Discussion Papers in Economics

Words Matter: Gender, Jobs and Applicant Behavior

Sugat Chaturvedi, Kanika Mahajan, and Zahra Siddique

June, 2021

Discussion Paper 21-03

Indian Statistical Institute, Delhi
Economics and Planning Unit
7, S. J. S. Sansanwal Marg, New Delhi, 110016, India

Words Matter: Gender, Jobs and Applicant Behavior*

Sugat Chaturvedi[†] Indian Statistical Institute

Kanika Mahajan[‡] Ashoka University

Zahra Siddique[§] University of Bristol

June 20, 2021

Abstract

We examine employer preferences for hiring men vs women using 160,000 job ads posted on an online job portal in India, linked with more than 6 million applications. We apply machine learning algorithms on text contained in job ads to predict an employer's gender preference. We find that advertised wages are lowest in jobs where employers prefer women, even when this preference is implicitly retrieved through the text analysis, and that these jobs also attract a larger share of female applicants. We then systematically uncover what lies beneath these relationships by retrieving words that are predictive of an explicit gender preference, or *gendered words*, and assigning them to the categories of hard and soft-skills, personality traits, and flexibility. We find that skills related female-gendered words have low returns but attract a higher share of female applicants while male-gendered words indicating decreased flexibility (e.g., frequent travel or unusual working hours) have high returns but result in a smaller share of female applicants. This contributes to a gender earnings gap. Our findings illustrate how gender preferences are partly driven by stereotypes and statistical discrimination.

Keywords: Gender, Job portal, Machine learning
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, Rossella Calvi, Stefano Caria, Rochana Chaturvedi, Shoumitro Chatterjee, Sekyu Choi, Seema Jayachandran, Aprajit Mahajan, Arnaud Philippe, Nishith Prakash, Hans Sievertsen, Shekhar Tomar, Basit Zafar and Yanos Zylberberg for useful discussions and feedback. We would also like to thank SOLE (2021) conference participants and seminar participants at Samvaad, Association for Mentoring and Inclusion in Economics (AMIE), IWWAGE-ISI, India Data Foundation, the University of Bristol, 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

Persistent gender disparities in the labor market could indicate that innately talented women are not pursuing their comparative advantage. The resulting misallocation has a detrimental effect on economic growth (Hsieh et al., 2019). One reason why such disparities can arise is due to gender stereotypes, e.g. regarding the relative ability of men vs women in different kinds of jobs (Bertrand, 2020). Notably, these stereotypes can lead to disparities within, and not just across, occupations. This can happen during the recruitment process, with job advertisements playing a key role. Employers can choose words within job ads to effectively recruit particular kinds of workers (e.g. men vs women). These words also reveal gender stereotypes associated with job roles held by employers. However, there is surprisingly little research looking into how words in job ads are associated with particular stereotypes, how they relate to different characteristics of jobs and the posted wage, or how they direct the job search behavior of men and women who are looking for work. We investigate these questions in this paper.

We examine employer's strategies aimed at recruiting male or female applicants either through an explicitly stated gender preference or through *implicit hints* within the job ad text. We make use of proprietary data on $\approx 160,000$ job ads posted between July 2018 and February 2020 on a leading Indian job portal, which are linked with 6.45 million applications. We first analyze job ads with explicitly stated gender preferences and show that such ads are highly effective in shaping the gender mix of the applicant pool. We then apply text analysis on detailed job descriptions to construct measures that tell us how predictive the job ad text is of an explicit gender preference. We find that implicit hints within the job text (even without an explicit gender preference) continue to be highly effective in directing where male and female job seekers send their applications. Lastly, we examine the role played by *gendered words* (or words predictive of an employer's explicit preference) related to desired hard and soft-skills, personality traits, or flexibility. We find that gendered words related to hard-skills and flexibility are the most important in explaining gender disparities in labor market outcomes. Importantly, we show that employer's gender preferences are partly driven by stereotypes and statistical discrimination.

The job portal we use primarily caters to young urban job seekers with a university education. Jobs advertised on the portal are high-skill jobs with posted wages that are, on average, 21%

higher than wages earned by a nationally representative and comparable sample of employed Indian workers. Consistent with low female labor force participation rates in India, we find that there are only half as many female as male applicants who search for jobs using the portal. However, these female applicants are more educated than male applicants and make a similar number of job applications, on average. Our work investigates disparities in the kinds of jobs these male and female job seekers direct their applications to.

We start by documenting two empirical results related to explicit gender requests which employers include in the job ads they post on the portal. First, we find that advertised wages are lowest in jobs that request women.[1] Second, we find that explicit gender preferences by employers reduce the total number of applications to a job ad while substantially changing the gender mix of the applicant pool in favor of the requested gender. Consequently, explicit gender requests result in female applicants applying to jobs with a *lower* advertised wage compared to men with similar characteristics. Our empirical results become attenuated but persist when we include detailed occupation and state fixed effects in our regressions. We do this by classifying job ads into 483 disaggregate occupation categories based on job titles using a topic model.[2]

We also construct measures that indicate whether the job ad text is predictive of an employer's explicit male or female preference using a multinomial logistic regression (LR) classifier. We refer to these as a job ad's implicit *maleness* and *femaleness*. We find that even among jobs without an explicit gender preference, higher implicit *femaleness* is associated with a substantial reduction in the advertised wage and increases the share of female applicants to a job ad. In other words, women continue to direct their applications toward job ads where the job text contains implicit hints that the employer might prefer a female and which have low advertised wages. All of our results persist when we use within occupation and state variation only. Our results also remain largely robust when we use firm $\times$ state fixed effects, or firm $\times$ occupation $\times$ state fixed effects.

We then systematically examine what lies beneath these relationships by using the Local Inter-

[1] A key advantage of the wage data we use is that employers provide an advertised wage for slightly over 87% of job ads in our sample, which is far higher compared to 20% of job ads in Marinescu and Wolthoff (2020) using *Careerbuilder* and 16.4% of job ads in Kuhn and Shen (2013).

[2] Specifically, we use a short text topic model (an unsupervised machine learning method) on text contained in job titles to categorize job ads in our sample into occupations. The use of a topic model provides dimension reduction compared to a manual classification using unigrams, bigrams, and trigrams in job titles used in the existing literature (Marinescu and Wolthoff, 2020; Banfi and Villena-Roldan, 2019). All of our results remain robust to using a manual classification on our sample which yields 747 occupations.

pretable Model-agnostic Explanations (LIME) algorithm from the literature on machine learning. This allows us to identify words in job ads that contribute to the classification decisions of the LR classifier. We refer to these words as *gendered words*. We assign gendered words to the categories of hard and soft-skills, personality traits, and job flexibility (vis-à-vis job timings and travel requirements). We examine the association of the constructed gender-categories with advertised wages and the female share of applicants to a job ad. This yields two important and novel findings. First, we find that job ads with female gendered words related to hard-skills are associated with a lower wage but attract a higher share of female applicants. This indicates a pattern whereby women acquire and respond to skills in stereotypical female job roles which are associated with lower wages contributing to a gender earnings gap. Second, we find that job ads with male gendered words related to (reduced) job flexibility are associated with higher wages but get a smaller share of female applicants. This further contributes to a gender earnings gap and is consistent with compensating differentials whereby women are willing to trade off higher wages for increased flexibility (Goldin and Katz, 2011; Goldin, 2014; Flory et al., 2015; Mas and Pallais, 2017; Wiswall and Zafar, 2018; He et al., 2019; Bustelo et al., 2020). Importantly, our results indicate that there is more than taste-based discrimination at work, so that employers (perhaps anticipating the importance of scheduling arrangements for women) frequently make explicit male requests in jobs with an inflexible schedule. Our list of gendered words reveals that employers' gender preferences are, therefore, partly driven by stereotypes and statistical discrimination.

We use ridge regression to directly identify words that attract a higher share of female applicants in job ads within an occupation and state. We find a positive correlation of the gendered words with words that attract more female applicants within the flexibility and hard-skills categories. However, there is a zero and negative correlation of gendered words with words that attract more female applicants within the soft-skills and personality categories respectively. So, while words such as *punctual*, *smile*, and *pleasant* are highly predictive of an employer's female preference, we find that they *reduce* the share of women in the applicant pool.

Our work extends and complements several strands of literature. This paper is the first to systematically uncover a list of gendered words that reveal gender stereotypes in job ad text and their association with posted wages and candidate search behavior. To do this, we use research in explainable artificial intelligence, employing a method hitherto not applied in the field of economics.

This allows us to gain new insights on the importance of skills and flexibility related gendered words, as well as highlighting the role played by stereotypes and statistical discrimination in the gender preferences of employers. This list is also likely to be useful to researchers interested in uncovering gender associations or stereotypes in a similar labor market setting who do not have information on explicit gender preferences. Therefore, it can be a useful resource for detecting bias in text data.[3]

We directly contribute to the literature comprising a relatively small number of papers that make use of job ad text to examine gender disparities in the labor market. Although our work also highlights the importance of words in job ads, the research questions we address are quite distinctive from this work. Abraham and Stein (2020) employ an RCT design using job postings for Uber's U.S. corporate positions and find that removal of optional qualifications and superfluous language in job ads increases applications by low-skill female and high-skill male applicants. Marinescu and Wolthoff (2020) show the significance of text contained in job titles in the matching process and use this text to assign jobs to disaggregate occupations. Kuhn et al. (2020) use text contained in job titles to construct measures that are predictive of an employer's explicit gender preference for a particular occupation. In contrast, we make use of the text contained in detailed job descriptions as well as job titles. We find that implicit *femaleness* and *maleness* can vary greatly across job ads with the same title (and occupation category). For example, consider two job ads titled 'sales executive' in our sample. Across the two ads, implicit *femaleness* is high when the job description emphasizes appearance or communication skills, while implicit *maleness* is high when the job requires fieldwork. Given our use of the entire job ad text, we are able to exploit variation in implicit associations *within* an occupation and state.

We also contribute to the literature documenting the presence of explicit gender preferences in job ads (Kuhn and Shen, 2013; Helleseter et al., 2020; Ningrum et al., 2020; Chowdhury et al., 2018). In contrast to much of this literature, with the exception of Kuhn et al. (2020), we observe applications data which allows us to investigate how job seekers *respond* to explicit gender requests. We extend this literature by deconstructing explicit gender preferences of employers into hard and soft-skills, personality, and flexibility related words. Our findings illustrate that explicit gender requests in job ads do not necessarily reflect employer's taste-based discrimination only.

[3]For instance, Burn et al. (2019) use word vectors to calculate cosine similarity of words from the industrial psychology literature with phrases in job ads to detect bias against older workers in the US.

More generally, we contribute to a growing literature on various aspects of labor markets using high-frequency data from online job portals.[4] In particular, we contribute to the 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 examine job search behavior of employed and unemployed job seekers. We also find evidence that higher posted wages attract more applicants; in addition, our work shows the key role played by employer's gender preferences in job search. Further, our use of a topic model as an alternative to a manual classification of job titles to disaggregate occupation categories is also likely to be useful and relevant for this literature.

Finally, we contribute to the literature examining gender differences in job search. Recent contributions include Morchio and Moser (2020) and Xiao (2020) who use matched employer-employee data from Brazil and Finland to estimate equilibrium search models which incorporate gender differences in preferences for job amenities and employer discrimination. Cortes et al. (2021) examine the importance of gender differences in risk aversion and optimism about prospective offers in job search using data collected from recent Boston University graduates in the U.S. Our analysis complements this literature by using micro-evidence to highlight the importance of words within job ad text in explaining gender differences in job search.

Apart from economics, our work is also related to a literature in social psychology. Born and Taris (2010) and Gaucher et al. (2011) show that women find job ads that contain *feminine* words more appealing. However, unlike us, these studies rely on student samples with none considering actual applications.[5] Additionally, within this literature, the characteristics that attract women

[4] Hershbein and Kahn (2018) use data from job vacancies to investigate how skills demand changed over the Great Recession in the US. Deming and Kahn (2018) classify skills related key words in the job ad text and examine their impact on aggregate regional pay measures. Several recent papers examine changes in labor demand with the onset of the Covid-19 pandemic using job portal data: 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.

[5] Born and Taris (2010) find that women respond more to feminine characteristics than men respond to masculine characteristics among 78 applicants. The study uses 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 are more likely to find job ads appealing where a greater proportion of feminine words were used and candidates were also more likely to anticipate gender diversity in roles advertised in such job ads.

and men to specific job ads are drawn from small, non-representative surveys.[6]

In Section 2 we provide a description of our data set and constructed variables. In Section 3 we discuss our empirical methodology and results when examining employer's gender preferences. Section 4 deconstructs the words used by employers to express a gender preference and examines their impact on different outcomes. Robustness checks are reported in Section 5. Section 6 examines words in job ads that attract a higher share of female applicants. Section 7 concludes.

2 Data

We analyze data from a leading job portal in India that primarily caters to young job seekers with job locations across all major Indian cities. 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 ($\approx$ USD 20 per ad). 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, location, education, job type (govt/private), and keywords. Job seekers who additionally register for a premium service are provided customized job recommendations and alerts on new jobs by e-mail. The proportion of job seekers who registered for this service in our data was $\approx 0.5\%$; hence, the chances that applications are driven by matching algorithms used by the portal are negligible.

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, along with data on all applications made to these ads. We use data on *active* job ads and job seekers, i.e. 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 period.

2.1 Job ads

In total 196,821 job ads were posted on the portal during our study period. We exclude vacancies outside India and those having 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 posted within a

[6] 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.

month of the original ad. This leaves us with 1,75,126 unique ads.[7] We further drop job ads that 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 (reducing the sample to 1,71,940 ads) and job ads that specify cities within a single Indian state as the location of the job (reducing the sample further to 1,58,249 ads).[8] We also restrict the sample to those jobs for which we are able to obtain an occupational classification based on the method described in Section 2.3, leaving us with a final sample of 1,57,888 job ads.

We construct variables indicating an employer's explicit gender preference. The portal does not have a separate field that allows employers to directly state the preferred gender for a position to applicants. However, employers indicate an explicit gender preference in the job title or description so we search this text for the following words which indicate an explicit female preference: 'female', 'females', 'woman', 'women', 'girl', 'girls', 'lady' and 'ladies'. Similarly, we search for the words 'male', 'males', 'man', 'men', 'guy', 'guys', 'boy', 'boys', 'gent' and 'gents' which indicate an explicit male preference.[9] Some job ads include words related to both genders. We categorize such ads as having no explicit gender preference, together with ads that 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).[10,11] Figure 1 shows word clouds of job titles in F, M and N jobs for our sample. As may

[7]Duplicate job ads are defined as ones which have exactly the same requirements and job description, as well as being posted by the same firm. Approximately 70% of duplicate job ads were posted within a month of the original ad. We keep duplicates posted more than a month after the original ad since these are likely to be new vacancies. When examining applicant behavior, we aggregate applications across duplicated ads to ensure we use data on *all* job seekers applying to a job.

[8]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.

[9]Our data contains 15,400 job ads where at least one of the words related to either gender is mentioned. However, just looking at the occurrence of these words may be misleading. Therefore, we first exclude the subset of job ads which combine these words with qualifiers that unambiguously indicate a gender preference—for instance, 'female only', 'female preferred', 'looking for female', 'require female', 'wanted female' etc. There are 10,001 job ads with phrases indicating a clear gender preference. The remaining job ads (5,399) were shown to two annotators who independently classified the job ads, based on the job title and description, as explicitly requiring a female, male, or having no explicit gender preference. The annotators agreed on the classification for 90% of the job ads. The remaining 10% were shown to a third annotator, whose judgment was used to classify the remaining ads into one of the three categories.

[10]The fractions of F and M jobs we find are smaller than those reported by Chowdhury et al. (2018) using data from *Babajob*. This is probably because, unlike the job portal we use, *Babajob* had a separate field where employers could directly state the preferred gender to applicants. A third of all job ads in their data used this field of which 21% preferred men and 14% preferred women.

[11]While the proportion of job ads with a gender request is about 8%, the absolute number of F and M in our sample is over 12,000 which is sufficiently large to train our ML models. To give equal importance to F, N, and M classes, we use balanced class weights while training the models. We are able to extrapolate the 'maleness' and

be seen, 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 suggests that explicit gender preferences operate to maintain existing occupational gender stereotypes.

A very small proportion of jobs advertised on the portal specified the education requirement as none (or illiterate). We group these with jobs requiring a secondary education or less as the base category in our empirical analysis. N jobs tend to have higher education requirements than F or M jobs while F jobs tend to have higher education requirements than M jobs. For instance, around 53% of N jobs require at least an undergraduate degree as opposed to 47% and 29% of F and M jobs respectively (Appendix Table A.1). F jobs are also far more likely than M jobs to require an undergraduate degree in a non-STEM subject. Consistent with the portal catering primarily to young job seekers, most job ads ($\approx 67\%$) require less than a year of experience. We also find that N jobs are more likely to require two or more years of experience compared to other jobs.[12]

Employers include a wage range in over 87% of the jobs advertised on the portal. 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 observe wages for a much larger fraction of job ads than existing studies. In comparison, wages are advertised in just 20% of job ads in Marinescu and Wolthoff (2020) using *Careerbuilder* and 16.4% of job ads in Kuhn and Shen (2013) using a Chinese job portal. We take the mid-point of the wage range as our measure of the posted wage. The mean of the posted wage is just above INR $213,000$ per year.[13] N jobs have higher mean posted wage than F and M jobs. Moreover, M jobs have a *higher* mean posted wage than F jobs despite having lower education requirements.

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

'femaleness' measures to N jobs since words contained in jobs with an explicit gender request comprise over 97% of words in N jobs by volume.

[12]The literature on gender targeting in job ads has documented that ads specifying a gender preference are more likely to specify other preferences, such as those related to age or beauty. We also derive the presence of age and beauty preferences from the text of the job ad (see Appendix B for details). We find that M jobs are more likely to specify an age preference and these jobs tend to specify a higher minimum and maximum required age than F or N jobs. On the other hand, F jobs are most likely to specify a beauty requirement.

[13]A very wide wage range is likely to be uninformative to job seekers. Therefore, we take the posted wage as missing if the range is greater than INR 2 million. We also replace the data at both top 0.5 percent and bottom 0.5 percent of the distribution to missing to mitigate the effects of any extreme outliers.

to 52% while for M jobs it falls to 13%. This indicates that there is some compliance with explicit gender requests but this compliance is far from perfect. Overall compliance with gender requests in F and M jobs, i.e. percent applications that are of the requested gender is 68%. 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.[14] 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% ($\kappa = 25\%$). Thus, compliance with gender requirements is lower than with education requirements but higher than with experience requirements.

On average, there are about 41 applications per job ad. 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 number of applications, particularly by job seekers of the opposite gender to the preferred one.

2.2 Job seekers

We also use data on 1.06 million job seekers who applied to at least one job using the portal. Appendix Table A.2 gives descriptive statistics for job seekers by gender. 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 women is less than a third of men, there are slightly more than half as many female job seekers as male job seekers on the portal. On average, female job seekers make a similar number of job applications as male job seekers. Most job seekers on the portal (86 percent) have an undergraduate or postgraduate degree though women are more likely to have a postgraduate degree. Job seekers are relatively young with an average age of 24 years and about 76% have less than a year of experience. Female job seekers are slightly younger than men and are less experienced. Lastly, women (unconditionally) apply to job ads with similar posted wages as men. However, conditional on candidate characteristics, female job seekers (who tend to be more educated than males) apply to jobs with 4 log points lower posted wages

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

than do male job seekers.[15]

We compare job seekers on the portal with the urban working-age population in India using the Periodic Labor Force Survey 2017–18 (PLFS), which is a nationally representative survey of employment in India. Appendix Table A.3, Panel A reports the average annual earnings in PLFS for casual or salaried workers among working-age adults (age 16–60) in urban Indian districts (with $\geq$ 70% urban population).[16] Advertised wages on the portal are higher than PLFS by about Rs. 14,000 per annum. However, wages in PLFS could be high because it has older and more experienced workers. To make the PLFS sample comparable to the age group catered to by the online portal we only keep adults aged 18–32 years in Appendix Table A.3, Panel B since around 95% of job seekers on the portal belong to this age group. This increases the gap in annual earnings to more than Rs. 37,000 per annum and the average advertised wage on the job portal is now 21% higher. Thus, the portal caters to young and inexperienced but more educated and skilled workers.

Figure 2 further confirms these patterns. The wage distributions for urban workers using the PLFS data are centered at a lower log wage and more dispersed compared to the distributions of posted wages on the job portal. This is particularly true for female wage distributions, indicating that gender wage disparities among Indian workers exceed disparities in posted wages across F and M jobs on the portal (Figures 2(a) and 2(c)). Restricting the PLFS sample to employed urban workers aged 18–32 with at least an undergraduate education, we find that gender wage disparities are now comparable to disparities in posted wages across F and M jobs on the portal (Figures 2(b) and 2(c)).

We also regress the log wage on a gender indicator, education, age, and occupation $\times$ state fixed effects using the sample of employed workers in PLFS aged 18–32 years having at least an undergraduate education. We find that the gender wage gap in this sample is 8 log points. On the portal, the gender wage gap in job applications is 2 log points after controlling for candidate education, age, and occupation $\times$ state fixed effects. This indicates that around 25 percent of the gender wage gap among employed workers, within an occupation in a given state, may be driven

[15] We estimate this specification at the application rather than candidate level to estimate the gender wage gap in applications. We regress the log of the posted wage for the applied job on candidate characteristics, giving each candidate equal weight. In another specification we also control for occupation controls within a state. Here we find that women, on average, apply to jobs with 2 log points lower wages than men.

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

by women applying to lower-wage jobs.

2.3 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.

We use an unsupervised machine learning technique to classify semantically similar job titles into 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 assumes that each document (or in our case, job title) comprises a single topic—an assumption suitable for short texts. The algorithm probabilistically combines job titles into occupation groups such that titles in the same group contain a similar set of words, whereas titles in different groups contain a different set of words. We provide details of the data pre-processing steps, algorithm, and the hyperparameter choice in Technical Appendix Section C.1. The final number of topics or occupation categories discovered by GSDMM for our sample of job ads is 483.

Our empirical results are robust to an alternative manual clustering of job ads to occupation categories based on word unigrams, bigrams, and trigrams in job titles as used in the existing literature (Banfi and Villena-Roldan, 2019; Marinescu and Wolthoff, 2020).[17] This gives us a total of 747 occupation categories. We prefer GSDMM to the manual classification as it provides dimension reduction based on the co-occurrence of words in the corpus of job titles. This is accomplished by probabilistically clustering together job titles that do not share any common word, but are linked

[17]To implement the manual clustering we first calculate n-gram counts after removing duplicate job ads, i.e. those posted by the same employer, with the same job title and description. We then classify jobs based on the most frequently occurring trigrams in job titles, subject to the trigram existing in at least 50 titles. The remaining ads are classified based on the most frequently occurring bigrams, and then unigrams in the titles with the restriction that the bigrams and unigrams occur in at least 100 job titles. The precedence given to higher-order n-gram based on the frequency of occurrence ensures that each ad is classified into no more than one cluster or occupation category. We discuss estimation results using the alternative categorization in Section 5.

together through sharing common word(s) with some other titles that act as a bridge between the two. For instance, ads titled 'english transcriber' and 'japanese translator' are assigned the same occupation cluster as they are linked through 'transcriber-translator'. These job ads cannot be assigned the same occupation using the manual classification as they do not share any common word. Our use of GSDMM also ensures that most of the job ads in our sample get assigned to meaningful occupation clusters. In contrast, over 5,800 job ads ($\approx 3\%$) could not be assigned any occupation using the manual classification because the word n-grams contained in them occur with a low frequency across the corpus.

2.4 Implicit *femaleness* and *maleness*

The text contained in a job ad may also convey an implicit signal to a job seeker about whether the employer posting the ad prefers a female or a male—even in the absence of an explicit gender preference. We define 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. Specifically, we train a Multinomial Logistic Regression (LR) classifier with balanced class weights where the output class can take three values depending on the employer making an explicit request for women, men, or no gender request.[18] The probabilities F_p and M_p returned by the model capture employer requests very well with correlations of 0.38 and 0.44 with binary variables indicating explicit female and male requests respectively.[19] Consistent with this, Appendix Figure

[18] We use the complete set of 196,857 job ads 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 due to a relatively smaller fraction of jobs having an explicit gender request. We concatenate the job title and description to get the complete job text and follow standard pre-processing steps (see Technical Appendix Section C.2). We then convert each processed document to its bag-of-n-grams representation using term frequency-inverse document frequency (TF-IDF) vectors which we use as inputs to the model (see Technical Appendix Section C.3). We use stratified 10-folds cross-validation wherein we split the data into 10 parts while preserving the class proportions in each split (see Technical Appendix Section C.4). F_p and M_p are then the estimated probabilities of a document belonging to the female or male class when the document belongs to the test split.

[19] Conditional on the employer making an explicit gender request, the model correctly predicts requests for women and men in 74.43% and 72.18% of job ads when they are part of the test set. It also correctly predicts the absence of a gender request in 79.50% of job ads, again when they are part of the test set.

A.1 shows that, on average, F_p has higher values for F jobs while M_p takes on higher values for M jobs. For N jobs, both F_p and M_p have a similar distribution.

We find that F_p is high for job ads with titles such as 'beautician', 'personal secretary', and 'school teacher', while M_p is high for job ads with titles such as 'cargo loader', 'delivery executive' and 'network engineer'. Even for the same job title, F_p and M_p can vary greatly based on the job description. For instance, consider two job ads titled 'business development executive' in the data; F_p is high when the job description mentions working from home or restarting a 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 while M_p is high for jobs requiring fieldwork.

We also train a Bernoulli Naive Bayes (NB) classifier on our data using the methodology in Kuhn et al. (2020) and find that it does not perform well in our context. It gives worse measures of F_p and M_p in our data with correlations of 0.23 and 0.22 with explicit employer requests for women and men respectively.[20] Additionally, the LR classifier allows us to better exploit a longer text which includes the job description as well as title for each job ad; NB is less suitable for longer texts as it uses *word occurrence*, rather than *word count* vectors.

3 Gender preferences of employers

We discuss the characteristics of job ads associated with gender requests by employers, both across and within occupations, in Appendix B. Broadly, our findings are in line with the existing literature that finds a negative skill-targeting relationship i.e. jobs with a higher skill requirement are *less* likely to have an explicit gender preference. We find that jobs with explicit gender preferences have lower education requirements (Columns (I)-(III), Appendix Table B.1). We also find that jobs with an explicit male preference require less education but offer higher wages than those with an explicit female preference (Columns (IV)-(VI), Appendix Table B.1). The latter is evident from our finding that a higher advertised wage is associated with an increased preference for men. Next, we examine if posted wages also vary with implicit gender associations in the job ad text and how

[20] The requests for women and men 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. This demonstrates that even though NB is a reasonable classifier, it does a poor job of estimating probabilities associated with the classes.

applicant behaviour responds to explicit gender requests as well as implicit gender associations.

3.1 Empirical methodology

To investigate whether wage differentials are associated with text predictive of employers' explicit gender preferences, we estimate variations of the following Mincer regressions separately for F, N, and M jobs:

$$\ln W_{ijst} = \alpha^W + \lambda^W F_{p,ijst} + \nu^W M_{p,ijst} + \beta^W X_{ijst} + \gamma_{j\times s} + \phi_t + \varepsilon_{ijst} \tag{3.1}$$

$\ln W_{ijst}$ is the log of the posted wage in job ad i advertising for a job of occupation j in state s and month-year t. $F_{p,ijst}$ and $M_{p,ijst}$ are measures of implicit *femaleness* and implicit *maleness* respectively (see Section 2.4). The coefficients on these variables (λ^W and ν^W) tell us how the advertised log wage changes as the implicit *femaleness* or *maleness* of a job ad increases from zero to one, everything else equal. X_{ijst} is a set of dummy variables for education and experience requirements. In our preferred specification, we include occupation and state fixed effects ($\gamma_{j\times s}$) as well as month-year fixed effects (ϕ_t). We use a detailed categorization of jobs to occupations with 483 distinct occupation categories derived from job titles (Section 2.3). The use of fixed effects ensures that we use *within* occupation and state variation only to identify the effect of different variables on the log wage. 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 regression specification:

$$Y^{TA}_{ijst} = \alpha^{TA} + \pi^{TA} F_{ijst} + \theta^{TA} M_{ijst} + \beta^{TA} X_{ijst} + \gamma_{j\times s} + \phi_t + \mu_{ijst} \tag{3.2}$$

where Y^{TA}_{ijst} is the total number of applications to a job ad. F_{ijst} is a binary variable and takes the value 1 if ad i has an explicit female preference, and 0 otherwise. Similarly, M_{ijst} is a binary variable taking the value 1 if ad i has an explicit male preference, and 0 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 dummy variables for education and

experience requirements. We include occupation and state fixed effects ($\gamma_{j\times s}$), month-year fixed effects (ϕ_t), and cluster standard errors by occupation and state.

To examine job seekers' compliance with the gender requirement set by the employer, we estimate variations of the following specification:

$$Y^S_{ijst} = \alpha^S + \pi^S F_{ijst} + \theta^S M_{ijst} + \beta^S X_{ijst} + \gamma_{j\times s} + \phi_t + \xi_{ijst} \tag{3.3}$$

where Y^S_{ijst} is the share of female applicants to a job ad. This is similar to (3.2) except that the regressions in (3.3) are weighted by the total number of male and female applications made to a job ad. The coefficients on the binary variables (π^S and θ^S) give the difference in the share of female applicants across ads exhibiting an explicit female or male preference relative to ads that exhibit no such preference (the base category), everything else equal.

We also examine how F_p and M_p derived from job ad text affect the female applicant share. To do this, we follow the strategy in Kuhn et al. (2020) and regress the share of female and male applicants to a job on explicit gender requests as well as quartics in F_p and M_p. We include the set of controls in equation (3.3) and use specifications with and without occupation and state fixed effects.[21] Further, we interact the quartics in F_p and M_p with explicit gender requests and use these as additional explanatory variables. We then use the regression estimates to predict and plot the share of female (male) applicants as a function of F_p (M_p) for each type of job (F, N and M).

3.2 Results

We examine the estimation results for equation (3.1)—estimated separately for F, N, and M jobs—which provide information on the effect of implicit *femaleness* and *maleness* on the advertised wage. The results are reported in Table 1. As expected, higher education and experience requirements increase advertised wages for all kinds of jobs. For N jobs, an increase in *femaleness* from 0 to 1 reduces the offered wage by 38 log points without occupation and state controls (Column (III), Table 1). After including occupation and state fixed effects, the effect of *femaleness* on offered wages drops to 26 log points but remains highly statistically significant (Column (IV), Table 1). This coefficient estimate translates to a decrease in the advertised wage of 5.2 log points with a

[21] We do not include wage controls to use the full sample of job ads.

one standard deviation increase in implicit *femaleness* ($SD = 0.2$). On the other hand, an increase in *maleness* from 0 to 1 is associated with a smaller decline in wages ($\approx 12 - 14$ log points); 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 the employer prefers a female from the job text) offer systematically lower wages even when the ad does not exhibit an explicit gender preference. We find similar patterns in F and M jobs but the negative effect of *femaleness* on log wage is smaller, though it is still statistically significant. The negative effect of *maleness* on log wage in F jobs is not significantly different from zero but becomes larger and statistically significant in M jobs.

To examine the effect of explicit gender preferences on applicant behaviors, we estimate and report regressions specified in equation (3.2), where the outcome variable is the total number of applications to a job ad; the results are reported in Columns (I)–(III), Table 2. We find that the number of applications decreases dramatically (≈ 21; 51% of mean) if a job ad exhibits an explicit female preference. The decline is smaller when we use within occupation-state variation only, but remains statistically significant ($\approx 5 - 8$; 13–20% of mean). On the other hand, an explicit male preference does not significantly reduce the total number of applications to a job ad. Consistent with directed search models, we also find that there is a statistically significant increase in the number of applications to a job ad as the advertised wage increases when we use within occupation and state variation; a 1% increase in the advertised wage increases the number of applications to a job ad by approximately 19 (Column (III), Table 2).

Next, we estimate and report the regressions specified by equation (3.3) where the outcome is the fraction of female applicants to a job ad; the results are reported in Columns (IV)–(VI) of Table 2. We find that, within an occupation and state, the fraction of female applicants to a job ad increases by $15.5 - 15.6$ percentage points when an ad exhibits an explicit female preference and reduces by $9.5 - 9.9$ percentage points when the ad has an explicit male preference (Columns (V)–(VI), Table 2). These translate into an increase of 48% and a decrease of 30% over the mean share of female applicants to a job ad respectively—which are substantially large effects. In addition, we find that a higher fraction of women apply to job ads that have a higher education or lower experience requirement. This is likely to be driven by more educated and younger women on the portal (Appendix Table A.2). We find that the advertised wage does not affect the share of

female applicants. However, this specification does not control for applicant characteristics which are likely to be important since female applicants on the portal are more educated than male applicants. Therefore, we also estimate regressions at the application rather than job ad level to control for applicant characteristics. The dependent variable, in this case, is whether or not an applicant to a job ad is female, with controls for job characteristics (as in equation (3.3)) as well as controls for candidate education and a quadratic in candidate age. Using this specification, we find that women are significantly *more* likely to apply to jobs with a lower advertised wage (Column (III), Appendix Table A.4). Within an occupation-state, a 10 percent increase in the posted wage decreases the probability that a female applicant applies by 6 percentage points (19% of the mean).

Lastly, we examine how the gender mix of applicants changes as the job text becomes more predictive of an explicit gender preference across F, N, and M jobs. Figure 3(a) gives the predicted share of female (male) applicants as F_p (M_p) changes while controlling for M_p (F_p) and using a specification without occupation and state fixed effects. Strikingly, it shows that the predicted share of female applicants increases as F_p rises not only for N jobs, but also for F and M jobs. This increase is almost linear for N jobs and as F_p increases from zero to one, the share of female applicants increases from 35 percentage points to 45 percentage points—a 29% increase (p-value < 0.01). 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 and at high F_p for M jobs, though the effects are imprecise for M jobs. The predicted share of male applicants also increases as M_p increases for M, N, and F jobs; however, there is a decline in this share at high M_p for F jobs. Again, the effect is highest and most consistent for N jobs, where an increase in M_p from zero to one increases the share of male applicants by 32%.

Figure 3(b) plots the predicted share of female (male) applicants as F_p (M_p) changes but using within occupation and state variation only. We find 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 points to 39 percentage points or by 15% (p-value < 0.01) for N jobs. The female applicant share increases with an increase in F_p for F and M jobs as well at low and high levels of F_p respectively.[22] Similarly, as M_p increases (or as we move along jobs with an increasingly male job description *within* the same

[22]Surprisingly, the female applicant share initially declines with higher F_p in M jobs, however this decline is noisy and not robust to the use of firm fixed effects (Section 5).

occupation and state) from zero to one, the predicted share of male applicants increases by 16% for N jobs. For M jobs, the male applicant share increases with M_p but this effect is imprecise.[23]

Our results bear similarities and differences from those reported by Kuhn et al. (2020). We too find that the difference in the predicted share of male applicants between M and N jobs is generally smaller and further declines as M_p increases in comparison with the difference in the predicted share of female applicants across F and N jobs as F_p increases. Thus, 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 implicit gender associations seem to play a role in changing the gender mix of the applicant pool even in F and M jobs.[24] Importantly, this persists even within a given occupation in a state, though the magnitudes decline.

4 Deconstructing gender preferences of employers

The analysis in Section 3 shows that higher *femaleness* relative to *maleness* is associated with a lower advertised wage and that both *femaleness* and *maleness* have an impact on the gender mix of the applicant pool. A natural question that arises is: what kind of words contribute to the implicit gender associations? In this section, we address this by deconstructing the gender preferences of employers. We examine words that contribute to *femaleness* and *maleness*, which we refer to as *gendered words*. We assign *gendered words* to meaningful categories (related to hard and soft-skills, personality traits, and flexibility) and examine which gendered word-categories (henceforth referred to as gender-categories for brevity) drive changes in advertised wages and affect female applicant shares. We also examine words in job descriptions which attract a larger share of female applicants for N jobs.

[23] In general, predictions at very high values of F_p and M_p are not precisely estimated since there are few job ads with these extreme values.

[24] This difference is not driven by the different ML classifier used in our paper. We also re-construct our measures of F_p and M_p using the Bernoulli NB classifier (Appendix Figure A.2(d)). We estimate similar regressions as before to find the predicted share of female (male) applicants using 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.

4.1 Empirical methodology

We use the Local Interpretable Model-agnostic Explanations (LIME) algorithm proposed by Ribeiro et al. (2016) to understand which words correspond to explicit gender preferences of employers. LIME can explain the predictions of any classifier and overcomes the *black box* nature of complex machine learning 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 given job ad) and then obtaining predictions $f(x)$ returned by the machine learning model f. This gives a new data set of inputs (i.e. perturbations of the job ad) with predictions for every perturbation on which an interpretable weighted model (or *surrogate* model) is trained.[25]

LIME has been used to explain predictions made by machine learning models in many applications ranging from the 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 systematically extract individual words that reflect gender associations. Explainability in itself might be desirable to assess the validity and generalizability of a model, and hence to gain trust in its predictions.[26] We use the LIME algorithm 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 the Multinomial Logistic Regression classifier and to assign contributions of individual words to the female, neutral and male class below.

Word scores: We map the classification scores returned by the Multinomial Logistic Regression classifier into the input space using the LIME algorithm over test set documents. This allows us to assign a relevance score $R^G_{i,w}$ to every word w in each job ad i which indicates the importance of that word to class $G \in \{F, N, M\}$.[27] Figure 4 shows a heat map visualization of words in distinctive F (Panel (a)) and M (Panel (b)) job ads. Job ads (i), (ii) and (iii) in both panels refer to jobs titled

[25] To approximate a model locally (instead of globally), the weights are assigned based on the similarity of the perturbed instance to the original job ad.

[26] A model can spuriously achieve high accuracy on the test data without learning anything meaningful due to some peculiar artifacts of the data.

[27] 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 the top 200 most relevant words for each class in a given job ad for our analysis.

'software trainee', 'business development manager', and 'sales market executive'. We find that words representing personality, appearance, communication skills and basic computer proficiency have a high relevance for the F class. On the other hand, working in rotational shifts, field work and travel requirements have a high relevance for the M class.

To construct an overall gender association for each word, we first take the median relevance score for a word towards each class $\overline{R}_w^G$, $G \in \{F, N, M\}$. A positive median score for the female (male) class ($\overline{R}_w^G > 0$) indicates that the word w is associated with requesting a female (male). However, a word that is associated with a female as well as a male request may not contribute *differentially* towards either 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 male class. A positive (negative) net score for the word reflects that it contributes more towards female (male) requests in job ads.

Gender-category scores: 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 3,113 words that meet this criteria. These words constitute 92% of all word occurrences by volume in N jobs as well. We manually classify these words into four categories (C): hard-skills (280 words), soft-skills (63), personality/appearance (91), and flexibility (12). We assign words to 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, 'flexibility' captures words related to job timings and travel requirements. The remaining words (including words that occur less than ten times in M and F jobs) 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 3,113 words and their median relevance scores $\overline{R}_w^G$. We sum the median relevance scores for all words within job ad i towards

the F class which are also assigned to a given category C, $S^C_{i,F} = \sum_{(w \in i) \wedge (w \in C)} \overline{R}^F_w$. Similarly we sum the median relevance scores for words towards the M class which are also assigned to a given category C, $S^C_{i,M} = \sum_{(w \in i) \wedge (w \in C)} \overline{R}^M_w$. We then take the difference between the two sums to arrive at a net score towards the F class ($NS^C_i = S^C_{i,F} - S^C_{i,M}$) in each category. A positive net score for a category could indicate either that the job ad contains *more words* that contribute towards a gender request for a female vs a male, or that the words have a *higher median relevance* for the female class vs the male class.

Estimation Strategy: We next examine which categories of words matter for the relationship between implicit gender association and the advertised wage, as well as between implicit gender association and the female applicant share. Rather than using individual words, we use aggregate category scores to derive meaningful interpretations. Specifically, we use net scores NS^C_i for each category to construct gender-category variables as below:

$$FW^C_i = \begin{cases} NS^C_i, & \text{if } NS^C_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$MW^C_i = \begin{cases} -NS^C_i, & \text{if } NS^C_i < 0 \\ 0, & \text{otherwise} \end{cases}$$

This gives us two separate variables for each category. For example, $FW^{hard-skills}$ or 'Female (hard-skills)' takes on the value of the net score for hard-skills when this score is positive and zero otherwise. This variable will be positive if words classified under the category of hard-skills in a job ad contribute more towards F than M jobs. Similarly, $MW^{hard-skills}$ or 'Male (hard-skills)' takes on the absolute value of net score for hard-skills when the score is negative and zero otherwise. This variable will be positive if words classified under the category of hard-skills in a given job ad contribute more towards M than F jobs. If a job ad does not have a word in a given category then it gets a zero score for both gender-categories. We use this procedure to construct ten gender-category variables—two for each of the five categories (including "others"). The gender-category variables are then standardized for ease of interpretation. We report the summary statistics for the

non-standardized gender-category variables in Table A.5 separately for F, N, and M jobs. The gender-category variables reflecting an implicit preference for women have the highest scores in F jobs; for instance, 'Female (hard-skills)' gets an average score of 0.17, 0.11 and 0.07 in F, N, and M jobs, respectively. However, gender-category variables reflecting a male preference do not always have the highest scores in M jobs. For instance, 'Male (hard-skills)' has the highest average score in N jobs ($= 0.16$), and then in M jobs ($= 0.12$). Nevertheless, male-category scores are consistently higher in M jobs than in F jobs.

To examine the association between the gender-category variables and the advertised wage, we estimate the following Mincer regressions separately for F, N, and M jobs:

$$\ln W_{ijst} = \rho^W + \sum_C \delta^{FW,C} FW^C_{ijst} + \sum_C \delta^{MW,C} MW^C_{ijst} + \tau^W X_{ijst} + \gamma_{j\times s} + \phi_t + \zeta_{ijst} \tag{4.1}$$

$\ln W_{ijst}$ is the log wage advertised in job ad i. In contrast to equation (3.1), the explanatory variables include standardized female-category (FW^C_i) and male-category (MW^C_i) variables rather than implicit *femaleness* and *maleness*. X_{ijst} is a set of dummy variables for education and experience requirements. The coefficients of interest, $\delta^{FW,C}$ and $\delta^{MW,C}$, give the log points change in wage for a standard deviation increase in the female and male-category scores respectively, everything else equal. We control for occupation and state fixed effects as well as time fixed effects and cluster standard errors by occupation and state.

To further examine the association between gender-category variables and the share of female applicants to a job we estimate the following regressions separately for F, N and M jobs:

$$Y^S_{ijst} = \rho^S + \sum_C \eta^{FW,C} FW^C_{ijst} + \sum_C \eta^{MW,C} MW^C_{ijst} + \tau^S X_{ijst} + \gamma_{j\times s} + \phi_t + \varsigma_{ijst} \tag{4.2}$$

Y^S_{ijst} is the share of female applicants to job ad i. This specification is similar to equation (4.1) except that the regressions in equation (4.2) are weighted by the total number of male and female applications made to a job ad. The coefficients of interest, $\eta^{FW,C}$ and $\eta^{MW,C}$, give the percentage point change in the female applicant share for a standard deviation increase in female and male-category scores respectively, everything else equal.

4.2 Results

4.2.1 Gendered words

To obtain the most relevant words for each gender-category we sort words based on their net scores within each category. The top words in our ordered list contribute relatively more towards female requests while the bottom words contribute relatively more towards male requests. We list a maximum of 20 words that are most highly associated with requests for women and men within each category in Table 3. 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)–(II), Panel A), words associated with a beautician (*facial, pedicure, manicure, makeup*), accounting tasks and software (*ledger, expense statements, tally*), 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, words 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)–(IV), Panel A) and again find a stark distinction 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, persuading, supervising, and motivating.

The gender contrast is particularly evident in different personality traits across jobs that request men and women (Columns (I)–(II), Panel B). Jobs requesting women require the applicant to be pleasant, presentable, confident, mature, careful, include physical traits such as height, and other characteristics such as politeness, patience, adaptability, and punctuality. At the same time, some contrasting words like being pro-active and entrepreneurial are also present. On the other hand, personality traits such as being 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 *skype* calls and possibility of work from *home* or *home* based work are associated with jobs requesting a

female (Column (III), Panel B).[28] On the other hand, night/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 soft and hard-skills, personality traits and flexibility related words are associated with jobs that request men and women.

4.2.2 Gendered words and the advertised wage

Table 4 reports the results from estimation of equation (4.1). Estimates for N jobs show that a standard deviation increase in 'Female (hard-skills)' decreases the advertised wage significantly by 3 log points, while an increase in 'Female (soft-skills)' and 'Female (personality)' increases the advertised wage by 0.6 and 3.2 log points (Column (III)). A standard deviation increase in 'Male (hard-skills)', 'Male (soft-skills)', 'Male (personality)' and 'Male (flexibility)' all increase the advertised wage by 1.8, 2, 1.3 and 3.3 log points respectively (Column (III), Table 4). Using within occupation and state variation only in Column (IV) of Table 4, we find that the negative effect of 'Female (hard-skills)' on the advertised wage as well as the positive effect of 'Female (soft-skills)', 'Female (personality)', and all male gender-category variables persists, albeit the magnitudes decline. One standard deviation increase in 'Male (flexibility)' is associated with the highest increase in wages, by 2.1 log points, while a similar increase in 'Female (hard-skills)' is associated with the largest decline in the advertised wage, by 1.3 log points.

The results for F jobs in Table 4, Column (II), show that a standard deviation increase in 'Female (hard-skills)' decreases the wage significantly by 2.1 log points using within occupation and state variation only. On the other hand, a standard deviation increase in 'Male (flexibility)' is associated with a significant increase in the wage, by 4.6 log points. Thus, if employers want a female for a position, they are willing to pay an even higher wage premium for jobs that require longer working hours, travel, relocation or night shifts. Lastly, the results for M jobs in Table 4, Column (VI), show that no category of female gendered words matters for advertised wages using within occupation and state variation. Only the presence of words related to 'Male (flexibility)' are associated with an increase in the advertised wage.

These results show that 'Female (hard-skills)' and 'Male (flexibility)' are the most important

[28] 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. We also check the robustness of our results to using context specific word scores for each job ad (Section 5).

correlates of advertised wages when we use within occupation and state variation only. Gendered words related to 'Female (hard-skills)' are associated with lower advertised wages, while those related to 'Male (flexibility)' and 'Male (soft-skills)' are associated with higher advertised wages.[29] There is also a wage premium for 'Female (soft-skills)' and 'Female (personality)', but these are significant only for N jobs. These results align with stereotypical female skills getting penalized in the labor market and also indicate a trade-off between job flexibility and wages.

4.2.3 Gendered words and the female applicant share

Table 5 reports the results from estimation of equation 4.2. The results for N jobs show that a standard deviation increase in 'Female (hard-skills)' and 'Male (hard-skills)' increases the fraction of female applicants by 0.9 percentage points (pp) or 2.8% of mean applicant share to N jobs and 1 pp (3.1% of mean) respectively (Column (III)). An increase in 'Male (personality)' does not deter women from applying. On the other hand, gendered words related to 'Male (soft-skills)' and 'Male (flexibility)' reduce the female applicant share by 0.5 pp (1.6% of mean) and 0.3 pp (0.9% of mean). Using within occupation and state variation, we find that only the positive effect of 'Female (hard-skills)' and the negative effect of 'Male (flexibility)' on the female applicant share persists, and both are almost equal at 0.4 pp or 1.3% of mean (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 (Columns (I)–(II)). A standard deviation increase in net scores within the categories of 'Male (hard-skills)', 'Male (soft-skills)' and 'Male (flexibility)' reduce the female applicant share by 6 pp (11.5% of mean applicant share to F jobs), 1.9 pp (3.7% of mean) and 2.6 pp (8.1% of mean) respectively—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% of mean) in Column (II). For M jobs, none of the gendered word categories significantly matter for the female applicant share (Column (VI)) within an occupation and state. An increase in net scores of words related to 'Male (flexibility)' decreases female applicant share by a similar magnitude as for N jobs, but this effect is imprecisely estimated.

[29] In additional analysis, we find that the negative association of 'Female (hard-skills)' is not driven by beautician related words, i.e. we continue to find the negative association even after we exclude beautician related words when constructing the net score for hard-skills. These results are available on request.

Our findings show that the proportion of female applicants increases when the job text uses female gendered words—especially those related to hard-skills; but these words are associated with a negative wage premium. This contributes to the gender earnings gap in the labor market. At the same time, the female applicant share decreases when the job text uses male gendered words related to flexibility (that largely reflect greater travel requirements, working on weekends or night shifts), and these words are associated with a positive wage premium in the labor market. This further contributes to the gender earnings gap, and is consistent with job attributes within an occupation—particularly those related to flexibility—being among the main drivers of gender wage gaps (Goldin, 2014). Indeed, as much as 20 percent of the gender wage gap in applications on the portal is driven by a larger proportion of women applying to jobs with higher female gendered words and lower male gendered words in job ads.[30]

5 Robustness checks

We examine the robustness of our results to several 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 2.3; we find that our results are largely robust. 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.6). 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.7). 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.2(a)). Lastly, our results related to employer's gendered word use in job ads and its consequences also continue to hold; we still find a decrease in the advertised wage and an increase in female applicant share with 'Female

[30]Female applicants apply to jobs that offer 2 log points lower annual wages after including controls for candidate education, age, occupation × location and month-year of job posting. This gap falls to 1.6 log points after accounting for the extent of gendered words present in the job ad by controlling for their standardized scores in the above regression. Thus, approximately, the gender wage gap in applications reduces by 20 percent.

(hard-skills)' as well as an increase in the advertised wage and a decrease in female applicant share with 'Male (flexibility)' for N jobs (Columns (I) and (IV), Appendix Table A.8).

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.[31] We continue to find that our results are largely robust. We still find 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.6). 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.7). 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.2(b)). However, when using firm × occupation × state fixed effects the confidence intervals on the predicted shares become quite wide (Appendix Figure A.2(c)). Lastly, our results on the positive (negative) effect of male (female) gendered words relating to job flexibility (hard-skills) on wages and the negative (positive) effect of these words on female applicant share for N jobs are largely robust (Columns (II)–(III) and (V)–(VI), Appendix Table A.8). We see a significant increase in the female applicant share when gendered words indicating greater flexibility occur in job ads posted by the same firm for the same occupation. However, the coefficients on female gendered words related to hard-skills are now insignificant, albeit still positive.

Applicant characteristics: We also estimate an alternative specification to regressions where the dependent variable is the share of female applicants in which we also control for applicant characteristics. We do this by estimating regressions at the application rather than job ad level, where the dependent variable takes the value one if an applicant to a job ad is female and zero if it is a male. Using these regressions, we are able to control for applicant characteristics such as

[31] In Appendix Tables A.6–A.8, 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.

the applicant's highest education level and a quadratic in applicant's age. We continue to control for job characteristics, occupation times state, and month-year fixed effects. We find statistically significant effects of an employer's explicit gender preference on the probability that a female applies (Appendix Table A.4). Similarly, we estimate the responsiveness of whether a female applicant applies to the job text being predictive of a gender preference; we find that our previous results continue to hold (results available on request). Finally, we estimate the effect of gender-category scores on the probability that a female applies to a job and find that our previous results on hard-skills and job flexibility persist (Appendix Table A.9).

Contextual gender-category scores: We also check the robustness of our results to an alternative way of constructing the gender-category scores. Rather than taking the median score for each word, we take the score associated with the word in a job ad (given the context in which the word appears in the job text). We find that our previous results on 'Female (hard-skills)' and 'Male (flexibility)', if anything, become stronger. On average, female gendered words are associated with a decline in the advertised wage and an increase in female applicant share. On the other hand, male gendered words—especially related to flexibility—are associated with an increase in the advertised wage and a decrease in the female applicant share (Appendix Table A.10).

6 Words that attract a higher fraction of women

In Section 4, we examined the impact of *gendered words* (or words that are predictive of an explicit gender preference by an employer) on the share of female applicants. In this Section we examine all words in job ads (not just *gendered words*) which attract a higher fraction of female applicants. We find a high correlation across the two lists in the categories of 'Hard-skills' (Spearman's rank correlation = 0.23) and 'Flexibility' (Spearman's rank correlation = 0.50) but a low or negative correlation in the categories of 'Soft-skills' (Spearman's rank correlation = 0.03) and 'Personality/Appearance' (Spearman's rank correlation = −0.12). This is consistent with our previous results, and indicates a higher correspondence between employer stereotypes and applicant responses for 'Hard-skills' and 'Flexibility' than for 'Soft-skills' or 'Personality/Appearance'. We discuss the construction of the new list as well as the words which appear in this list below.

We use N jobs to first estimate the part of applicant share variation within a given occupation-location which is not due to job characteristics. We do this by regressing the female applicant share on job characteristics (education and experience requirements, month-year of posting) and occupation $\times$ state fixed effects.[32] We then predict the residual applicant share and use it as the dependent variable to estimate a ridge regression model using word unigrams (with TF-IDF scores) as features.[33] The model gives a coefficient for each word which can be interpreted as a marginal effect of the presence of that word on the female applicant share. Words with a positive marginal effect (or which increase the female applicant share) are included in the female list while those with a negative marginal effect are added to the male list. Table 6 displays the top 20 words further classified into each of the four categories ('Hard-skills', 'Soft-skills', 'Personality/Appearance', 'Flexibility'), with the marginal effect for the word in parentheses.[34]

Within the category of 'Hard-skills' (Column (I), Panel A, Table 6), words related to beauty service, accounting, and architectural skills still appear among words attracting a larger share of female applicants. In addition, we also find words related to legal professions, software and database management, automation, and content creation in this list. Within this category, words that attract the highest fraction of male candidates continue to be dominated by words related to engineering, analytics and quantitative skills such as *python, machine learning, robotics, plc, server, desktop, configuration, network management, es, ui*, and *seo* (Column (II), Panel A). Within the category of 'Soft-skills', female applicant shares increase with words related to communication skills such as coordination, counseling, and managing customer relations (Column (III), Panel A), while words related to team-work and collaboration, negotiation, and supervision still dominate for attracting a larger share of male applicants (Column (IV), Panel A). Within the category of 'Personality' (Columns (I)–(II), Panel B), there are several deviations from the list of gendered words. Female applicant share increases with words reflecting *determination*, being *pro-active*, willing to go to the last *mile*, *ethical, creative, thinker,* taking *initiative*, and being *motivated* appear in the job ad. In

[32] This regression is weighted by the total number of applicants to a job.

[33] 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 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.

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

contrast, from the employers' perspective, gendered words in this category included appearance-related words as well as words such as *patience*, being *careful* and *punctual* (Table 3). This is consistent with 'Female-personality' scores in job ads not having an impact on female applicant shares (Table 5). Similarly, we find little overlap between personality-related words that attract a lower share of female applications and that are predictive of an employer's male preference. Lastly, we examine 'Flexibility' related words (Column (III)–(IV), Panel B). For women, we see that the most important words are again those related to being able to take skype calls and working on weekday which increase the female share of applicants by approximately 2.5 percentage points, while words that reflect job characteristics involving night shift and travel decrease the female applicant share by 10 and 5 percentage points respectively, which are very large effects.

7 Conclusion

Our results bear significant relevance, given the low female labor force participation rates in India, and the absence of effective legal bans on gender requests in job ads (unlike the US or China). They indicate that placing restrictions on gender targeting in job ads can increase the share of job applications by women towards relatively high skill and remunerative jobs, thus reducing the gender wage gap at the application stage. Further, we use explicit gender preferences to derive implicit employer gender associations. We show that job ads with higher female association use words in the job text that are associated with gender stereotypes in job attributes, offer lower wages, and see a higher fraction of female applications. Importantly, our analyses show that among the category of gendered words that contribute to these implicit associations, those related to female hard-skills and (reduced) job flexibility increase and decrease the female applicant share while being associated with a lower and higher wage respectively. Thus, words contained in a job ad matter for job search.

These results have broader implications for the literature on gender wage gaps and the labor market. Recent evidence (primarily from developed countries) shows women's willingness to pay for flexible working hours and documents that sorting of workers across jobs with differential flexibility requirements can generate a gender wage gap. In addition to finding evidence for this using high skill jobs in a developing country setting with large gender disparities in labor market outcomes, we also find that skills required for jobs within (narrowly-defined) occupations may matter for the

gender wage gap. Thus, both skills and job flexibility related attributes matter during the search stage, and therefore may have implications on final sorting into jobs by gender. Given the lack of matched employer-employee data in most developing country settings, we show how applications data from job portals may be particularly useful to analyse job search behaviour. Lastly, these results using data from primarily entry-level job ads are striking. We show that job attributes matter at a stage when young people are entering the labor market. Extant literature shows that early career shocks—such as recessions—have important cumulative consequences for future labor market returns (Kahn, 2010; Oreopoulos et al., 2012; Oyer, 2006). In fact, job attributes that matter early on are likely to matter for future job switches too and may even affect returns to experience which are likely to depend on the initial pay.

References

Abraham, L. and A. Stein (2020): "Words Matter: Experimental Evidence from Job Applications," Unpublished manuscript.

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.

Bertrand, M. (2020): "Gender in the Twenty-First Century," *American Economic Review Papers and Proceedings*, 110(5), 1–24.

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.

Cortes, P., J. Pan, L. Pilossoph, and B. Zafar (2021): "Gender differences in job search and the earnings gap: Evidence from business majors," NBER working paper 28820.

Deming, D. and L. B. Kahn (2018): "Skill requirements across firms and labor markets: Evidence from job postings for professionals," *Journal of Labor Economics*, 36, S337–S369.

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.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019): "The allocation of talent and U.S. economic growth," *Econometrica*, 87, 1439–1474.

Kahn, L. B. (2010): "The long-term labor market consequences of graduating from college in a bad economy," *Labour economics*, 17, 303–316.

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.

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).

Morchio, I. and C. Moser (2020): "The Gender Pay Gap: Micro Sources and Macro Consequences," Unpublished manuscript.

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.

Oreopoulos, P., T. Von Wachter, and A. Heisz (2012): "The short-and long-term career effects of graduating in a recession," *American Economic Journal: Applied Economics*, 4, 1–29.

Oyer, P. (2006): "Initial labor market conditions and long-term outcomes for economists," *Journal of Economic Perspectives*, 20, 143–160.

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.

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.

Wiswall, M. and B. Zafar (2018): "Preference for the workplace, investment in human capital, and gender," *The Quarterly Journal of Economics*, 133, 457–507.

Xiao, P. (2020): "Wage and Employment Discrimination by Gender in Labor Market Equilibrium," Unpublished manuscript.

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: Advertised wages

Sample:	*F* jobs		*N* jobs		*M* jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
Femaleness	–0.185***	–0.202***	–0.379***	–0.264***	–0.320***	–0.192***
	(0.052)	(0.039)	(0.023)	(0.017)	(0.069)	(0.069)
Maleness	–0.107	–0.085	–0.123***	–0.136***	–0.116*	–0.151***
	(0.064)	(0.062)	(0.019)	(0.013)	(0.052)	(0.045)
Education requirements:						
Senior secondary	0.058*	0.045**	0.068***	0.043***	0.094***	0.013
	(0.028)	(0.020)	(0.009)	(0.007)	(0.036)	(0.024)
Diploma	0.117***	0.101***	–0.020	0.026***	0.056	0.096**
	(0.035)	(0.029)	(0.014)	(0.008)	(0.050)	(0.038)
Undergrad degree, STEM	0.112	0.132	0.156***	0.173***	0.116	0.115**
	(0.068)	(0.078)	(0.018)	(0.012)	(0.080)	(0.050)
Undergrad degree, non-STEM	0.095***	0.104***	0.046***	0.062***	0.127***	0.090***
	(0.027)	(0.023)	(0.012)	(0.007)	(0.038)	(0.025)
Postgrad degree, STEM	0.720	0.000	0.438***	0.352***	0.927	1.115**
	(0.507)	(0.205)	(0.055)	(0.048)	(0.507)	(0.455)
Postgrad degree, non-STEM	–0.047	0.089	0.241***	0.275***	–0.037	–0.088
	(0.115)	(0.076)	(0.036)	(0.032)	(0.112)	(0.055)
Experience requirements:						
$1-2$ years	0.101***	0.115***	0.066***	0.075***	0.110***	0.083***
	(0.019)	(0.015)	(0.009)	(0.007)	(0.026)	(0.022)
>2 years	0.248***	0.253***	0.319***	0.308***	0.290***	0.261***
	(0.024)	(0.021)	(0.013)	(0.011)	(0.037)	(0.031)
Fixed Effects	month	month, occ × state	month	month, occ × state	month	month, occ × state
Femaleness = Maleness, p-value	0.226	0.033	0.000	0.000	0.001	0.472
N	5727	5727	124654	124654	4795	4795

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. The omitted category among experience requirement categories is 0 to <1 year of experience. 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 on the portal which advertise a wage range, subject to the restrictions described in Section 2.1. All columns 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: Applications

Dependent variable:	total applications			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female preference	–20.686***	–8.079***	–5.455***	0.206***	0.156***	0.155***
	(2.654)	(0.821)	(0.803)	(0.014)	(0.006)	(0.007)
Male preference	–3.677	–0.996	–2.710	–0.133***	–0.099***	–0.095***
	(4.542)	(4.691)	(2.955)	(0.009)	(0.005)	(0.005)
Education requirements:						
Senior secondary	–0.547	2.428***	1.761**	0.047***	0.027***	0.028***
	(0.809)	(0.716)	(0.781)	(0.004)	(0.003)	(0.003)
Diploma	24.756***	3.766*	2.084	0.001	0.021***	0.023***
	(2.095)	(1.725)	(1.584)	(0.010)	(0.004)	(0.004)
Undergrad degree, STEM	108.789***	55.382***	49.773***	0.077***	0.047***	0.046***
	(14.747)	(7.445)	(6.806)	(0.013)	(0.004)	(0.004)
Undergrad degree, non-STEM	24.861***	11.162***	7.810***	0.125***	0.054***	0.055***
	(4.371)	(1.802)	(1.373)	(0.007)	(0.004)	(0.004)
Postgrad degree, STEM	6.882	1.425	–1.491	0.177***	0.112***	0.122***
	(5.124)	(7.273)	(15.745)	(0.011)	(0.013)	(0.016)
Postgrad degree, non-STEM	–3.627***	1.024	–9.934***	0.154***	0.079***	0.085***
	(1.305)	(2.391)	(2.482)	(0.020)	(0.011)	(0.014)
Experience requirements:						
$1-2$ years	–25.235***	–24.511***	–18.039***	–0.024***	–0.015***	–0.016***
	(4.116)	(3.626)	(2.408)	(0.004)	(0.002)	(0.003)
>2 years	–40.138***	–46.762***	–35.800***	–0.064***	–0.037***	–0.035***
	(5.757)	(6.829)	(4.331)	(0.005)	(0.003)	(0.003)
Advertised wage:						
ln(wage)			18.927***			–0.000
			(2.744)			(0.002)
Fixed Effects	month	month, occ × state	month, occ × state	month	month, occ × state	month, occ × state
N	157888	156221	136453	157888	156221	136453

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. The omitted category among experience requirement categories 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, 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 2.1. 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 3: Gendered words

(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		**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, Soft-skills, Personality/Appearance, 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. 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 4: Gendered words and the advertised wage

Sample:	*F* Jobs		*N* Jobs		*M* Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.037***	–0.021***	–0.030***	–0.013***	–0.033***	–0.018
	(0.006)	(0.006)	(0.003)	(0.002)	(0.010)	(0.009)
Female (soft-skills)	–0.006	–0.004	0.006**	0.004*	0.009	–0.002
	(0.005)	(0.004)	(0.002)	(0.002)	(0.009)	(0.009)
Female (personality)	0.014**	0.007	0.032***	0.009***	0.034***	0.006
	(0.005)	(0.005)	(0.003)	(0.002)	(0.007)	(0.005)
Female (flexibility)	0.008	–0.001	0.000	–0.000	–0.009	–0.007
	(0.007)	(0.007)	(0.002)	(0.002)	(0.008)	(0.008)
Female (others)	–0.022***	–0.018***	–0.043***	–0.021***	0.050***	0.032
	(0.006)	(0.006)	(0.004)	(0.003)	(0.018)	(0.021)
Male (hard-skills)	0.006	0.022	0.018***	0.016***	–0.019	0.011
	(0.022)	(0.020)	(0.004)	(0.003)	(0.010)	(0.010)
Male (soft skills)	0.026*	0.016	0.020***	0.014***	0.019***	0.021***
	(0.012)	(0.012)	(0.003)	(0.002)	(0.007)	(0.006)
Male (personality)	0.002	0.004	0.013***	0.010***	0.002	0.004
	(0.010)	(0.008)	(0.003)	(0.003)	(0.007)	(0.006)
Male (flexibility)	0.059***	0.046***	0.033***	0.021***	0.019***	0.015**
	(0.014)	(0.007)	(0.003)	(0.002)	(0.007)	(0.007)
Male (others)	0.003	0.014	0.017***	0.008**	0.035***	0.014***
	(0.019)	(0.019)	(0.004)	(0.004)	(0.005)	(0.005)
Fixed Effects	month	month, occ × state	month	month, occ × state	month	month, occ × state
N	5727	5727	124654	124654	4795	4795

Notes: The dependent variable is the log of the mid-point of the wage range advertised 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 includes other, illiterate, and secondary education. The omitted category among experience requirement categories is 0 to < 1 year of experience. 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 2.1. All columns 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 5: Gendered words and the 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 (softskills)	–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	4945	4945

Notes: The dependent variable is the fraction of female applicants to 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 includes other, illiterate, and secondary education. The omitted category among experience requirement categories is 0 to $<$ 1 year of experience. All regressions 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.

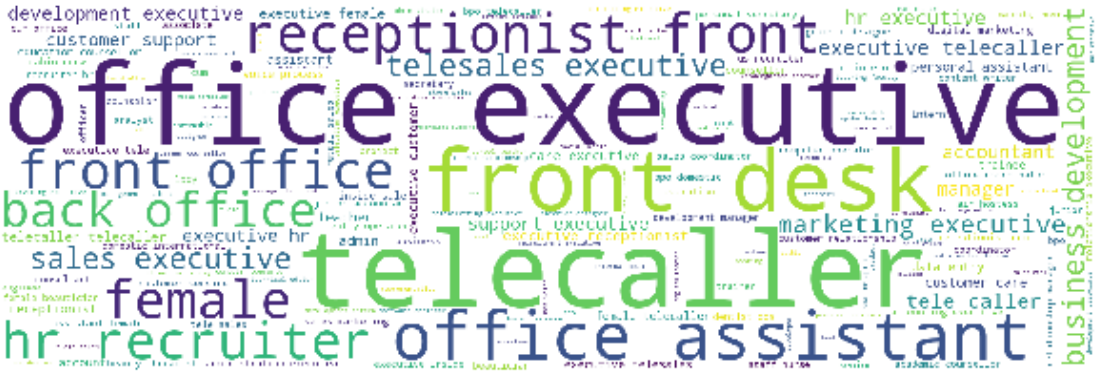
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2.1. All columns report the effective number of observations after incorporating occ × state fixed effects which exclude job ads for which there is no variation in the dependent variable within an occ × state cell.

Table 6: Words which affect the share of female applications

(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		**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, Soft-skills, Personality/Appearance, 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). We only keep words with a marginal effect exceeding one percentage point in the table.
Source: Data from the population of all job ads that do not specify a gender request and applicants on the portal, subject to the restrictions described in section 2.1.

Figure 1: Word clouds of job titles



(a) Female preference (F jobs)

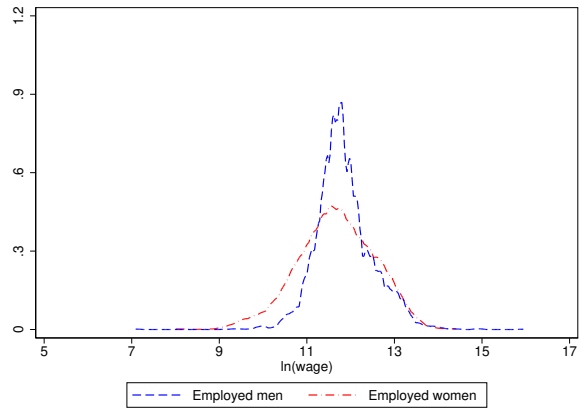
(b) Male preference (M jobs)

(c) No gender preference (N jobs)

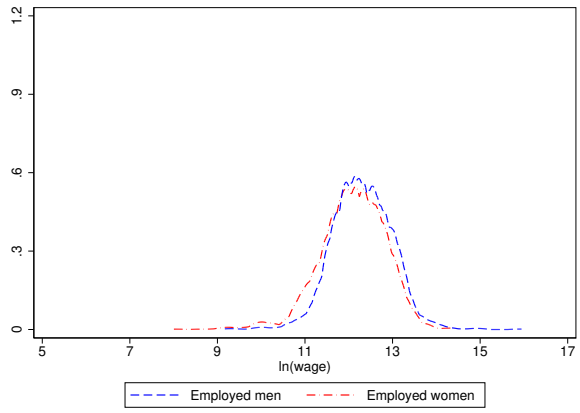
Notes: The word clouds are constructed based on words contained in job titles of ads displaying an explicit female preference, an explicit male preference and no explicit gender preference.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2.1. The final number of job ads is 157888.

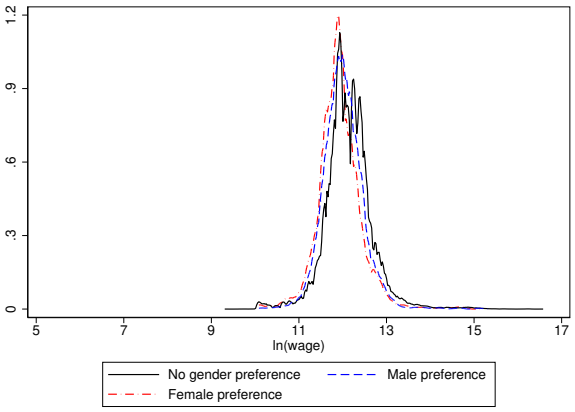
Figure 2: Wage distributions



(a) Wage distributions by gender, PLFS



(b) Wage distributions by gender (undergraduates or higher), PLFS

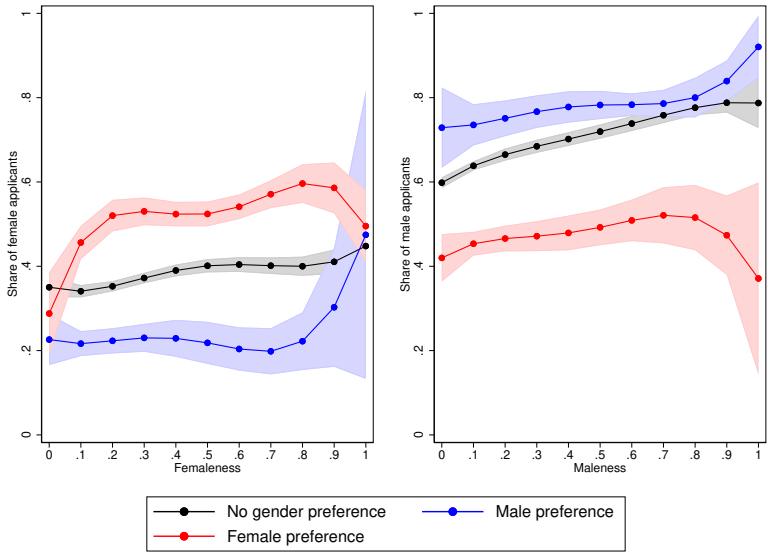


(c) Wage distributions by gender preference, job portal

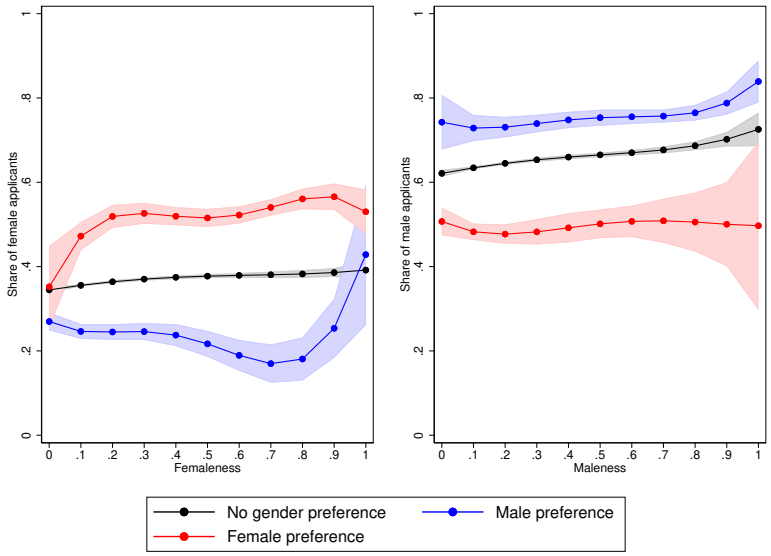
Notes: Distributions are the kernel density estimates. Figure (c) uses the mid-point of the posted wage range in job ads on the job portal.

Source: Figure (a) includes all urban workers while Figure (b) includes urban workers with an undergraduate or postgraduate degree who are 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 (c) includes data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2.1.

Figure 3: 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 (3.3), 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 2.1. The final number of job ads is 157888.

Figure 4: 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. The job ads are shown after removing stopwords and lemmatization. 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
Undergraduate degree, STEM	0.034	0.089	0.054	0.086
Undergraduate 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
Advertised wage:				
Wage not specified	0.021	0.134	0.033	0.126
Annual wage, if wage specified in job ad	177100	216807	183293	213648
N (jobs with advertised wage)	6413	126152	5407	137972
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 (INR). Wages and experience are the mid-point of the range specified in the job ad. Other job requirements are constructed from the job ad text, the details of which are provided in Appendix B.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2.1.

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
Undergraduate degree, STEM	0.535	0.545	0.541
Undergraduate 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	257177	256810	256939
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 applicants. Experience is given in years and is divided into three 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
Undergraduate 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
Undergraduate 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 employed or seeking work for majority of the year. This proportion is calculated for all individuals in the respective gender-age group.
Source: Periodic Labour Force Survey (PLFS) conducted in 2017-18.

Table A.4: Female applicant share, robustness checks (with controls for applicant characteristics)

	(I)	(II)	(III)
Female preference	0.204***	0.167***	0.166***
	(0.012)	(0.006)	(0.006)
Male preference	–0.118***	–0.090***	–0.092***
	(0.007)	(0.006)	(0.005)
Education requirements:			
Senior secondary	0.022***	0.010***	0.011***
	(0.004)	(0.002)	(0.002)
Diploma	–0.038***	–0.005	–0.004
	(0.007)	(0.004)	(0.004)
Undergraduate degree, STEM	0.012	0.010***	0.008
	(0.010)	(0.004)	(0.005)
Undergraduate degree, non-STEM	0.050***	0.020***	0.022***
	(0.005)	(0.003)	(0.003)
Postgraduate degree, STEM	0.070***	0.042***	0.049***
	(0.009)	(0.009)	(0.012)
Postgraduate degree, non-STEM	0.079***	0.045***	0.054***
	(0.020)	(0.013)	(0.015)
Experience requirements:			
$1-2$ years	–0.020***	–0.014***	–0.013***
	(0.004)	(0.002)	(0.003)
>2 years	–0.046***	–0.027***	–0.025***
	(0.005)	(0.003)	(0.003)
Other job requirements:			
Age requirement present	–0.029***	–0.010***	–0.008*
	(0.007)	(0.004)	(0.004)
Beauty requirement present	–0.005	–0.001	0.000
	(0.006)	(0.003)	(0.003)
Advertised wage:			
ln(wage)			–0.006***
			(0.002)
Fixed Effects	month	month, occ × state	month, occ × state
N	6401972	6401972	5332833

Notes: The dependent variable is a dummy variable that takes a value one if an applicant to a job ad is female 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 0 to <1 year of experience. All regressions control for education level, age and age squared of the applicant. 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 2.1.

Table A.5: Descriptive statistics: Gendered words

	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 as described in Section 4.1.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in section 2.1.

Table A.6: Advertised wages, robustness checks

	(I)	(II)	(III)
Femaleness	–0.225***	–0.283***	–0.127***
	(0.013)	(0.019)	(0.018)
Maleness	–0.105***	–0.076***	–0.095***
	(0.012)	(0.017)	(0.019)
Education requirements:			
Senior secondary	0.034***	–0.018	–0.025**
	(0.006)	(0.013)	(0.010)
Diploma	0.017*	0.038**	0.008
	(0.008)	(0.016)	(0.018)
Undergrad degree, STEM	0.145***	0.143***	0.107***
	(0.011)	(0.028)	(0.019)
Undergrad degree, non-STEM	0.052***	0.019	–0.003
	(0.006)	(0.011)	(0.010)
Postgrad degree, STEM	0.360***	0.177***	0.141*
	(0.044)	(0.065)	(0.070)
Postgrad degree, non-STEM	0.216***	0.207***	0.254***
	(0.034)	(0.050)	(0.074)
Experience requirements:			
$1-2$ years	0.065***	0.045**	0.013
	(0.005)	(0.019)	(0.011)
>2 years	0.289***	0.261***	0.179***
	(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	0.152
N	121931	74729	42059

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. The omitted category among experience requirement categories is 0 to <1 year of experience. Standard errors are clustered at the (state, occupation) level (column (I)), the (state, firm) level (column (II)), or the (state, occupation, 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 2.1. 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: Applications, robustness checks

Dependent variable:	total applications			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female preference	–6.291***	–8.499***	–4.105***	0.150***	0.195***	0.139***
	(0.690)	(0.926)	(0.920)	(0.006)	(0.010)	(0.010)
Male preference	1.235	–7.468***	1.163	–0.087***	–0.120***	–0.091***
	(3.720)	(2.702)	(3.827)	(0.005)	(0.009)	(0.009)
Education requirements:						
Senior secondary	2.055***	–0.232	1.697**	0.025***	0.023***	0.016***
	(0.732)	(0.883)	(0.684)	(0.002)	(0.004)	(0.004)
Diploma	1.811	12.522***	4.384***	0.021***	–0.003	0.028***
	(1.559)	(1.615)	(1.596)	(0.003)	(0.010)	(0.006)
Undergrad degree, STEM	42.619***	35.182***	14.816***	0.041***	0.041***	0.053***
	(5.295)	(3.817)	(3.054)	(0.003)	(0.013)	(0.006)
Undergrad degree, non-STEM	7.658***	2.792*	1.638	0.048***	0.082***	0.056***
	(1.351)	(1.276)	(0.877)	(0.003)	(0.009)	(0.005)
Postgrad degree, STEM	–3.611	1.878	–9.664	0.107***	0.122***	0.115***
	(7.690)	(7.198)	(16.131)	(0.018)	(0.023)	(0.027)
Postgrad degree, non-STEM	–4.209	–3.667	–2.379	0.081***	0.111***	0.069***
	(2.371)	(7.313)	(4.219)	(0.017)	(0.015)	(0.014)
Experience requirements:						
$1-2$ years	–23.301***	–10.626***	–10.978***	–0.012***	–0.015***	–0.008***
	(3.284)	(1.642)	(1.347)	(0.002)	(0.004)	(0.003)
>2 years	–42.704***	–19.196***	–20.303***	–0.033***	–0.043***	–0.029***
	(5.208)	(2.726)	(1.428)	(0.002)	(0.007)	(0.003)
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. The omitted category among experience requirement categories 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, occupation) level (columns (I) and (IV)), the (state, firm) level (columns (II) and (V)), or the (state, occupation, 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 2.1. 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.8: Gendered words, advertised wages and applicant behaviors, robustness checks

Dependent variable:	log of advertised wage			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.008***	–0.011***	–0.006*	0.002***	0.007***	0.002
	(0.002)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)
Female (soft-skills)	0.003	–0.002	0.003	0.000	0.001	0.000
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Female (personality)	0.008***	0.004**	–0.000	0.001	0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Female (flexibility)	–0.001	–0.003	0.002	–0.000	0.000	0.002*
	(0.002)	(0.004)	(0.003)	(0.000)	(0.001)	(0.001)
Female (others)	–0.012***	–0.040***	–0.017***	0.006***	0.021***	0.007***
	(0.003)	(0.003)	(0.003)	(0.001)	(0.003)	(0.002)
Male (hard-skills)	0.016***	0.018***	0.005	0.000	0.002***	–0.000
	(0.002)	(0.003)	(0.004)	(0.001)	(0.001)	(0.001)
Male (soft skills)	0.011***	0.011***	0.007**	–0.001	–0.003	–0.001
	(0.002)	(0.003)	(0.003)	(0.001)	(0.002)	(0.001)
Male (personality)	0.009***	–0.001	–0.004	0.000	–0.003***	–0.001
	(0.002)	(0.002)	(0.003)	(0.000)	(0.001)	(0.001)
Male (flexibility)	0.020***	0.009***	0.006***	–0.003***	–0.007***	–0.003***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Male (others)	0.015***	0.005*	–0.002	–0.007***	–0.023***	–0.007***
	(0.003)	(0.002)	(0.003)	(0.001)	(0.003)	(0.001)
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	122163	74913	42141	140763	93930	57427

Notes: The dependent variable in columns (I)-(III) is the log of the mid-point of the wage range advertised in a job ad and in columns (IV)-(VI) is the fraction of female applicants. All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories includes other, illiterate, and secondary education. The omitted category among experience requirement categories 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, occupation) level (columns (I) and (IV)), the (state, firm) level (columns (II) and (V)), or the (state, occupation, 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 without an explicit gender preference and which advertise a wage, subject to the restrictions described in Section 2.1. 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.9: Gendered words and the female applicant share, robustness checks (with controls for applicant characteristics)

Sample:	F Jobs		N Jobs		M Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female (hard-skills)	–0.004	–0.003	0.009***	0.003***	0.002	–0.000
	(0.005)	(0.004)	(0.002)	(0.001)	(0.004)	(0.004)
Female (softskills)	–0.001	–0.003	0.003*	–0.000	0.006*	0.003
	(0.003)	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)
Female (personality)	0.001	0.002	–0.001	0.001	0.005	–0.000
	(0.003)	(0.002)	(0.001)	(0.001)	(0.003)	(0.002)
Female (flexibility)	0.000	–0.001	0.000	–0.001	–0.004	0.001
	(0.002)	(0.002)	(0.001)	(0.000)	(0.003)	(0.003)
Female (others)	0.006	–0.003	0.017***	0.008***	0.038***	0.015
	(0.003)	(0.002)	(0.003)	(0.001)	(0.010)	(0.008)
Male (hard-skills)	–0.073***	–0.025*	0.009***	–0.001	0.011**	–0.001
	(0.015)	(0.012)	(0.002)	(0.001)	(0.004)	(0.002)
Male (soft-skills)	–0.019**	–0.008	–0.005***	–0.001	–0.003	–0.001
	(0.008)	(0.005)	(0.001)	(0.001)	(0.004)	(0.002)
Male (personality)	0.005	0.003	0.003*	0.000	0.007	0.002
	(0.007)	(0.005)	(0.001)	(0.001)	(0.004)	(0.003)
Male (flexibility)	–0.027***	–0.022***	–0.003*	–0.004***	0.005	–0.007*
	(0.006)	(0.005)	(0.002)	(0.001)	(0.003)	(0.003)
Male (others)	–0.084***	–0.048***	–0.024***	–0.008***	–0.010***	–0.004***
	(0.020)	(0.013)	(0.002)	(0.001)	(0.002)	(0.001)
Fixed Effects	month	month, occ × state	month	month, occ × state	month	month, occ × state
N	112869	112869	6115984	6115984	173186	173186

Notes: The dependent variable is a dummy variable that takes a value one if an applicant to a job ad is female and zero otherwise. All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories includes other, illiterate, and secondary education. The omitted category among experience requirement categories is 0 to <1 year of experience. All regressions control for education level, age and age squared of the applicant. 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 2.1. All columns 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.

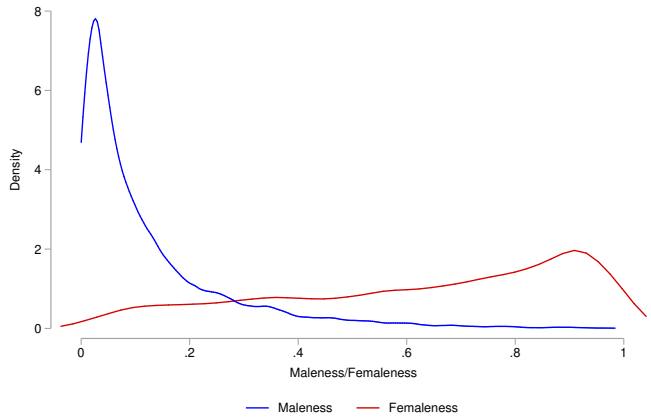
Table A.10: Gendered words, advertised wages and applicant behaviors, robustness checks (contextual scores)

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)	(0.006)	(0.002)	(0.001)	(0.003)
Female (personality)	0.005	0.005***	–0.001	0.001	0.001	0.002
	(0.005)	(0.002)	(0.005)	(0.002)	(0.001)	(0.003)
Female (flexibility)	0.003	0.003	0.002	–0.000	0.001	0.001
	(0.008)	(0.002)	(0.005)	(0.004)	(0.001)	(0.001)
Female (others)	–0.019***	–0.027***	0.006	–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***	0.005	–0.013**	–0.001	0.003
	(0.012)	(0.002)	(0.007)	(0.005)	(0.001)	(0.001)
Male (soft-skills)	–0.003	0.011***	0.010	0.001	0.001	0.001
	(0.005)	(0.002)	(0.010)	(0.003)	(0.001)	(0.004)
Male (personality)	0.003	0.005***	–0.002	0.000	–0.001	–0.005**
	(0.006)	(0.002)	(0.005)	(0.003)	(0.001)	(0.002)
Male (flexibility)	0.027***	0.018***	0.010*	–0.014***	–0.006***	–0.007**
	(0.007)	(0.002)	(0.005)	(0.004)	(0.001)	(0.003)
Male (others)	–0.000	–0.005	–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	5727	124654	4795	5839	144117	4945

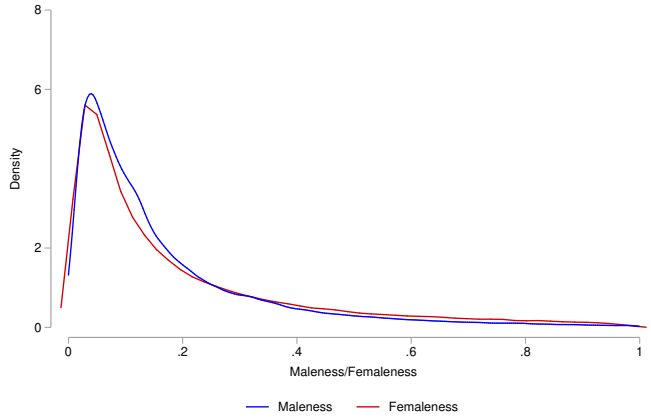
Notes: The dependent variable in columns (I)-(III) is the log of the mid-point of the wage range advertised in a job ad and in columns (IV)-(VI) is the fraction of female applicants. All regressions control for a set of education and experience requirement categories given in a job ad. The omitted category among education requirement categories includes other, illiterate, and secondary education. The omitted category among experience requirement categories 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, 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 2.1. All columns 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.

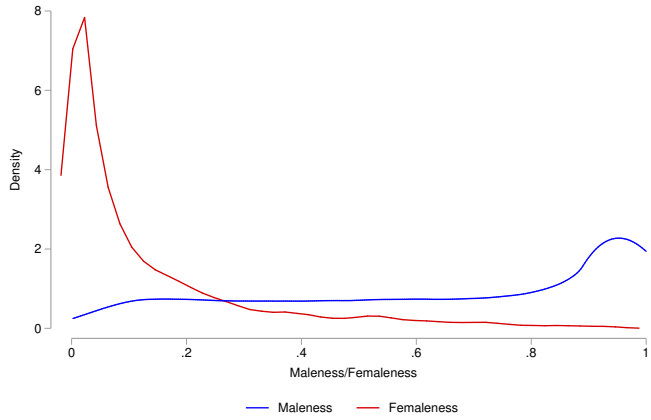
Figure A.1: Kernel Density Maleness and Femaleness



(a) F Jobs



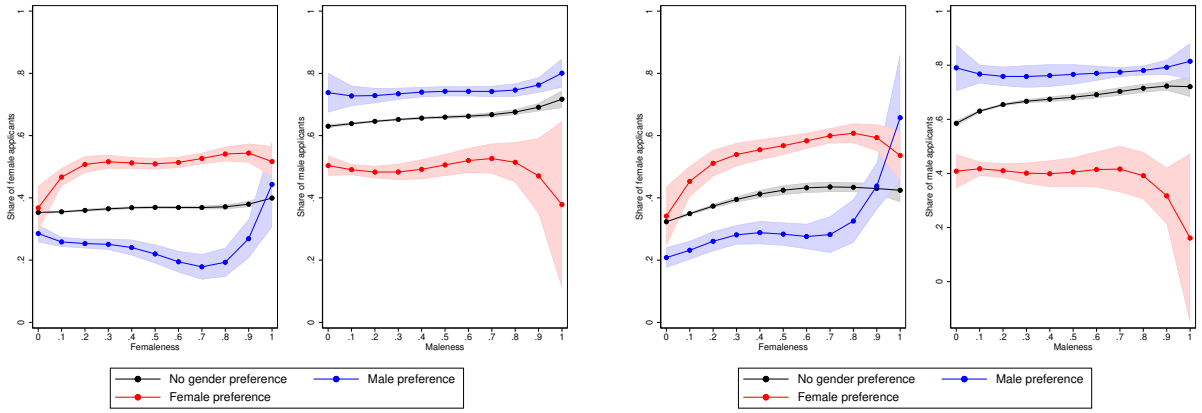
(b) N Jobs



(c) M Jobs

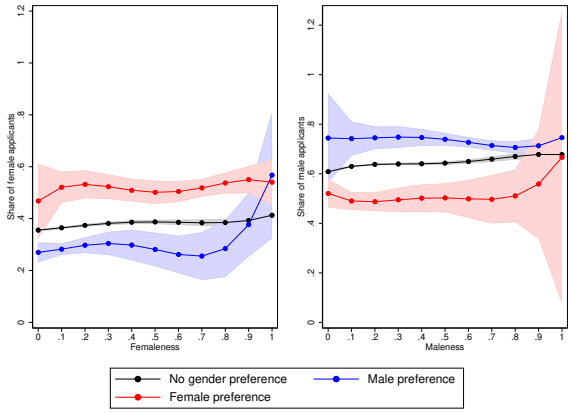
Notes: Distributions are the kernel density plots of estimated maleness and femaleness in F, N and M jobs.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions described in Section 2.1.

Figure A.2: Predicted share of female (male) applicants, robustness checks

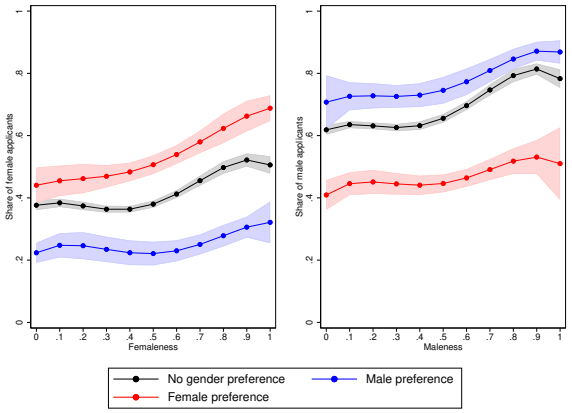


(a) Month and alternative occupation × state fixed effects

(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 (3.3), 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 2.1. The final number of job ads is 157888.

B Gender Requests in Job Ads

To detect the presence of an age requirement, we search the job text for the phrases 'years of age', 'years old', 'years to', 'age', or 'age limit' and also determine the minimum and maximum age requirements. We examine 25 characters before and after these phrases and search for numbers from 18 to 45 (since 45 is the maximum number found across all ads). If an ad has two numbers, the minimum of these is coded as the minimum age requirement and the maximum is taken as the maximum age requirement. In jobs where only one number appears, we check for words such as 'above', 'below', 'more than' and 'not above', 'not below', 'not less' to determine whether the age specified is the minimum or maximum required age.

We also create a dummy variable indicating the presence of a beauty requirement in a job ad by searching for the words 'height', 'weight', 'beautiful', 'charming', 'delightful', 'pretty', 'attractive' (ignoring cases specifying an attractive salary or package), 'good looking', 'nice looking', 'complexion', 'pleasing', 'appearance' and 'handsome' in the job text.[1]

We examine characteristics of jobs in which employers exhibit explicit gender preferences; the regressions we estimate are variations of the following specification:

$$Y_{ijst}^{k} = \alpha^{k} + \beta^{k} X_{ijst} + \gamma_{j\times s} + \phi_{t} + \epsilon_{ijst}^{k} \tag{B.1}$$

where $k \in \{FM, M\}$ indicates two different dependent variables capturing the *presence* and *direction* of explicit gender preferences. Y_{ijst}^{FM} is a binary outcome which takes the value 1 if there is an explicit male or female preference in job ad i advertising for a job of occupation j in state s and month-year t. The second dependent variable Y_{ijst}^{M} can take three values: -1 if there is an explicit female preference, 0 if there is no gender preference, and 1 if there is an explicit male preference.[2] X_{ijst} are job ad specific variables including dummy variables indicating education requirements, experience requirements, the presence of age and beauty requirements, and log ad-

[1] To find beauty-related words, we started with an initial list of words such as 'beautiful' and 'handsome'. We append this list by considering 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 were obtained by training the algorithm on web data from a 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, to identify synonyms.

[2] While we estimate and report linear regressions in this Appendix, we also estimate non-linear models (probit and ordered probit) with coarser job role and state fixed effects. Our results are largely unchanged; available on request.

vertised wage. Again, our preferred specification includes occupation and state fixed effects ($\gamma_{j\times s}$) as well as month-year fixed effects (ϕ_t). We use a detailed categorization of jobs to occupations with 483 distinct occupation categories derived from job titles as described in Section 2.3. The use of fixed effects ensures that 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.

Columns (I)–(III), Table B.1 give estimation results for equation (B.1) when the dependent variable is Y_{ijst}^{FM}. Column (I) includes all controls apart from the advertised wage as well as time (or month and year) fixed effects. Column (II) adds occupation × state fixed effects while Column (III) additionally controls for log advertised wage.[3] 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 results for experience.[4] We also find that the presence of an age or beauty requirement increases the probability of a job having an explicit gender preference (Columns (II) and (III), Table B.1).

Columns (IV)–(VI) in Table B.1 give results from estimation of equation (B.1) when the outcome of interest is male preference in a job ad (Y_{ijst}^{M}). We find that jobs with an explicit male preference are less likely to require a higher education; this effect becomes attenuated when we use within occupation-location variation only but still remains highly statistically significant. We also find that the presence of an age requirement leads to an increased preference for men, while the presence of a beauty requirement is associated with a reduced preference for men.[5] Jobs with an explicit male preference also offer higher wages than those with an explicit female preference; this is evident from our finding that a higher advertised wage is associated with an increased preference for men.

We check the robustness of the above findings to using the manual classification of occupations.

[3]Since wages are not posted for all jobs, we lose some observations when moving from Column (II) to (III).

[4]When occupation and state fixed effects are not included, jobs that specify a higher experience category (> 2 years relative to $0-1$ years) are less likely to exhibit a gender preference. However, after including occupation × state fixed effects and wage controls, higher experience requirement is associated with an *increased* probability of a job ad 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. We do not find the positive coefficients on higher experience requirements to be robust to the use of firm fixed effects (Columns (II) and (III), Appendix Table B.2).

[5]We also 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). For this, we estimate regressions on the sub-set of ads that specify a maximum required age and use the maximum required age instead of a dummy for the presence of age requirement as the explanatory variable of interest. While maximum required age has a positive association with preference for men, this effect is not statistically significant. These results are available on request.

We continue to find that explicit gender preferences are less likely in high skill jobs with a higher education requirement (Column (I), Appendix Table B.2). 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 B.2). Using either firm $\times$ state fixed effects or firm $\times$ occupation $\times$ state fixed effects, we continue to 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 B.2).

Table B.1: Explicit gender preferences

Dependent variable:	any gender preference			male preference		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Education requirements:						
Senior secondary	–0.0642***	–0.0273***	–0.0249***	–0.0709***	–0.0361***	–0.0376***
	(0.0104)	(0.0077)	(0.0078)	(0.0118)	(0.0080)	(0.0082)
Diploma	–0.0796***	–0.0299***	–0.0277***	–0.0569***	–0.0378***	–0.0405***
	(0.0129)	(0.0076)	(0.0077)	(0.0151)	(0.0079)	(0.0080)
Undergrad degree, STEM	–0.1014***	–0.0371***	–0.0261***	–0.0486***	–0.0338***	–0.0323***
	(0.0129)	(0.0074)	(0.0075)	(0.0153)	(0.0079)	(0.0080)
Undergrad degree, non-STEM	–0.0810***	–0.0325***	–0.0255***	–0.0745***	–0.0397***	–0.0415***
	(0.0127)	(0.0073)	(0.0075)	(0.0148)	(0.0080)	(0.0083)
Postgrad degree, STEM	–0.1148***	–0.0549***	–0.0454***	–0.0836***	–0.0338***	–0.0299*
	(0.0146)	(0.0093)	(0.0128)	(0.0168)	(0.0100)	(0.0142)
Postgrad degree, non-STEM	–0.0901***	–0.0403***	–0.0045	–0.0884***	–0.0366***	–0.0442**
	(0.0147)	(0.0107)	(0.0176)	(0.0169)	(0.0118)	(0.0194)
Experience requirements:						
$1-2$ years	0.0191***	0.0129***	0.0214***	–0.0006	–0.0017	–0.0023
	(0.0039)	(0.0025)	(0.0029)	(0.0041)	(0.0023)	(0.0028)
>2 years	–0.0111***	–0.0035	0.0125***	0.0090***	0.0043	0.0026
	(0.0025)	(0.0022)	(0.0030)	(0.0025)	(0.0023)	(0.0032)
Other job requirements:						
Age requirement present	0.0233	0.0501***	0.0675***	0.0579***	0.0381***	0.0446***
	(0.0122)	(0.0091)	(0.0107)	(0.0155)	(0.0073)	(0.0085)
Beauty requirement present	0.0295***	0.0286***	0.0280**	–0.0584***	–0.0550***	–0.0576***
	(0.0108)	(0.0106)	(0.0112)	(0.0072)	(0.0081)	(0.0084)
Advertised wage:						
ln(wage)			–0.0363***			0.0063*
			(0.0035)			(0.0032)
Fixed Effects	month	month, occ × state	month, occ × state	month	month, occ × state	month, occ × state
N	157888	156221	136453	157888	156221	136453

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. The omitted category among experience requirement categories is 0 to <1 year of experience. 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 on the portal, subject to the restrictions described in Section 2.1. 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 B.2: 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***	–0.022***	–0.020***	–0.068***	–0.024***
	(0.004)	(0.006)	(0.006)	(0.004)	(0.006)	(0.007)
Diploma	–0.012**	–0.072***	–0.019**	–0.022***	–0.065***	–0.027***
	(0.005)	(0.009)	(0.008)	(0.006)	(0.007)	(0.008)
Undergrad 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)
Undergrad degree, non-STEM	–0.013***	–0.075***	–0.024***	–0.022***	–0.083***	–0.031***
	(0.004)	(0.009)	(0.007)	(0.005)	(0.007)	(0.008)
Postgrad degree, STEM	–0.030***	–0.081***	–0.013	–0.030***	–0.076***	–0.037***
	(0.009)	(0.010)	(0.010)	(0.009)	(0.009)	(0.012)
Postgrad degree, non-STEM	–0.026**	–0.067***	–0.014	–0.014	–0.084***	–0.043***
	(0.010)	(0.008)	(0.010)	(0.011)	(0.009)	(0.010)
Experience requirements:						
$1-2$ years	0.012***	0.012*	0.006	–0.002	–0.008**	–0.007
	(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	0.030***	0.008	–0.004	–0.048***	–0.038***	–0.041***
	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
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. The omitted category among experience requirement categories is 0 to <1 year of experience. Standard errors are clustered at the (state, occupation) level (columns (I) and (IV)), the (state, firm) level (columns (II) and (V)), or the (state, occupation, 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 2.1. 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.

C Technical Appendix

C.1 GSDMM: Pre-processing and hyperparameter choice

Prior to implementing GSDMM, 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. This leaves us with $D = 28,957$ documents and $V = 3,127$ unique words.

Tokenization splits a character sequence into tokens, which are meaningful semantic units for processing, while lemmatization reduces words to their base form or lemma. To implement tokenization and lemmatization 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.[1]

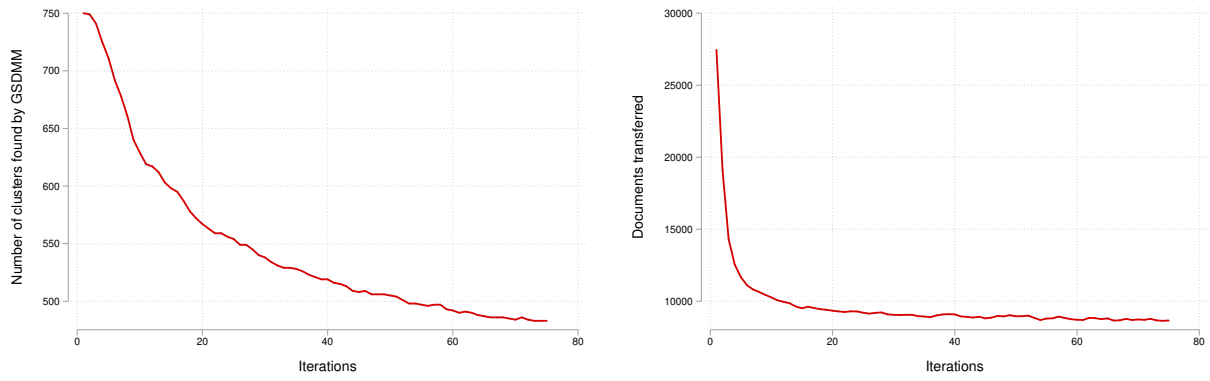
Next, we implement the GSDMM algorithim. This algorithm first randomly assigns all documents (job titles) to K clusters where K is a pre-defined upper limit on the number of topics (occupations) given as a human input to the algorithm. As long as K is larger than the 'true' number of clusters, the algorithm 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

[1] See https://spacy.io for more details.

Figure C.1: 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).

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 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.[2]

Yin and Wang (2014) use $\alpha = 0.1$, $\beta = 0.1$ and 30 iterations. 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 β. 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

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

clusters obtained in the manual classification using n-grams. Figure C.1 shows that the number of clusters and the number of documents transferred across clusters initially falls sharply, and then tends to stabilize after a few iterations.[3]

C.2 Pre-processing bag-of-n-grams Logistic Regression

For pre-processing the data, we first remove all special characters and numbers as well as extra spaces, i.e. we retain only alphabets. We convert all characters in the job text to lowercase. We also remove all words indicating an explicit gender preferences as mentioned in Section 2.1. If we were to retain these words, our algorithm's accuracy will be artificially inflated by classifying jobs largely on the basis of words that we originally used to code employers' gender preferences. We 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.[4] We remove words having length less than 2 or greater than 15 characters, and 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.[5] A generative

[3]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.

[4]For more details see https://www.nltk.org/.

[5]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 Bernoulli Naive Bayes (NB) classifier, the input x corresponds to binary-valued feature vectors indicating the presence or absence of n-grams in each job title.

model such as Bernoulli Naive Bayes (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 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).

C.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 token t in document d, the $TF - IDF$ score is computed as follows:

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

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 is 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. $TF - IDF$ vectors for each document are also 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.

C.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 LR 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.