023***	0.019**
	(0.014)	(0.007)	(0.004)	(0.002)	(0.008)	(0.007)
Male (others)	0.003	0.011	0.018***	0.010***	0.035***	0.013**
	(0.019)	(0.019)	(0.004)	(0.004)	(0.005)	(0.005)
Fixed Effects	month	month,	month	month,	month	month,
		occ × state		occ × state		occ × state
N	5729	5729	124892	124892	4800	4801

Notes: The dependent variable is the log of wage offered in a job. All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories is other (education not specified), illiterate, and secondary education. The omitted category among experience requirement categories is none to less than a year of experience. All regressions include (month,year) of job posting fixed effects. Standard errors are clustered at the (state, occupation) level and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2. The final number of job ads is 157888 and those with non-missing wage data are 138216. Each column reports the effective number of observations used for estimations after incorporating occ × state fixed effects for the respective subsamples. This excludes job ads for which there is no variation in the dependent variable within an occ × state cell.

Table 6: Employer gender word representations and share of female applications

Sample:	F Jobs		N Jobs		M Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.006	–0.004	0.009***	0.004***	0.006	–0.001
	(0.006)	(0.004)	(0.002)	(0.001)	(0.004)	(0.004)
Female (soft-skills)	–0.003	–0.004	0.003	0.000	0.011***	0.004
	(0.004)	(0.002)	(0.002)	(0.001)	(0.004)	(0.004)
Female (personality)	0.001	0.003	–0.002	0.001	0.002	–0.000
	(0.004)	(0.002)	(0.002)	(0.001)	(0.004)	(0.003)
Female (flexibility)	–0.002	–0.002	0.000	–0.001	–0.004	0.002
	(0.003)	(0.003)	(0.001)	(0.001)	(0.004)	(0.004)
Female (others)	0.006	–0.003	0.021***	0.011***	0.051***	0.021
	(0.004)	(0.003)	(0.004)	(0.001)	(0.016)	(0.011)
Male (hard-skills)	–0.060***	–0.018	0.010***	–0.000	0.012***	0.000
	(0.014)	(0.010)	(0.002)	(0.001)	(0.003)	(0.002)
Male (soft-skills)	–0.019**	–0.008	–0.005***	–0.001	–0.007	–0.001
	(0.008)	(0.005)	(0.001)	(0.001)	(0.004)	(0.003)
Male (personality)	0.005	0.003	0.003***	0.000	0.008**	0.002
	(0.007)	(0.004)	(0.001)	(0.001)	(0.003)	(0.002)
Male (flexibility)	–0.026***	–0.021***	–0.003*	–0.004***	0.006**	–0.006
	(0.006)	(0.005)	(0.002)	(0.001)	(0.002)	(0.003)
Male (others)	–0.106***	–0.067***	–0.035***	–0.011***	–0.019***	–0.006***
	(0.024)	(0.017)	(0.003)	(0.002)	(0.002)	(0.002)
Fixed Effects	month	month, occ × state	month	month, occ × state	month	month, occ × state
N	5839	5839	144117	144117	4944	4945

Notes: The dependent variable is the fraction of female applicants in a job ad. All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories is other (education not specified), illiterate, and secondary education. The omitted category among experience requirement categories is none to less than a year of experience. All regressions include (month,year) of job posting fixed effects and are weighted by the the total number of applications made to a job ad. Standard errors are clustered at the (state, occupation) level and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2. The final number of job ads is 157888. Each column reports the effective number of observations used for estimations after incorporating occ × state fixed effects for the respective subsamples. This excludes job ads for which there is no variation in the dependent variable within an occ × state cell.

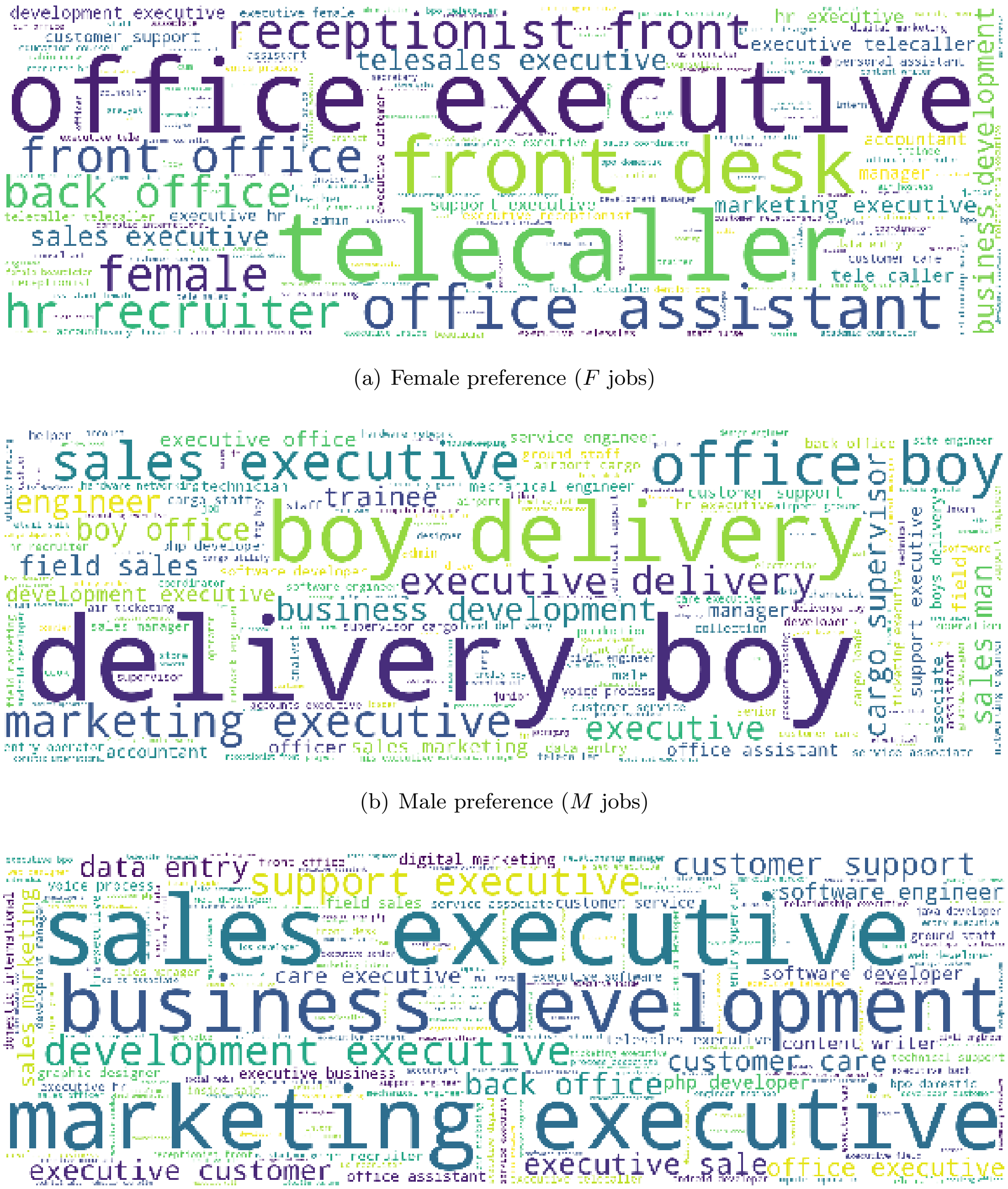
Table 7: Direct responsiveness of share of female applications to words

(I)	(II)	(III)	(IV)
Panel A			
Hard skills		**Soft skills**	
Female	Male	Female	Male
makeup (0.106)	python (-0.115)	write (0.057)	collaborate (-0.048)
legal (0.076)	desktop (-0.061)	bengali (0.055)	ar (-0.040)
facial (0.066)	robotic (-0.055)	guide (0.053)	telugu (-0.039)
architectural (0.062)	quantitative (-0.047)	counsel (0.052)	negotiate (-0.032)
rf (0.061)	install (-0.043)	coordinate (0.043)	speak (-0.030)
manuscript (0.057)	machine (-0.039)	rapport (0.037)	fluency (-0.026)
compute (0.051)	server (-0.038)	relationship (0.036)	supervise (-0.023)
court (0.048)	plc (-0.036)	english (0.035)	speech (-0.023)
cnc (0.045)	guest (-0.036)	story (0.030)	verbal (-0.021)
content (0.044)	statement (-0.034)	coordination (0.029)	read (-0.020)
proofread (0.044)	configuration (-0.033)	french (0.028)	edit (-0.017)
draft (0.040)	repair (-0.032)	crm (0.025)	marathi (-0.016)
database (0.038)	adobe (-0.032)	ordinate (0.025)	articulate (-0.015)
software (0.038)	es (-0.031)	fluent (0.025)	persuade (-0.015)
risk (0.036)	network (-0.031)	communicate (0.022)	neutral (-0.013)
cake (0.034)	knowledgeable (-0.030)	feedback (0.021)	engage (-0.013)
demonstration (0.033)	erp (-0.030)	verbally (0.020)	pitch (-0.012)
animation (0.032)	ui (-0.030)	influence (0.018)	clientele (-0.011)
automation (0.031)	collate (-0.028)	liaise (0.016)	malayalam (-0.011)
regulation (0.031)	seo (-0.027)	color (0.016)	etiquette (-0.010)
Panel B			
Personality/Appearance		**Job Flexibility**	
Female	Male	Female	Male
personality (0.053)	punctual (-0.034)	skype (0.026)	night (-0.103)
appearance (0.046)	smile (-0.032)	weekday (0.020)	travel (-0.049)
ethic (0.042)	adapt (-0.028)	outstation (0.015)	petrol (-0.041)
mile (0.042)	tone (-0.026)		fuel (-0.019)
resourceful (0.040)	dedicate (-0.024)		rotational (-0.016)
initiative (0.039)	keen (-0.024)		relocate (-0.013)
motivation (0.039)	pleasant (-0.021)		shift (-0.012)
determination (0.031)	neat (-0.021)		
proactively (0.031)	chest (-0.019)		
zeal (0.027)	entrepreneurial (-0.019)		
responsive (0.027)	adaptability (-0.019)		
proactive (0.026)	confident (-0.018)		
creative (0.026)	vigilant (-0.017)		
passionate (0.022)	enthusiasm (-0.017)		
rejection (0.021)	hardworke (-0.017)		
thinker (0.021)	height (-0.017)		
attitude (0.020)	initiate (-0.017)		
persuasive (0.019)	learner (-0.016)		
professionalism (0.018)	empathy (-0.015)		
creatively (0.016)	dedication (-0.013)		

Notes: The table shows the top 20 words in each of the four categories - Hard skills/Skills, Soft skills, Personality/Appearance, Job Flexibility - for females (Column I and III) and males (column II and IV). Words are sorted in decreasing order of importance within each gender-category combination. Parentheses show the effect on female applicant share. Abbreviations - rf (radio frequency), cnc (computerized numerical control), plc(programmable logic controller), es(engineering science), erp (enterprise resource planning), ui(user interface), seo (Search Engine Optimization), ar(augmented reality), crm (customer relationship management)

Source: Data from the population of all job ads that do not specify a gender request, after dropping duplicates.

Figure 1: Word clouds of job titles



(a) Female preference (F jobs)

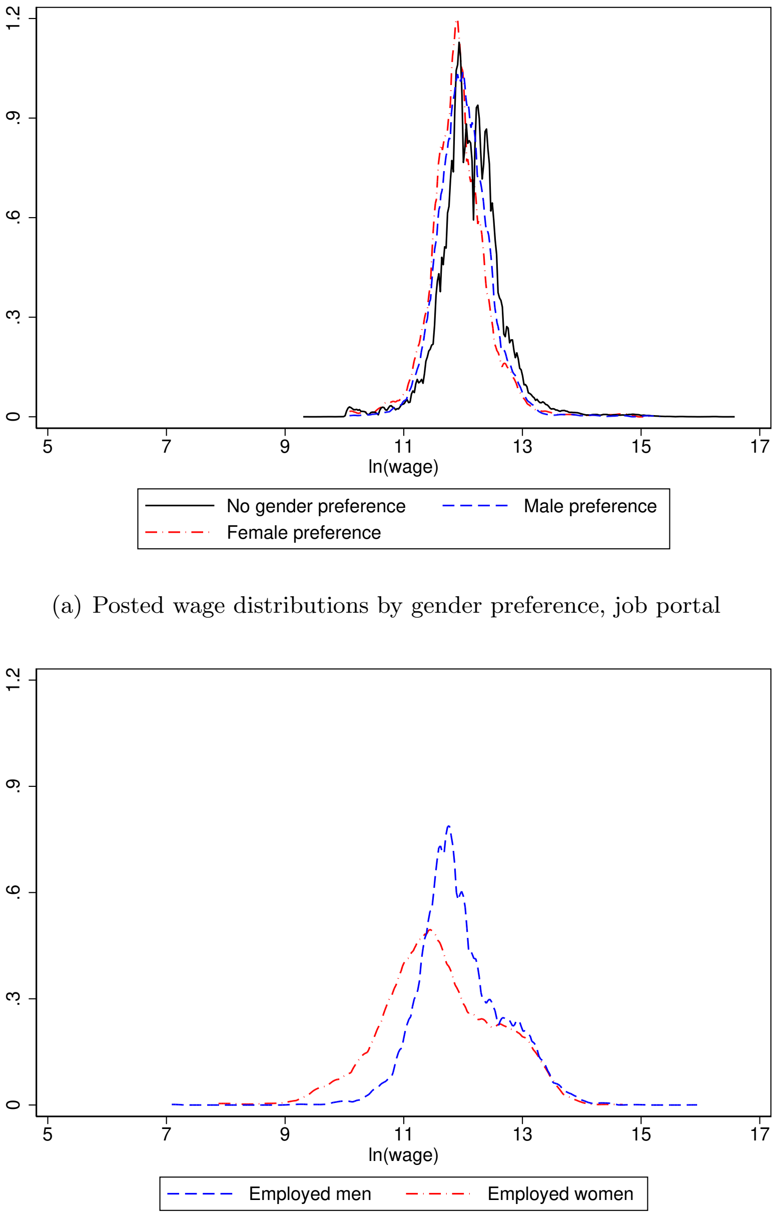
(b) Male preference (M jobs)

(c) No gender preference (N jobs)

Notes: The word clouds are constructed based on words contained in job titles of ads displaying an explicit female preference, an explicit male preference and no explicit gender preference.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2. The final number of job ads is 157888.

Figure 2: Wage distributions



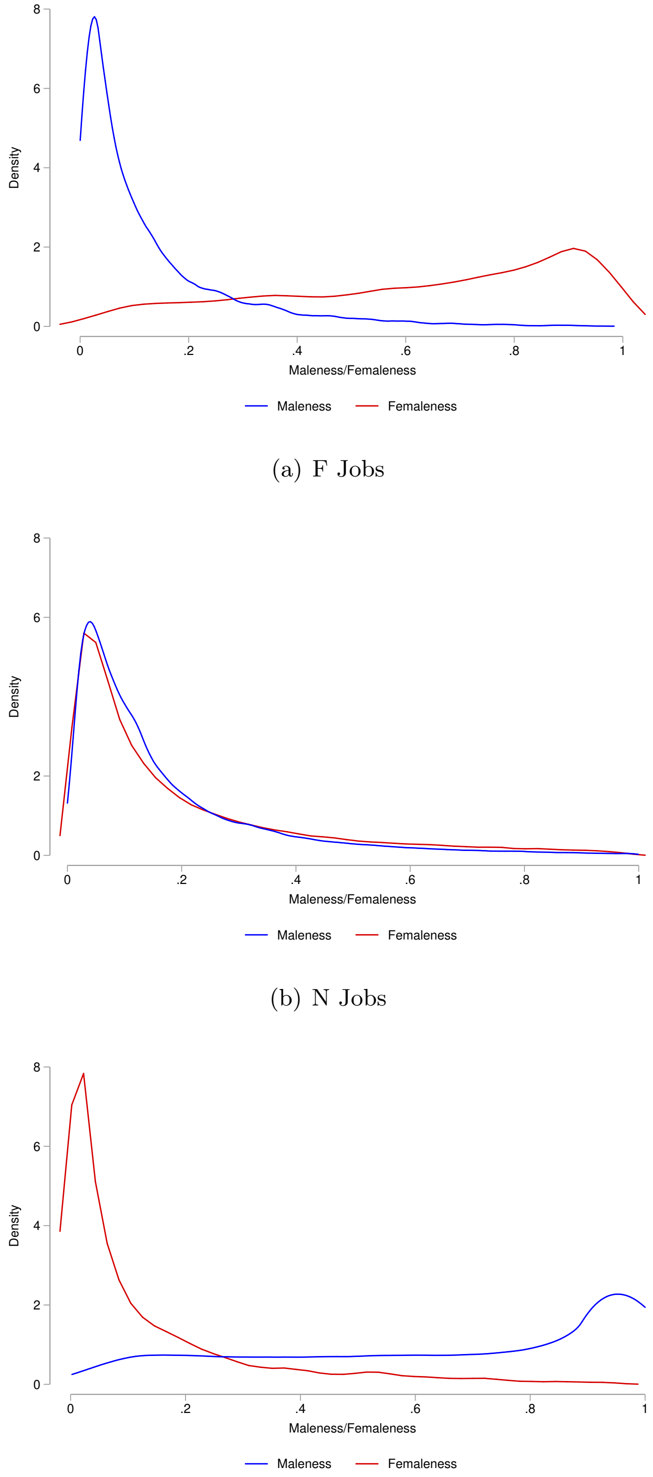
(a) Posted wage distributions by gender preference, job portal

(b) Wage distributions by gender, PLFS

Notes: Distributions are the kernel density estimates. Figure (a) uses the mid-point of the posted wage range in job ads on the job portal.

Source: Figure (a) includes data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2. Figure (b) includes all urban workers aged 18-32 in 63 majority urban districts (having at least 70% urban population) in India and reporting a wage in the Periodic Labor Force Survey for India (2017-18).

Figure 3: Kernel Density Maleness and Femaleness



(a) F Jobs

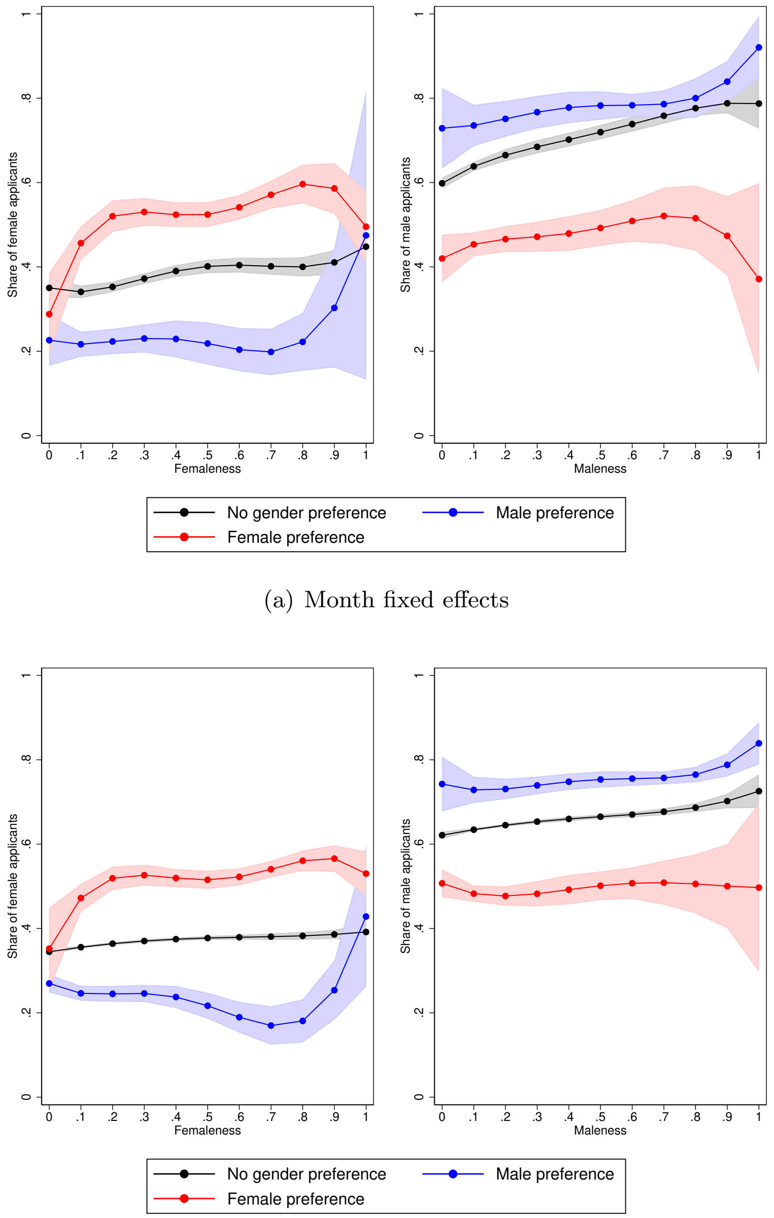
(b) N Jobs

(c) M Jobs

Notes: Distributions are the kernel density plots of estimated maleness and femaleness in F, N and M jobs.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2.

Figure 4: Predicted share of female (male) applicants



(a) Month fixed effects

(b) Month and occupation × state fixed effects

Notes: Shaded areas give the 95% confidence intervals around predicted values. The measure of implicit femaleness (maleness) is constructed using a Logistic Regression classifier as described in Section 2.4. Predictions are based on regressing the share of female (male) applicants on explicit gender preferences, quartics in implicit femaleness (maleness), their interactions and the set of controls specified in equation (3.4), as well as time (month and year) fixed effects. Predictions used to construct the Figures in (b) also include occupation × state fixed effects. These regressions are weighted by the total number of female and male applications, with standard errors clustered by occupation and state.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2. The final number of job ads is 157888.

Figure 5: Heat map visualization of words in distinctive job ads

i. **SOFTWARE TRAINEE**: faculty follow subject basic computer complete knowledge ms office friendly internet advance english grammar personality development class comunication skill basic account taly gst

ii. **BUSINESS DEVELOPMENT MANAGER**: language bengali fluently speak english read write fluently speak hindi fluently speak groom look air hostess manager hr student counsel employee handle eod report share total office management bond applicable employee qualification preferable minimum graduate mba market master psychology applicable look smart computer knowledge power point mail communication excel presentation skill age height weight proportionate height

iii. **SALES MARKET EXECUTIVE:** smart intelligent look sale experience aviation experience sell tour operator hotel corporate client complete cabin crew train add advantage communication skill english malayalam speak regional language it add advantage smart look able handle high client business development manage exist client day day flight manage customer relationship support head sale addition entitle incentive achieve set target

(a) female preference

i. **SOFTWARE TRAINEE**: qualification tech sc bca mca sc fresh pass it computer science background verbal write communication skill basic knowledge it technologie quick learner able work rotational shift

ii. **BUSINESS DEVELOPMENT MANAGER**: look energetic post bdm experience sale communication skill wheeler jd set deliver sale presentation demo daily identify potential client implement innovative business strategy

iii. **SALES MARKET EXECUTIVE**: fix incentive call field work education degree diploma experience fresh experience designation market manager shift general shift wheeler mandatory language tamil

(b) male preference

Notes: Panel (a) shows correctly classified job ads with an explicit female preference; panel (b) shows correctly classified job ads with an explicit male preference. Words highlighted in red reflect female associations, and those in blue correspond to male associations as returned by LIME. The color intensity reflects the strength of the attached gender association, with darker shades showing a higher strength.

A Additional Tables & Figures

Table A.1: Descriptive statistics, job ads

	Prefer female	No pref.	Prefer male	Total
Education requirements:				
Other (education not specified)	0.006	0.004	0.004	0.004
None (illiterate)	0.018	0.014	0.042	0.015
Secondary education	0.113	0.099	0.322	0.108
Senior secondary education	0.318	0.263	0.259	0.265
Diploma	0.075	0.090	0.077	0.089
Graduate degree, STEM	0.034	0.089	0.054	0.086
Graduate degree, non-STEM	0.425	0.424	0.237	0.417
Postgraduate degree, STEM	0.003	0.007	0.000	0.006
Postgraduate degree, non-STEM	0.006	0.007	0.002	0.006
Experience requirements:				
$0-1$ years	0.688	0.663	0.687	0.665
$1-2$ years	0.215	0.177	0.202	0.179
> 2 years	0.096	0.160	0.111	0.155
Other job requirements:				
Age requirement present	0.073	0.083	0.187	0.086
Minimum age requirement present	0.059	0.075	0.173	0.078
Maximum age requirement present	0.066	0.078	0.168	0.080
Beauty requirement present	0.118	0.057	0.060	0.059
Working night shifts specified	0.008	0.021	0.039	0.021
Advertised wage:				
Wage not specified	0.021	0.133	0.032	0.125
Annual wage, if wage specified in job ad	178261	224649	186640	221008
N (jobs with advertised wage)	6415	126389	5412	138216
Applications:				
Share of female applicants	0.521	0.319	0.129	0.321
Number of applications	17.416	42.274	31.296	40.854
N (all jobs)	6551	145748	5589	157888

Notes: Each cell gives the average value of a variable in the respective sub-sample of job ads. Wages are annual wages in Rupees. Wages and experience are the mid-point of the range specified in the job ad.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2.

Table A.2: Descriptive statistics, job applicants

	Female	Male	Total
Education:			
Other (education not specified)	0.002	0.002	0.002
None (illiterate)	0.000	0.000	0.000
Secondary education	0.004	0.016	0.012
Senior secondary education	0.030	0.068	0.054
Diploma	0.030	0.087	0.066
Graduate degree, STEM	0.535	0.545	0.541
Graduate degree, non-STEM	0.155	0.135	0.142
Postgraduate degree, STEM	0.122	0.067	0.087
Postgraduate degree, non-STEM	0.122	0.080	0.095
Experience:			
$0-1$ years	0.799	0.736	0.758
$1-2$ years	0.069	0.079	0.075
> 2 years	0.132	0.185	0.166
Age:			
Age at registration	23.460	23.863	23.720
Applied wage:			
Mean annual wage	307208	294428	298931
Number of applications:			
Number of applications	6.148	6.048	6.083
N (Applicants)	374804	685927	1060731

Notes: Each cell gives the average value of the variable in the respective sub-sample of job applications. Experience is given in years and is divided into four categories to correspond to the job advertisements sample.
Source: The applicant sample includes those who applied to at least one job in our job advertisement sample, and disclosed their gender.

Table A.3: Descriptive statistics, PLFS Urban workers

	Female	Male	Total
Panel A: Age 16-60			
Education:			
None (illiterate)	0.159	0.075	0.094
Less than Secondary education	0.254	0.335	0.317
Secondary education	0.074	0.147	0.131
Senior secondary	0.075	0.117	0.108
Diploma	0.020	0.026	0.025
Graduate degree	0.263	0.216	0.226
Postgraduate degree	0.155	0.083	0.098
Age:			
Age	35.417	36.030	35.897
Salary:			
Annual Wage	167983	207824	199217
Observations	2954	10853	13807
LFPR	0.226	0.821	0.529
Panel B: Age 18-32			
Education:			
None (illiterate)	0.089	0.052	0.060
Less than Secondary education	0.170	0.321	0.288
Secondary education	0.075	0.140	0.125
Senior secondary	0.079	0.129	0.118
Diploma	0.028	0.035	0.033
Graduate degree	0.361	0.244	0.270
Postgraduate degree	0.196	0.079	0.105
Age:			
Age	26.417	26.436	26.432
Salary:			
Annual Wage	167490	178405	176001
Observations	1166	4382	5548
LFPR	0.242	0.774	0.518

Notes: The sample includes all urban workers in 63 majority urban districts (having at least 70% urban population) in India. Panel A includes all workers aged 16-60 while Panel B includes all workers aged 18-32. Each cell gives the average value of the variable in the respective sub-sample of workers. Age is given in years. The Labour force participation rate (LFPR) refers to proportion of individuals working majority of the year. This proportion is calculated for all individuals in the respective gender-age group.
Source: Periodic Labour Force Survey (PLFS) conduted in 2017-18.

Table A.4: Explicit gender preferences, robustness checks

Dependent variable:	any gender preference			male preference		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Education requirements:						
Senior secondary	–0.012***	–0.060***	–0.021***	–0.020***	–0.068***	–0.024***
	(0.004)	(0.006)	(0.006)	(0.004)	(0.006)	(0.007)
Diploma	–0.012**	–0.072***	–0.018**	–0.022***	–0.065***	–0.027***
	(0.005)	(0.009)	(0.008)	(0.006)	(0.007)	(0.008)
Graduate degree, STEM	–0.018***	–0.089***	–0.025***	–0.016***	–0.064***	–0.023***
	(0.005)	(0.009)	(0.007)	(0.005)	(0.009)	(0.008)
Graduate degree, non-STEM	–0.013***	–0.075***	–0.024***	–0.021***	–0.082***	–0.030***
	(0.004)	(0.009)	(0.007)	(0.005)	(0.007)	(0.008)
Postgrad degree, STEM	–0.030***	–0.080***	–0.012	–0.029***	–0.075***	–0.037***
	(0.009)	(0.010)	(0.010)	(0.009)	(0.009)	(0.012)
Postgrad degree, non-STEM	–0.026**	–0.067***	–0.014	–0.013	–0.083***	–0.042***
	(0.010)	(0.008)	(0.010)	(0.011)	(0.009)	(0.010)
Experience requirements:						
$1-2$ years	0.012***	0.012*	0.005	–0.002	–0.008**	–0.008
	(0.002)	(0.006)	(0.005)	(0.002)	(0.003)	(0.004)
>2 years	–0.004*	–0.007	–0.004	0.006***	–0.001	–0.000
	(0.002)	(0.004)	(0.004)	(0.002)	(0.004)	(0.004)
Other job requirements:						
Age requirement present	0.048***	0.039*	0.058***	0.031***	0.041***	0.064***
	(0.006)	(0.019)	(0.012)	(0.006)	(0.007)	(0.009)
Beauty requirement present	0.029***	0.008	–0.004	–0.049***	–0.039***	–0.042***
	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
Working night shifts specified	0.011	0.015	0.024**	0.056***	0.062***	0.041***
	(0.006)	(0.010)	(0.010)	(0.007)	(0.010)	(0.010)
Fixed Effects	month, alt occ × state	month, firm × state	month, firm × occ × state	month, alt occ × state	month, firm × state	month, firm × occ × state
N	152568	102203	62089	152568	102203	62089

Notes: The dependent variable in columns (I)-(III) takes the value 1 if a job ad shows a male or female preference and 0 otherwise. The dependent variable in columns (IV)-(VI) takes the value −1 if a job ad shows a female preference, 0 if it does not show any gender preference and 1 if it shows a male preference. The omitted category among education requirement categories includes other, illiterate, and secondary education; among experience requirement categories it is 0 to <1 year of experience. Standard errors are clustered at the state and occupation level (columns (I) and (IV)), state and firm level (columns (II) and (V)), or state, occupation and firm level (columns (III) and (VI)), and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.
Source: Data from the population of all job ads on the portal, subject to the restrictions described in Section 2. Each column reports the effective number of observations after incorporating fixed effects which exclude job ads for which there is no variation in the dependent variable within an alt occ × state, firm × state or firm × occ × state cell, depending on the fixed effects used.

Table A.5: Advertised wages, robustness checks

	(I)	(II)	(III)
Femaleness	–0.233***	–0.286***	–0.129***
	(0.014)	(0.019)	(0.018)
Maleness	–0.110***	–0.080***	–0.095***
	(0.013)	(0.017)	(0.019)
Education requirements:			
Senior secondary	0.032***	–0.019	–0.029***
	(0.006)	(0.013)	(0.010)
Diploma	0.017*	0.036**	0.006
	(0.008)	(0.016)	(0.019)
Graduate degree, STEM	0.148***	0.141***	0.101***
	(0.011)	(0.028)	(0.019)
Graduate degree, non-STEM	0.056***	0.018	–0.007
	(0.006)	(0.011)	(0.011)
Postgrad degree, STEM	0.359***	0.176***	0.135
	(0.046)	(0.068)	(0.070)
Postgrad degree, non-STEM	0.222***	0.220***	0.250***
	(0.036)	(0.054)	(0.074)
Experience requirements:			
$1-2$ years	0.064***	0.046**	0.014
	(0.005)	(0.019)	(0.011)
>2 years	0.287***	0.262***	0.181***
	(0.010)	(0.026)	(0.013)
Fixed Effects	month, alt occ × state	month, firm × state	month, firm × occ × state
Femaleness = Maleness, p-value	0.000	0.000	0.137
N	122163	74913	42141

Notes: The dependent variable is the log of the mid-point of the wage advertised in a job ad. The omitted category among education requirement categories includes other, illiterate, and secondary education; among experience requirement categories it is 0 to <1 year of experience. Standard errors are clustered at the state and occupation level (column (I)), state and firm level (column (II)), or state, occupation and firm level (column (III)), and reported in parentheses; * p-value <0.05, ** p-value <0.025, *** p-value <0.01.
Source: Data from the population of all job ads on the portal without an explicit gender preference and which advertise a wage, subject to the restrictions described in Section 2. Each column reports the effective number of observations after incorporating fixed effects which exclude job ads for which there is no variation in the dependent variable within an alt occ × state, firm × state or firm × occ × state cell, depending on the fixed effects used.

Table A.6: Applications, robustness checks

Dependent variable:	total applications			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female preference	–5.858***	–8.368***	–4.109***	0.150***	0.195***	0.139***
	(0.704)	(0.924)	(0.919)	(0.006)	(0.010)	(0.010)
Male preference	1.131	–7.507***	1.512	–0.087***	–0.119***	–0.090***
	(3.690)	(2.704)	(3.863)	(0.005)	(0.009)	(0.009)
Education requirements:						
Senior secondary	2.084***	–0.242	1.622**	0.025***	0.023***	0.016***
	(0.729)	(0.889)	(0.678)	(0.002)	(0.004)	(0.004)
Diploma	1.739	12.477***	4.261***	0.020***	–0.004	0.028***
	(1.560)	(1.608)	(1.604)	(0.003)	(0.010)	(0.006)
Graduate degree, STEM	42.580***	35.091***	14.696***	0.040***	0.041***	0.053***
	(5.307)	(3.779)	(3.054)	(0.003)	(0.013)	(0.006)
Graduate degree, non-STEM	7.599***	2.744*	1.506	0.048***	0.081***	0.056***
	(1.364)	(1.238)	(0.881)	(0.003)	(0.009)	(0.005)
Postgrad degree, STEM	–3.121	1.742	–9.956	0.106***	0.121***	0.114***
	(7.700)	(7.194)	(16.155)	(0.018)	(0.022)	(0.027)
Postgrad degree, non-STEM	–3.937	–3.732	–2.616	0.081***	0.109***	0.069***
	(2.396)	(7.319)	(4.241)	(0.017)	(0.015)	(0.014)
Experience requirements:						
$1-2$ years	–23.435***	–10.603***	–11.112***	–0.013***	–0.015***	–0.008***
	(3.288)	(1.649)	(1.359)	(0.002)	(0.004)	(0.003)
>2 years	–42.761***	–19.227***	–20.443***	–0.033***	–0.043***	–0.030***
	(5.210)	(2.769)	(1.437)	(0.002)	(0.006)	(0.003)
Other job requirements:						
Age requirement present	–5.315***	0.498	–1.901	–0.005*	–0.013	–0.003
	(1.282)	(1.720)	(1.056)	(0.002)	(0.011)	(0.005)
Beauty requirement present	–3.016	–4.231***	–0.597	0.002	–0.004	–0.001
	(2.070)	(1.184)	(0.707)	(0.003)	(0.004)	(0.004)
Working night shifts specified	17.998*	–3.581	–6.806**	–0.017***	–0.028***	–0.020***
	(8.184)	(4.148)	(2.706)	(0.007)	(0.007)	(0.005)
Fixed Effects	month, alt occ × state	month, firm × state	month, firm × occ × state	month, alt occ × state	month, firm × state	month, firm × occ × state
N	152568	102203	62089	152568	102203	62089

Notes: The dependent variable in columns (I)-(III) is the number of applicants to a job ad and in columns (IV)-(VI) is the share of female applicants. The omitted category among education requirement categories includes other, illiterate, and secondary education; among experience requirement categories it is 0 to < 1 year of experience. Regressions in columns (IV)-(VI) are weighted by the total number of applications made to a job ad. Standard errors are clustered at the state and occupation level (columns (I) and (IV)), state and firm level (columns (II) and (V)), or state, occupation and firm level (columns (III) and (VI)), and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2. Each column reports the effective number of observations after incorporating fixed effects which exclude job ads for which there is no variation in the dependent variable within an alt occ × state, firm × state or firm × occ × state cell, depending on the fixed effects used.

Table A.7: Whether an applicant is a female: Role of explicit gender preference

	(I)	(II)	(III)
Female preference	0.153***	0.152***	0.151***
	(0.006)	(0.006)	(0.006)
Male preference	–0.093***	–0.093***	–0.088***
	(0.005)	(0.005)	(0.004)
Education requirements:			
Senior secondary	0.011***	0.011***	0.012***
	(0.002)	(0.002)	(0.002)
Diploma	0.002	0.002	0.004
	(0.003)	(0.003)	(0.003)
Graduate degree, STEM	0.017***	0.016***	0.017***
	(0.003)	(0.003)	(0.004)
Graduate degree, non-STEM	0.017***	0.018***	0.018***
	(0.003)	(0.003)	(0.003)
Postgraduate degree, STEM	0.048***	0.047***	0.058***
	(0.009)	(0.009)	(0.011)
Postgraduate degree, non-STEM	0.035***	0.035***	0.046***
	(0.011)	(0.011)	(0.014)
Experience requirements:			
$1-2$ years	–0.011***	–0.011***	–0.013***
	(0.002)	(0.002)	(0.002)
>2 years	–0.023***	–0.023***	–0.023***
	(0.002)	(0.002)	(0.003)
Other job requirements:			
Age requirement present		–0.004	–0.001
		(0.003)	(0.003)
Beauty requirement present		0.009	0.008
		(0.008)	(0.008)
Working night shifts specified		–0.031***	–0.028***
		(0.007)	(0.008)
Advertised wage:			
ln(wage)			–0.040*
			(0.021)
ln(wage)^2			0.002
			(0.001)
Candidate Controls	✓	✓	✓
Fixed Effects	occ × state	occ × state	occ × state
N	6402149	6402149	5392307

Notes: The dependent variable in columns (I)-(III) is a dummy variable that takes a value one if a female applied for the job and zero otherwise. The omitted category among education requirement categories is other (education not specified), illiterate, and secondary education. The omitted category among experience requirement categories is none to less than a year of experience. The regressions also control for education level and age and age squared of the applicant. All regressions include (month,year) of job posting fixed effects. Standard errors are clustered at the (state, occupation) level and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2. The final number of job ads is 157888. Each column reports the effective number of observations used for estimations after incorporating the respective fixed effects, which vary across columns. This excludes job ads for which there is no variation in the dependent variable within the fixed effect cell.

Table A.8: Descriptive statistics: Stereotypes

	F Jobs	N Jobs	M Jobs	All jobs
Female (hard-skills)	0.170	0.114	0.068	0.114
Male (hard-skills)	0.037	0.163	0.127	0.157
Female (soft-skills)	0.217	0.109	0.091	0.112
Male (soft-skills)	0.012	0.033	0.020	0.032
Female (personality)	0.093	0.055	0.046	0.056
Male (personality)	0.023	0.035	0.031	0.035
Female (flexibility)	0.005	0.003	0.003	0.003
Male (flexibility)	0.103	0.093	0.161	0.096
Female (others)	2.595	0.765	0.155	0.816
Male (others)	0.096	0.675	3.490	0.750
N	6791	158946	6009	171746

Notes: Each cell gives the average (non-standardized) value of a variable in the respective sub-sample of job ads. The gender association scores for each word are obtained by applying LIME technique to the multinomial logistic regression model based on explicit preferences of employers. The score for each gender × stereotype category is then obtained for each job ad.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2.

Table A.9: Employer gender word representations, advertised wage and applicant behavior, robustness (manual occupation classification)

Dependent variable:	log of advertised wage			share of female applications		
Sample:	*F* Jobs	*N* Jobs	*M* Jobs	*F* Jobs	*N* Jobs	*M* Jobs
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.020***	–0.008***	–0.002	0.005	0.002***	0.002
	(0.006)	(0.002)	(0.009)	(0.004)	(0.001)	(0.003)
Female (soft-skills)	–0.001	0.003	–0.005	–0.004	0.000	–0.002
	(0.005)	(0.002)	(0.009)	(0.002)	(0.001)	(0.003)
Female (personality)	0.008	0.008***	0.005	–0.000	0.001	0.002
	(0.005)	(0.002)	(0.007)	(0.002)	(0.001)	(0.003)
Female (flexibility)	0.004	–0.001	–0.004	–0.005	–0.000	0.004
	(0.007)	(0.002)	(0.009)	(0.003)	(0.000)	(0.003)
Female (others)	–0.014**	–0.012***	0.019	–0.005	0.006***	0.005
	(0.006)	(0.003)	(0.024)	(0.003)	(0.001)	(0.009)
Male (skills)	0.028	0.016***	0.010	–0.006	0.000	0.004
	(0.025)	(0.002)	(0.012)	(0.013)	(0.001)	(0.004)
Male (soft-skills)	0.012	0.011***	0.015	–0.009	–0.001	0.006
	(0.010)	(0.002)	(0.008)	(0.005)	(0.001)	(0.003)
Male (personality)	0.003	0.009***	0.004	–0.001	0.000	0.003
	(0.008)	(0.002)	(0.006)	(0.004)	(0.000)	(0.003)
Male (flexibility)	0.044***	0.020***	0.019**	–0.011**	–0.003***	0.000
	(0.008)	(0.002)	(0.008)	(0.004)	(0.001)	(0.002)
Male (others)	0.033	0.015***	0.015***	–0.067***	–0.007***	–0.004**
	(0.019)	(0.003)	(0.004)	(0.021)	(0.001)	(0.002)
Fixed Effects	alt occ × state	alt occ × state	alt occ × state	alt occ × state	alt occ × state	alt occ × state
N	5374	122163	4438	5484	140763	4582

Notes: All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories is other (education not specified), illiterate, and secondary education. The omitted category among experience requirement categories is none to less than a year of experience. All regressions include (month,year) of job posting fixed effects. Standard errors are clustered at the (state, occupation) level and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2. The final number of job ads is 157888 and those with non-missing wage data are 138216. Each column reports the effective number of observations used for estimations after incorporating occ × state fixed effects for the respective subsamples. This excludes job ads for which there is no variation in the dependent variable within an occ × state cell.

Table A.10: Employer gender word representations, advertised wage and applicant behavior, robustness (firm × occupation × state)

Dependent variable:	log of advertised wage			share of female applications		
Sample:	*F* Jobs	*N* Jobs	*M* Jobs	*F* Jobs	*N* Jobs	*M* Jobs
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.023	–0.006*	–0.000	–0.006	0.002	0.004
	(0.012)	(0.003)	(0.027)	(0.009)	(0.001)	(0.007)
Female (soft-skills)	–0.001	0.003	0.002	–0.002	0.000	0.012
	(0.011)	(0.002)	(0.025)	(0.007)	(0.001)	(0.007)
Female (personality)	–0.007	–0.000	–0.010	0.005	0.000	–0.005
	(0.008)	(0.002)	(0.013)	(0.006)	(0.001)	(0.005)
Female (flexibility)	0.003	0.002	0.000	0.018***	0.002*	0.002
	(0.013)	(0.003)	(0.009)	(0.005)	(0.001)	(0.002)
Female (others)	–0.015	–0.017***	–0.035	0.012	0.007***	–0.007
	(0.011)	(0.003)	(0.040)	(0.008)	(0.002)	(0.029)
Male (skills)	–0.053	0.005	0.016	0.025	–0.000	0.003
	(0.049)	(0.004)	(0.017)	(0.020)	(0.001)	(0.004)
Male (soft skills)	0.023	0.007**	0.029***	–0.027	–0.001	–0.002
	(0.024)	(0.003)	(0.011)	(0.023)	(0.001)	(0.005)
Male (personality)	0.002	–0.004	0.004	–0.009	–0.001	–0.004
	(0.013)	(0.003)	(0.012)	(0.011)	(0.001)	(0.007)
Male (flexibility)	0.026	0.006***	0.029*	–0.037***	–0.003***	–0.002
	(0.024)	(0.002)	(0.014)	(0.013)	(0.001)	(0.004)
Male (others)	0.050	–0.002	0.027***	–0.214***	–0.007***	–0.003
	(0.078)	(0.003)	(0.005)	(0.064)	(0.001)	(0.002)
Fixed Effects	firm × occ	firm × occ	firm × occ	firm × occ	firm × occ	firm × occ
	× state	× state	× state	× state	× state	× state
N	1328	42141	1591	1342	57427	1659

Notes: All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories is other (education not specified), illiterate, and secondary education. The omitted category among experience requirement categories is none to less than a year of experience. All regressions include (month,year) of job posting fixed effects. Standard errors are clustered at the (state, occupation) level and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.

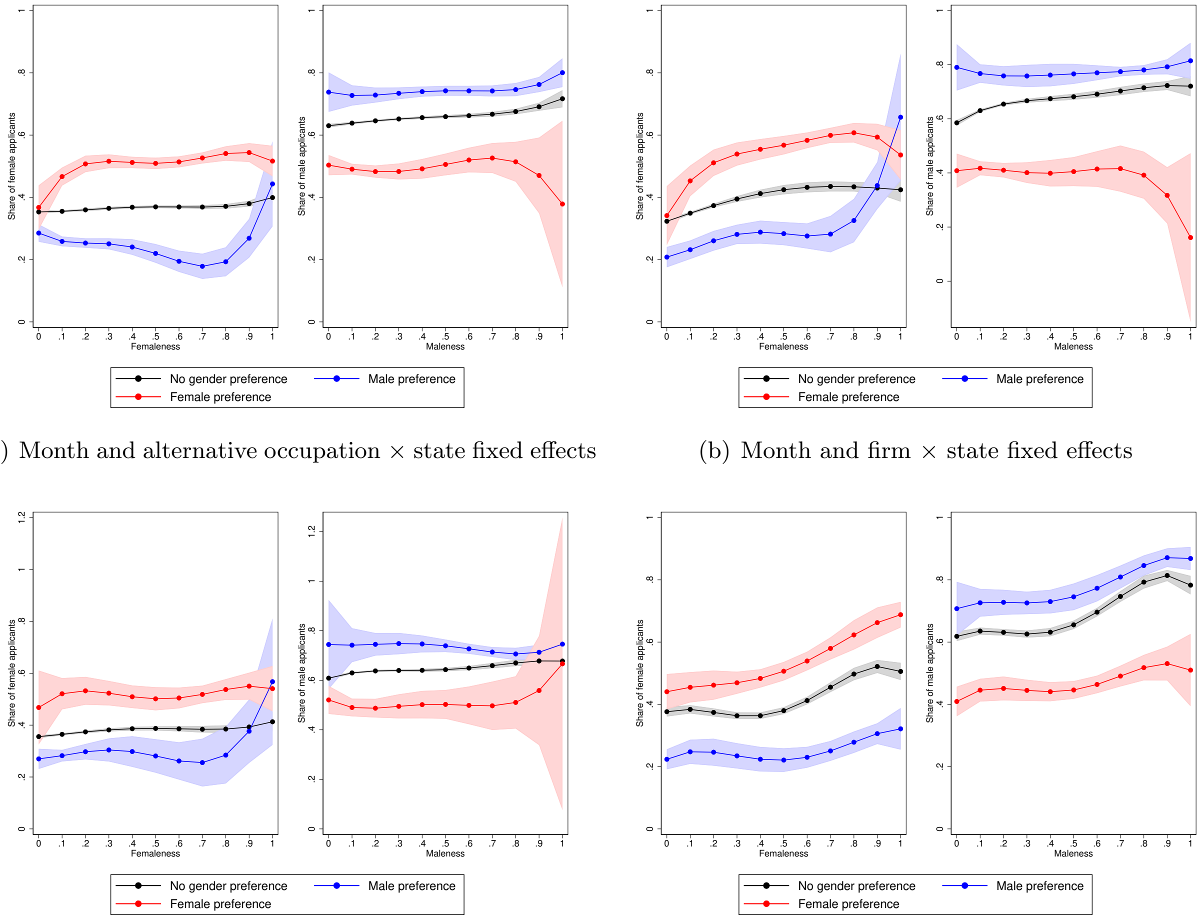
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2. The final number of job ads is 157888 and those with non-missing wage data are 138216. Each column reports the effective number of observations used for estimations after incorporating occ × state fixed effects for the respective subsamples. This excludes job ads for which there is no variation in the dependent variable within an occ × state cell.

Table A.11: Employer gender word representations, advertised wage and applicant behavior, robustness (contextual scores)

Dependent variable:	log of advertised wage			share of female applications		
Sample:	*F* Jobs	*N* Jobs	*M* Jobs	*F* Jobs	*N* Jobs	*M* Jobs
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.025***	–0.014***	–0.021***	0.002	0.004***	–0.001
	(0.005)	(0.002)	(0.008)	(0.003)	(0.001)	(0.003)
Female (soft-skills)	–0.009*	–0.001	–0.004	–0.003	0.002**	0.002
	(0.004)	(0.002)	(0.007)	(0.002)	(0.001)	(0.003)
Female (personality)	0.005	0.006***	0.000	0.001	0.001	0.002
	(0.005)	(0.002)	(0.005)	(0.002)	(0.001)	(0.003)
Female (flexibility)	0.003	0.003	0.004	–0.000	0.001	0.001
	(0.008)	(0.002)	(0.006)	(0.004)	(0.001)	(0.001)
Female (others)	–0.019***	–0.028***	0.007	–0.002	0.007***	0.018***
	(0.006)	(0.002)	(0.017)	(0.002)	(0.001)	(0.007)
Male (hard-skills)	–0.014	0.006**	0.013	–0.013**	–0.001	0.003
	(0.012)	(0.002)	(0.009)	(0.005)	(0.001)	(0.001)
Male (soft-skills)	–0.003	0.012***	0.015	0.001	0.001	0.001
	(0.005)	(0.002)	(0.011)	(0.003)	(0.001)	(0.004)
Male (personality)	0.003	0.006***	0.001	0.000	–0.001	–0.005**
	(0.006)	(0.002)	(0.006)	(0.003)	(0.001)	(0.002)
Male (flexibility)	0.027***	0.019***	0.012**	–0.014***	–0.006***	–0.007**
	(0.007)	(0.002)	(0.005)	(0.004)	(0.001)	(0.003)
Male (others)	–0.000	–0.004	–0.003	–0.027**	–0.011***	–0.005***
	(0.015)	(0.003)	(0.004)	(0.011)	(0.002)	(0.002)
Fixed Effects	month, occ × state	month, occ × state	month, occ × state	month, occ × state	month, occ × state	month, occ × state
N	5729	124892	4801	5839	144117	4945

Notes: The dependent variable is the log of wage offered in a job. All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories is other (education not specified), illiterate, and secondary education. The omitted category among experience requirement categories is none to less than a year of experience. All regressions include (month,year) of job posting fixed effects. Standard errors are clustered at the (state, occupation) level and reported in parentheses; * p-value < 0.05, ** p-value < 0.025, *** p-value < 0.01.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2. The final number of job ads is 157888 and those with non-missing wage data are 138216. Each column reports the effective number of observations used for estimations after incorporating occ × state fixed effects for the respective subsamples. This excludes job ads for which there is no variation in the dependent variable within an occ × state cell.

Figure A.1: Predicted share of female (male) applicants, robustness checks



(a) Month and alternative occupation × state fixed effects

(b) Month and firm × state fixed effects

(c) Month and firm × occupation × state fixed effects

(d) Month and state fixed effects using NB classifier

Notes: Shaded areas give the 95% confidence intervals around predicted values. The measure of implicit femaleness (maleness) in Figures (a)-(c) is constructed using a Logistic Regression classifier and in Figure (d) is constructed using a Bernoulli Naive Bayes classifier, as described in Section 2.4. Predictions are based on regressing the share of female (male) applicants on explicit gender preferences, quartics in implicit femaleness (maleness), their interactions and the set of controls specified in equation (3.4), as well as time (month and year) fixed effects. Predictions used to construct the Figures in (a) also include alternative occupation × state fixed effects, in (b) include firm × state fixed effects, in (c) include fim × occupation × state fixed effects and in (d) include state fixed effects. These regressions are weighted by the total number of female and male applications.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2. The final number of job ads is 157888.

B Technical Appendix

B.1 GSDMM, Preprocessing and Hyperparameter Choice

We use the following pre-processing steps on the text contained in job titles: (a) convert letters to lowercase; (b) remove non-latin characters, multiple occurrences of the same word in a job title, stop words, and words unrelated to job positions such as proper nouns; (c) remove words whose length is smaller than 2 or larger than 30 characters; (d) tokenize and lemmatize the job titles and (e) remove duplicate job titles as well as words that occur only once in the entire corpus. In our data, the resulting number of documents $D = 28,957$ and the number of unique words $V = 3,127$.

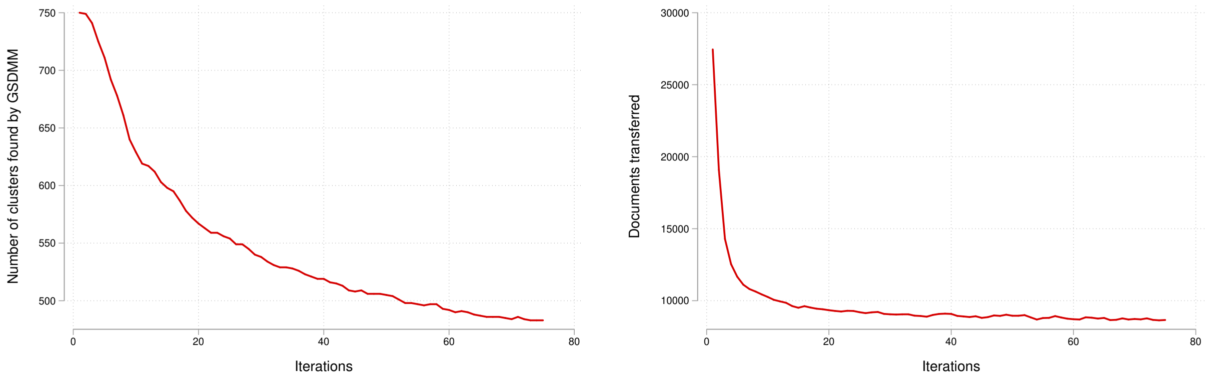
Note that tokenization splits a character sequence into tokens, which are meaningful semantic units for processing. Lemmatization reduces words to their base form or lemma. To implement these we use the small English model of *spaCy* trained on written text on the web such as blogs, news, comments etc. *spaCy* is an open source library used for advanced natural language processing in Python and Cython, and has pre-trained statistical models for over 60 languages. See https://spacy.io for more details.

The GSDMM algorithm first randomly assigns all documents to K clusters where K is a predefined upper limit on the number of topics given as a human input to the algorithm. As long as K is larger than the 'true' number of clusters, the algorithm can automatically infer the appropriate number of clusters. In each subsequent iteration the algorithm probabilistically re-assigns each document one-by-one to a cluster based on two considerations: (a) sharing a more similar set of words, and (b) having more documents. As the algorithm proceeds, some clusters grow larger and others disappear until finally each cluster contains a similar set of documents. Mathematically, a document d is assigned to cluster z with probability:

$$p(z_d = z|\vec{z}_{\neg d}, \vec{d}) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d}(n^w_{z,\neg d} + \beta)}{\prod_{i=1}^{N_d}(n_{z,\neg d} + V\beta + i - 1)}$$

where $\vec{z}$ is the cluster label of each document, m_z is the number of documents in cluster z, n_z is the number of words in cluster z and n^w_z represents the number of occurrences of word w in cluster z. $\neg d$ denotes that cluster label of document d is removed from $\vec{z}$. D refers to the total number of documents in the corpus, N_d is the number of words in document d and V is the total number of

Figure A.2: GSDMM Iterations and Clusters



(a) Number of clusters

(b) Number of transferred documents

Notes: Number of clusters found by GSDMM in each iteration (subfigure a) and number of documents transferred across clusters in each iteration (subfigure b).

words in the vocabulary.

The parameter α is related to the prior probability of choosing an empty cluster. For example, when $\alpha = 0$, the probability of choosing an empty cluster is 0. The parameter β relates to homogeneity of clusters. If $\beta = 0$, a document will never be assigned to a cluster if any particular word in the document is not contained within any document in a cluster, even if the other words of the document may appear in multiple documents in that cluster. Therefore, a positive value of β should be chosen. We set the initial number of clusters $K = 750$, $\alpha = 0.005$, $\beta = 0.005$ and run the model for 75 iterations.[45]

Yin and Wang (2014) use $\alpha = 0.1$, $\beta = 0.1$ and 30 iterations. However, we choose a smaller value of β to get more homogeneous clusters. We find that the overall performance of the algorithm is not sensitive to α in range [0,1], and therefore, choose $\alpha = 0.005$ to maintain the same ratio between α and *beta*. We choose the number of iterations such that the number of clusters becomes stable and the number of documents transferred across clusters also become very small post that number. We tried up to 100 iterations and found that at approximately 75 iterations both these criteria are met. Lastly, the initial number of clusters (K) were chosen to be approximately equal to the number of clusters obtained in the manual classification using n-grams. Figure A.2 shows that the number of clusters and the number of documents transferred across clusters initially falls sharply,

[45] We use the python implementation of GSDMM available at https://github.com/rwalk/gsdmm.

and then tends to stabilize after a few iterations.[46]

B.2 Preprocessing Bag-of-n-grams Logistic Regression

For preprocessing the data, we first remove all special characters, numbers as well as extra spaces, i.e. we retain only alphabets, and convert all the characters in the job text to lowercase. We remove all the words indicating explicit gender preferences as mentioned in sub-section 2.1. If we retain these words, our algorithm's accuracy will be artificially inflated by classifying jobs largely on the basis of words which were originally used to code employers' gender preferences. We also filter out stop words such as "the", "are", "and" which are uninformative in representing the text. We use the Stopwords corpus of the Natural Language Toolkit (NLTK) version 3.5. NLTK is a python package used for NLP. For more details, see https://www.nltk.org/. We also remove words having length less than 2 or greater than 15 characters, and then lemmatize the job text using the large English model of *spaCy*.

In a bag-of-words (BOW) representation, each document is represented as a vector based on the occurrence of words in it, without taking into account their relative position in the document. This generates a matrix where each row represents a document and each column indexes a word or a set of words (also known as a token) that occurs in the corpus.

A discriminative classifier such as LR directly learns the mapping from inputs x to the class label y by fitting a hyperplane in the input feature space to separate the classes.[47] A generative model such as NB (McCallum et al., 1998), on the other hand, tries to solve a more general problem of modeling the joint probability Prob(x,y) as an intermediate step and then uses Bayes rule to

[46] There is no direct way to assess objectively whether short text topic model or manual clustering performs better. Existing measures such as homogeneity and completeness used in the literature are not appropriate in our context since the true occupation categories are not known. The variable depicting job roles has very few categories to reflect true occupation categorization. In many cases two jobs involving similar tasks can often be assigned two or three different job roles. For example, the job ads titled "customer care executive" and "customer care professional" are both assigned job roles "BPO/Telecaller" as well as "Customer Service/Tech Support". While our topic model assigns them to the same cluster, the manual classification assigns them to different topics—"customer care executive" and "customer care" respectively. Similarly, "software engineer" and "software test engineer" are both assigned job roles "IT Software Engineer" as well as "Engineer (Core, Non IT)". These are assigned to same cluster by our topic model, but again assigned different occupations by the manual classification. Therefore, job role is an imperfect gold standard for measuring homogeneity. Nonetheless, we compute the homogeneity score and find that it has a value of close to 75% for the short text model. This indicates that job ads within a cluster largely belong to the same job role.

[47] The output y in our models is a variable indicating the presence and direction of explicit gender preferences of employers and can take three values. The input x is the bag-of-n-gram representation of text in job ads using $TF - IDF$ vectors for the LR model. In case of the NB classifier, the input x corresponds to binary-valued feature vectors indicating the presence or absence of n-grams in each job title.

calculate Prob(y|x). Consequently, LR has a lower asymptotic error, and is expected to outperform NB when the number of training examples is high enough, as in our case (Ng and Jordan, 2002).

B.3 TF-IDF Implementation

TF-IDF captures how important a token (or a set of words) is to a document with respect to its importance in the corpus based on its frequency. Therefore, it improves text classification by scaling down the weights of common tokens which are likely to be uninformative in capturing employers' preferences. We consider word unigrams, bigrams and trigrams, i.e., $n \in \{1, 2, 3\}$. For a token t in document d, the $TF - IDF$ score is computed as follows in our implementation:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

such that,

$$TF(t, d) = \frac{N_{t,d}}{N_d} \quad \text{and} \quad IDF(t) = ln\frac{1 + n}{1 + DF(t)} + 1$$

where, $N_{t,d}$ is the number of occurrences of token t in document d; N_d refers to the length of document d; $DF(t)$ is the number of documents in which token t appears; and n is the total number of documents in the corpus. Additionally, the $TF - IDF$ vectors for each document are normalized to have Euclidean norm 1. Therefore, TF captures how important a token is to a document, whereas IDF scales down the weight of tokens that occur very frequently in the corpus, and hence are less informative for our classification.

B.4 Stratified k-folds Cross Validation

In stratified 10-folds cross-validation, for each of the 10 “folds”, the model is trained on 9 folds (or 90% of the sample) and its performance is assessed using the remaining fold (or 10% of the sample) as the test set. If we use the same data for learning the parameters of the logistic regression model as well as evaluation, this will lead to overfitting, i.e. the model will perform exceptionally well on the training data, but will not generalize well. We also use $L2$ regularization to prevent overfitting with regularization parameter (inverse of regularization strength) equal to 0.35 and 0.45 to calculate F_p and M_p respectively. To do this the sum of squared weights (i.e. coefficients) are multiplied by a constant C and added to the loss function. This adds a quadratic penalty to the weights as they move away from zero to prevent overfitting. A methodological issue may arise when two documents with exactly the same text are assigned different probabilities if they belong to different test sets for which slightly different training data is used. This, however, does not pose a significant challenge for us as over 99% of the overall variance in the probabilities is explained between job texts, with the remainder explained within job texts.

Supplementary References

McCallum, A., K. Nigam, et al. (1998): "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, Citeseer, vol. 752, 41–48.

Ng, A. Y. and M. I. Jordan (2002): "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in neural information processing systems*, 841–848.

Yin, J. and J. Wang (2014): "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 233–242.