Ashoka University Economics
Discussion Paper 63

Words Matter: Gender, Jobs and Applicant Behavior

October 2022

Sugat Chaturved, University of Sussex
Kanika Mahajan, Ashoka University
Zahra Siddique, University of Bristol

https://ashoka.edu.in/economics-discussionpapers

Words Matter: Gender, Jobs and Applicant Behavior*

Sugat Chaturvedi[†] University of Sussex

Kanika Mahajan[‡] Ashoka University

Zahra Siddique[§] University of Bristol

October 14, 2022

Abstract

We examine employers' gender preferences using 157,888 job ads posted on an online job portal in India which received 6.45 million applications. We find that explicit gender requests by employers explain 7% of the gender wage gap in applications after controlling for job location and occupation. Implicit gender associations in job ad text—indicating how predictive the text is of employers' gender preferences—together with explicit gender requests explain 17% of this gap. We retrieve words predictive of gender requests and find that skills and flexibility-related gendered words play an important role in observed gender disparities.

Keywords: Gender, Stereotypes, Discrimination, Job search, 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 are also grateful for comments and feedback from Farzana Afridi, Patrick Bennett, Sonia Bhalotra, Ian Burn, Rossella Calvi, Stefano Caria, Rochana Chaturvedi, Shoumitro Chatterjee, Nancy Chau, Sekyu Choi, Leandro De Magalhaes, Siddharth George, Seema Jayachandran, Thomas Le Barbanchon, Aprajit Mahajan, Arnaud Philippe, Nishith Prakash, Roland Rathelot, Hans Sievertsen, Shekhar Tomar, Basit Zafar, and Yanos Zylberberg, as well as numerous seminar and conference participants. All errors are our own.

[†]Address: Science Policy Research Unit, University of Sussex, Sussex House, Falmer, Brighton, BN1 9RH, United Kingdom. Email: sc2057@sussex.ac.uk

[‡]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 can be an indication that innately talented women are not pursuing their comparative advantage with the resulting misallocation having a detrimental effect on economic growth (Hsieh et al., 2019). Identifying the sources of such disparities, thus, continues to be an important aim of economics research. We examine gender disparities that arise at the recruitment stage when job ads play 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 relatively little research looking into how words in job ads are associated with gender stereotypes, how they relate to different characteristics of jobs and the posted wage, or how they direct where male and female job-seekers send their applications. We investigate these questions in this paper.[1]

We use proprietary data on 157,888 job ads posted between July 2018 and February 2020 on a leading Indian job portal together with 6.45 million applications made to these ads. We find that women apply to lower wage jobs than comparable men even when applying for jobs within the same occupation, or there is a gender wage gap in applications. We find that 7% of this gap can be explained by explicitly stated gender preferences since employers request women in lower wage jobs. We also apply text analysis on job titles and detailed job descriptions to construct measures that tell us how predictive the job ad text is of an explicit gender preference. We find that, together with explicitly stated gender preferences, the constructed measures or implicit gender associations within the job ad text explain as much as 17% of the gender wage gap in 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, and flexibility. We find that gendered words related to hard-skills and flexibility are particularly important in explaining gender disparities in labor market outcomes.

The job portal we use primarily caters to young urban job seekers with a university education. Jobs advertised on the portal are relatively 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

[1] Recent work by Burn et al. (2019) and Burn et al. (2021) examines similar questions, but focuses on age rather than gender stereotypes in job ads.

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. A key advantage of the data we use is that employers provide an informative wage range for slightly over 87% of job ads in our sample which is substantially higher in comparison to existing studies using job portal data, as we discuss in Section 2.

We use a multinomial logistic regression (LR) classifier from the literature on machine learning on text contained in a job ad's title and description to construct measures that indicate whether this text is predictive of an employer's explicit male or female preference. We refer to these measures as a job ad's implicit *maleness* and *femaleness*. We find that advertised wages are lowest in jobs that request women and that, even among jobs without an explicit gender preference, higher implicit *femaleness* is associated with a substantially lower advertised wage. We also find that an explicit female preference by employers is associated with fewer applications to a job ad as well as a higher fraction of female applicants. Among jobs without an explicit gender preference, higher implicit *femaleness* is associated with a higher share of female applicants. Consequently, explicit gender requests and implicit gender associations in job ad text (implicit *femaleness* and *maleness*) contribute to women applying to jobs with a lower advertised wage compared to men with similar characteristics. Our 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] Our results also remain largely robust when we use (firm $\times$ state) fixed effects, or (firm $\times$ occupation $\times$ state) fixed effects.

We find that women apply to jobs with 3.7% lower wages than men with the same education, age, and state of residence. We use the semi-parametric decomposition of DiNardo et al. (1996) to find that explicit gender requests explain 7% of the gender wage gap in applications, while implicit gender associations together with explicit gender requests explain 17% of the gap after accounting for differential applications to occupations by gender.

Next, we use the Local Interpretable Model-agnostic Explanations (LIME) algorithm to identify words in job ads that contribute to the decisions of the LR classifier; we refer to these words as

[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 an n-gram based classification using distinct unigrams, bigrams, and trigrams in job titles used in the existing literature (Marinescu and Wolthoff, 2020; Banfi and Villena-Roldan, 2019). Our results remain robust to using an n-gram based classification on our sample which yields 747 occupations.

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 gendered words 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. Second, we find that job ads with male gendered words related to (lower) job flexibility are associated with higher wages but get a smaller share of female applicants; this is consistent with compensating differentials whereby women are willing to trade off higher wages for increased flexibility.

We also use ridge regression to directly identify words which are associated with 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. So, while words such as *punctual*, *smile*, and *pleasant* are highly predictive of an employer's female preference, we find that they are associated with a lower share of women in the applicant pool.

Our work is related to several strands of the literature. We draw inspiration from a growing number of studies that document and examine explicit gender preferences in job ads across different countries (Kuhn and Shen, 2013; Helleseter et al., 2020; Ningrum et al., 2020; Chowdhury et al., 2018; Kuhn et al., 2020; Kuhn and Shen, 2021; Card et al., 2021). We extend this literature in several ways. First, we quantify the contribution of explicit gender preferences and implicit gender associations in job ad text to the gender wage gap in applications. Second, we deconstruct implicit gender associations in job ad text into a list of hard and soft-skills, personality, and flexibility related gendered words using machine learning methods. This allows us to highlight the importance of hard-skills and flexibility related gendered words in job ads.

Our work is also related to the large literature on gender wage gaps (see Olivetti and Petrongolo, 2016 and Blau and Kahn, 2017 for a review), as well as the literature on the sources of gender wage gaps beyond occupation selection (Goldin and Katz, 2011; Goldin, 2014). In a recent study, Fluchtmann et al. (2022) use data on applications by Unemployment Insurance recipients in the Danish labor market along with predicted typical wage of a job to show that the gender wage gap at the application stage explains up to 73% of the residual gender wage gap in realized starting

wages. While we do not observe realised starting wages, the actual posted wages are available for most job ads in our setting which allows us to look at factors which affect the gender wage gap in applications. Additionally, our setting allows us to show that employers' gender preferences can also explain why women apply to lower wage jobs than comparable men; indeed, we find that implicit gender associations within the job ad text are also important. In contrast employers cannot include gender requests in job ads in the setting studied in Fluchtmann et al. (2022). Our results are also consistent with studies which explain gender wage gaps as arising from women's willingness to pay for flexibility-related workplace attributes such as working part-time, working from home, and scheduling flexibility (Mas and Pallais, 2017; Wiswall and Zafar, 2018; He et al., 2019; Bustelo et al., 2020; Fluchtmann et al., 2022).

Moreover, our findings are consistent with studies that empirically verify the prediction of directed search models that job ads posting higher wages attract more applicants (Dal Bó et al., 2013; Belot et al., 2017; Banfi and Villena-Roldan, 2019; Marinescu and Wolthoff, 2020). Additionally, we find that higher posted wages attract better or more able applicants which is consistent with the theory and evidence in Dal Bó et al. (2013) and Marinescu and Wolthoff (2020). A further implication of directed search models with heterogeneity is endogenous segmentation with applicants targeting their search towards sub-markets where they meet the selection criteria set by employers (Shi, 2002; Menzio et al., 2016). Our work confirms this by providing evidence that applicants direct their search based on the explicit and implicit gender requirements of employers as given in job ad text.

Apart from economics, our work is also related to a literature in social psychology; within this literature, 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.[3] Additionally, within this literature, the characteristics that attract men and women to specific job ads are drawn from small, non-representative surveys.[4]

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

[4] Taris and Bok (1998) compile 20 characteristics based on 512 job ads judged by 40 students as being typically masculine or feminine while Gaucher et al. (2011) use lists of words denoted as feminine and masculine (based on gender differences in linguistic style) from existing studies.

In Section 2 we briefly describe our data and constructed variables; a detailed discussion of both can be found in Appendix A. Section 3 describes our empirical methodology and results when examining employer's gender preferences, with additional empirical results provided in Appendix B. Section 4 deconstructs the words used by employers to express a gender preference and examines their association with different outcomes. Section 5 briefly discusses robustness checks, with details provided in Appendix C. Section 6 describes words in job ads that are associated with a higher female applicant share while Section 7 concludes.

2 Data

We use data from a leading Indian job portal that advertises jobs located in all major Indian cities. Job seekers can create a profile on the portal for free and start applying to posted ads while employers need to pay a fee in order 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.[5]

We use data on jobs advertised on the portal with a last date of application between 24^{th} July 2018 and 25^{th} February 2020, together with all applications made to these ads. Our estimation sample consists of $157,888$ job ads posted on the portal over this time to which $1,060,731$ job-seekers made applications; see Appendix A.1.1 for details on the sample restrictions we make. Of the $157,888$ job ads in our sample, approximately 4.2% include an explicit female preference by the employer (F jobs), 3.5% include an explicit male preference (M jobs), and the rest have no explicit gender preference (N jobs).[6] Details on how we categorise jobs as F, N and M are given in Appendix A.1.2 while Appendix A.1.3 gives descriptive statistics for F, N and M jobs. Figure 1 shows word clouds of job titles in F, N and M jobs. As may 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

[5] Job seekers who register for a premium service with the portal are provided customized job recommendations and alerts on new jobs by e-mail. The proportion of job seekers who register for this service in our data is very low ($\approx$ 0.5%); therefore, the chances that applications are influenced by matching algorithms used by the portal are negligible.

[6] The fraction of F and M jobs we find are smaller than those reported by Chowdhury et al. (2018) using data from *Babajob*, another Indian job portal. 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.

preferences operate to maintain existing occupational gender stereotypes.

Employers post an informative wage range to job seekers in just over 87% of jobs advertised on the portal. Employers are required to provide a minimum wage for every job ad they post on the portal. Though they can choose to hide this information, the default option is to display wages to job-seekers.[7] We find that wages are less likely to be posted publicly in relatively high skill jobs which require more education and experience. This is similar to Brenčič (2012) and Banfi and Villena-Roldan (2019), as well as consistent with the model in Michelacci and Suarez (2006) where hidden wages might be used as a signal to high skill applicants that the employer is open to ex-post bargaining. Therefore, our sample of job ads with wage information is a somewhat selected sample of relatively lower skill jobs; nevertheless, we observe wages for a much higher fraction of job ads than existing studies. In contrast, wages are advertised in just 13.4% of job ads in Banfi and Villena-Roldan (2019) using *trabajando.com*, 20% of job ads in Marinescu and Wolthoff (2020) using *Careerbuilder*, 24.8% of job ads in Brenčič (2012) using *monster* and 16.4% of job ads in Kuhn and Shen (2013) using *zhaopin.com*. While the number of applications per job ad is not very different in our sample from these studies, it is possible that the portal's requirement to post a minimum wage in combination with a default option of displaying this wage to job-seekers, higher screening costs per application, or differences in corporate culture and HR practices in India compared to other contexts are behind a higher proportion of job ads in our sample posting a wage. Corporate culture and HR practices may be important since Chowdhury et al. (2018) also report that almost all job ads on *Babajob*, another Indian job portal, include an informative wage range.

We take the mid-point of the wage range as our measure of the posted wage.[8] N jobs have higher education requirements and mean posted wages than F or M jobs while F jobs have higher education requirements but a lower mean posted wage than M jobs (Appendix Table A.1). Consistent with lower female vs male labor force participation in India, there are 0.37 million female vs 0.69 million male job-seekers on the portal (see details in Appendix A.2). Job-seekers are relatively young (with an average age of 24 years) and 86% have an undergraduate or postgraduate degree. On average, female job-seekers are more educated in comparison to male job-seekers. Mean posted wages on

[7] We only have wage data in cases where employers choose not to hide posted wage information from job seekers.

[8] 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 wage data at both the top 0.5 percent and bottom 0.5 percent of the distribution to missing in order to mitigate the effects of any extreme outliers.

the portal are 21% higher than mean earnings of nationally representative and comparable samples of urban Indian workers (see details in Appendix A.2). This comparison indicates that the portal primarily caters to young and relatively skilled workers in the Indian labor market. A comparison by gender shows that female job seekers on the platform have higher education and are also more skilled than their urban peers. This is not surprising, since the female labor force participation in urban India follows a U-shaped relation with female education, declining till higher secondary and thereafter rising for graduate and above educated women Afridi et al. (2019).

We use text in job titles to categorise job ads to dis-aggregate occupations. We use two different methods: first, we use a short text topic model on job title text to assign all job ads in our sample to 483 occupations; second, we use distinct unigrams, bigrams and trigrams within job titles to assign all job ads to 747 occupations. Details on both methods are provided in Appendix A.3. While our main estimations make use of the occupation categorisation obtained using the first method our results are also robust to using the second (Appendix C.2).

We use a Multinomial Logistic Regression classifier on text contained in a job ad's title and description to construct measures indicating whether this text is predictive of an employer's explicit female or male preference which we refer to as a job ad's implicit *femaleness* (F_p) or *maleness* (M_p) (see details in Appendix A.4).[9] We find that F_p is high in job ads with titles such as **beautician**, **personal secretary**, and **school teacher**, while M_p is high in 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 based on the job description. For instance, consider two job ads titled **business development executive** in the data (Figure 2, job ad (ii) in both Panels); 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** (Figure 2, job ad (iii) in both Panels), high F_p is associated with jobs emphasizing appearance or communication skills while M_p is high in jobs requiring fieldwork.

[9]While the proportion of job ads with a gender request is about 8%, the absolute number of F and M jobs in our sample is over 12,000 which is sufficiently large to train our ML models. We are able to extrapolate the *maleness* and *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.

3 Gender preferences of employers

We examine characteristics of job ads associated with gender requests by employers in Appendix B.1.[10] We find, in line with the literature, that there is a **negative skill-targeting** relationship i.e., jobs with a higher skill requirement are less likely to have an explicit gender preference. This is consistent with the model in Kuhn and Shen (2013) where employers search broadly (or do not include an explicit gender preference in job ads) when the job's skill level is high and the number of applications are not plentiful (the "high frictions" case), since the marginal value of identifying the best candidate in such jobs is greater. Results in Appendix B.1 show that job ads with explicit gender preferences, especially for females, have lower posted wages. Next, we examine if posted wages are also correlated with implicit gender associations in the job ad text (F_p and M_p), and how applicant behavior varies with explicit gender requests as well as implicit gender associations.

3.1 Posted wages and applicant behavior

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

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

where $\ln W_{ijst}$ 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 *maleness*; see Appendix A.4 for details on how we construct these measures. 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 a job ad. In our preferred specification we include occupation and

[10] Though there are provisions within the Indian legal framework that could prohibit employers from posting job ads that explicitly request a male or female, the implementation of labor laws is generally inadequate. Article 16 of the Constitution of India prohibits discrimination on the basis of sex in public employment, while Article 39 guides the state to direct its policy towards ensuring "equal pay for equal work for both men and women". The Equal Remuneration Act, 1976 implements the provisions of Article 39 and prohibits sex based discrimination in payment of salary for same work (or work of similar nature) as well as in recruitment, promotion, training and transfer. Given the ambiguity about legal status of gender requests in job ads, it is not unusual for employers to express explicit gender preferences in the job ads they post online.

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; see Appendix A.3 for additional information on how we carry out this categorization. The use of fixed effects ensures we use **within** occupation and state variation to identify the effect of different variables on the log posted wage. We cluster standard errors by occupation and state.

Estimation results for equation (3.1) are reported in Table 1. As expected, higher education and experience requirements in a job ad are associated with an increase in the posted wage for all kinds of jobs. In jobs without an explicit gender preference (N jobs) an increase in implicit *femaleness* from zero to one is associated with a reduction in the posted wage by 38 log points, without occupation and state controls (column (III)). After including occupation and state fixed effects this coefficient drops to 26 log points but remains highly statistically significant (column (IV)). This translates to a decrease in the posted wage of 5.2 log points with a one standard deviation increase in implicit *femaleness* ($SD = 0.2$). On the other hand, an increase in *maleness* from zero to one 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. We find similar patterns in jobs with an explicit gender preference (F and M jobs) but the negative effect of *femaleness* on the log wage is smaller, though it is still statistically significant. The negative effect of *maleness* on the log wage in jobs with an explicit female preference (F jobs) is not significantly different from zero but becomes larger and statistically significant in jobs with an explicit male preference (M jobs).

Next we examine how explicit gender preferences are associated with job seeker's responses to an ad by estimating variations of the following regression specification:

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

where Y_{ijst}^{TA} is the total number of applications to a job ad. $F_{e,ijst}$ is a binary variable taking the value 1 if ad i has an explicit female preference, and 0 otherwise. Similarly, $M_{e,ijst}$ is a binary variable taking the value 1 if ad i has an explicit male preference, and 0 otherwise. 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.

Estimation results are reported in columns (I)–(III) of Table 2. We find that the number of applications is dramatically lower (≈ 21; 51% of mean) when 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 is not associated with a significant decline in the total number of applications to a job ad. Consistent with directed search models, we also find that there are a statistically significant **larger** 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 is associated with an increase in the number of applications to a job ad by approximately 19 (column (III)).

We further examine job seekers' behavior by estimating variations of the following regression specification:

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

where Y^S_{ijst} is the share of female applicants to a job ad. This is similar to equation (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. Coefficients on the binary variables $F_{e,ijst}$ and $M_{e,ijst}$ (or 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.[11]

Estimation results for equation (3.3) 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 is higher by $15.5-15.6$ percentage points (ppt) when an ad exhibits an explicit female preference and is lower by $9.5-9.9$ ppt when the ad has an explicit male preference (columns (V)–(VI)). These translate to 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

[11] In additional estimations using similar regression specifications as equation (3.3) we also investigate whether explicit gender requests are associated with applicant and match quality, but find the effects to be economically small; see details in Appendix B.2.

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 there is no association between the advertised wage and the share of female applicants. However, equation (3.3) does not control for applicant characteristics which are likely to be important since female applicants on the portal are more educated than male applicants. To check whether women apply to lower-wage jobs than observationaly similar men, or whether there is a gender wage gap in applications, we estimate application level regressions where the dependent variable is the log advertised wage of the job that a job-seeker applies to; see Appendix B.3 for details on the regression equations we estimate. The gender wage gap in applications is given by the coefficient on applicant gender in these regressions which also control for applicant characteristics such as education, age and state of residence. Using this strategy we find that women on the portal apply to jobs with 3.7% **lower** posted wages than comparable men (Appendix Table B.3, column (I)).

We also examine how implicit gender associations within the job text (F_p and M_p) are associated with the female applicant share in different kinds of jobs (F, N and M). We find that the share of female applicants increases as F_p increases for all kinds of jobs. We also find that explicit female requests matter more for female applicant shares than explicit male requests matter for male applicant shares; see details in Appendix B.4.

3.2 Gender wage gap in applications: Wage decomposition

In the previous sub-section and Appendix B.3 we show that women apply to lower-wage jobs than observationaly similar men, or that there is a gender wage gap in applications. As a next step, we quantify the importance of gender requests and implicit gender associations within the job ad text in explaining this gap by carrying out a semi-parametric decomposition using the method introduced in DiNardo et al. (1996).

Let the log posted wage in a job ad be w and the gender of an applicant be G where $G \in \{F, M\}$ i.e., female (F) or male (M). Let the set of observable characteristics of an applicant be denoted by a vector x which includes the applicant's education, age and location (or state of residence). Similarly, let the set of observable attributes of a job ad the job-seeker applies to be denoted by a vector a which includes occupation and gender requests in the job ad. The density of the log posted

wage in job ads that applicants of gender G apply to is

$$f^G(w) = \int\int f^G(w|x,a) f^G_{a|x}(a|x) f^G_x(x) dadx$$

where f^G_x is the distribution of observable applicant characteristics and $f^G_{a|x}$ is the conditional distribution of applications given these characteristics. The two step estimator we employ first conditions on observable characteristics of applicants. The base log wage gap between male and female applicants is given by $E(w|G=M) - E_{c_x}(w|G=F)$, where $E(w|G=M)$ is the average male log wage and $E_{c_x}(w|G=F)$ is the average female log wage given female applicants have the same observable characteristics as male applicants. The density function to obtain $E_{c_x}(w|G=F)$ is

$$\begin{aligned} f^F_{c_x}(w) &= \int\int f^F(w|x,a) f^F_{a|x}(a|x) f^M_x(x) dadx \\ &= \int\int f^F(w|x,a) f^F_{a|x}(a|x) f^F_x(x) \frac{f^M_x(x)}{f^F_x(x)} dadx \end{aligned}$$

where $\frac{f^M_x(x)}{f^F_x(x)}$ (also known as the re-weighting factor) can be estimated using a propensity score re-weighting method

$$\frac{f^M_x(x)}{f^F_x(x)} = \frac{Pr(G=M|x)}{1-Pr(G=M|x)} \frac{Pr(G=F)}{Pr(G=M)}$$

We use a logit model to estimate the propensity score and then use the obtained re-weighting factor to estimate $E_{c_x}(w|G=F)$.[12] This allows us to estimate the average base log wage gap between male and female applicants.

In a second step, we decompose the base gender log wage gap in applications into a part explained by differential application behavior of male and female job-seekers across jobs based on observable job attributes, such as gender requests made by employers in job ads. To do this we estimate the average log wage in job ads applied to by female applicants who have the same characteristics as male applicants when they also apply to the same jobs as the male applicants (based on the job attributes vector a). This wage is denoted by $E_{c_{x,a}}(w|G=F)$ and the associated wage density

[12] $Pr(G=M|x)$ can be estimated using a logit regression where the dependent variable is whether an applicant is a male based on observable applicant characteristics. $Pr(G=F)$ and $Pr(G=M)$ are the proportion of women and men in the sample.

required to estimate it is given by

$$f^F_{c_{x,a}}(w) = \int\int f^F(w|x,a) f^M_{x,a}(x,a) dadx$$
$$= \int\int f^F(w|x,a) f^F_{x,a}(x,a) \frac{f^M_{x,a}(x,a)}{f^F_{x,a}(x,a)} dadx$$

As before, we obtain the re-weighting factor $\frac{f^M_{x,a}(x,a)}{f^F_{x,a}(x,a)}$ by matching applicants both on their observable characteristics and their application behavior when responding to jobs with attributes a.[13] This allows us to estimate a counterfactual average female log wage $E_{c_{x,a}}(w|G = F)$. The gender wage gap if female and male applicants apply to jobs with the same attributes a is now given by $E(w|G = M) - E_{c_{x,a}}(w|G = F)$.

Using the above decomposition method, the baseline gender wage gap in applications after controlling for observable characteristics across applicants is 3.5% (Table 3, Panel A, Model 1 in column (I)); observable characteristics include education (indicator variables for different education levels attained by applicants), age (quadratic in applicant age) and location (indicator variables for applicant's state of residence).[14] Table 3, column (II), reports the explained component of the gender wage gap across different model specifications as we include different dimensions across which application behavior of men and women can differ. Model 1 controls for occupation and state in which the job is located, Model 2 controls for gender requests (F_e and M_e) as well as occupation and state whereas Model 3 controls for gender requests interacted with quartics in F_p and M_p as well as occupation and state. Estimates for Model 1 show that 45% of the baseline gender wage gap is explained by differential applications across occupations and state by men and women on the portal ($= \frac{0.0156}{0.0349}$). Estimates for Model 2 show that an additional 7% of the gender wage gap is explained by explicit gender requests on the portal.[15] Estimates for Model 3 show that an additional 17% of the baseline gender wage gap (over and above 45% from Model 1) is explained by explicit

[13] The expression for estimating the re-weighting function in this case is given by

$$\frac{f^M_{x,a}(x,a)}{f^F_{x,a}(x,a)} = \frac{Pr(G = M|x,a)}{1 - Pr(G = M|x,a)} \frac{Pr(G = F)}{Pr(G = M)}$$

.

[14] Note that this is the gender wage gap in applications among young and skilled job-seekers in the urban Indian labor market. The gender wage gap in applications is likely to be higher when examining older and/or unskilled job-seekers in India whom we cannot study since fewer such job-seekers use the portal (see also the discussion in Appendix A.2 and Appendix Figure A.1). We leave the study of the gender wage gap in applications among these important sub-populations to future work.

[15] This is obtained by subtracting 0.0156 from 0.0180 and dividing by the baseline gender wage gap.

gender requests and implicit gender associations in the job ad text (F_p and M_p). Overall, differential applications across occupations and state, gender requests and implicit gender associations in the job ad text explain 62% of the gender wage gap in applications on the portal.[16]

Similarly, Panel B of Table 3 reports decomposition results for N jobs. Unsurprisingly, the baseline gender wage gap is lower in N jobs than in all jobs at 3%. Results for Model 1 show that 44% of the gender wage gap in N jobs can be explained by differential applications across occupations and state by male and female applicants on the portal while Model 2 shows that a further 11% can be explained by implicit gender associations in the job ad text (F_p and M_p).

We also examine differences in the application wage across male and female applicants along the wage distribution; Figure 3 plots the differences in the application wage density between male and female applicants (male density - female density) who have the same observable characteristics (education, age and state of residence). As may be seen, the negative gender wage gap (female wage - male wage) at baseline arises from the lower to mid part of the wage distribution. Once we control for differences in applications across occupations and the state in which jobs are located (Model 1), the difference in application wages between men and women is reduced. It reduces further once we control for employer's gender requests (Model 2) and implicit gender associations in a job ad's text (Model 3).

The above decomposition does not provide a causal estimate of the effect of gender requests on the gender wage gap in applications since we are not able to fully account for all applicant characteristics (e.g. marital status and children, etc.) that may be correlated with applications to jobs with gender requests. Nevertheless, given the lack of previous evidence around this, the decomposition is able to quantify the importance of gender requests and implicit gender associations in explaining the gender wage gap relative to other factors that researchers have investigated. Our findings indicate that as much as 7% of the gender wage gap in applications can be explained by the presence of gender requests and as much as 17% can be explained by the presence of gender requests together with implicit gender associations in job ad text. In contexts where a higher fraction of job ads display an explicit gender request these proportions are likely to be even greater. Investigating

[16]The remainder of the gap is explained by differential applications across firms by male and female applicants. We do not control for firm fixed effects in the decomposition directly since the non-linear re-weighting function does not converge with firm fixed effects. However, we use an alternative specification to estimate the gender wage gap after controlling for firm fixed effects. This is described in Appendix B.3.

and reporting these estimates in different contexts is useful for understanding and tackling gender disparities within the labor market.

4 Deconstructing implicit gender associations

The previous sub-section highlights the importance of implicit gender associations within job ad text (F_p and M_p) in explaining the gender wage gap in applications. A natural question that arises is: what kind of words contribute to these implicit gender associations? In this Section we address this by deconstructing implicit gender associations, or job ad text which is predictive of explicit gender requests.

We use the Local Interpretable Model-agnostic Explanations (LIME) algorithm to obtain a measure of the importance of words in a job ad to the classification decisions of the Multinomial Logistic Regression (LR) classifier, which we used earlier to construct implicit gender associations within job ad text (F_p and M_p). This allows us to assign *contextual* relevance score (depending on which bigrams and trigrams it occurs in) to every word in each job ad indicating the importance of that word to the F, N, and M classes. To illustrate, Figure 2 gives a heat map visualization of words in distinctive F (Panel (a)) and M (Panel (b)) job ads in our data. Job ads (i), (ii) and (iii) in both panels refer to jobs titled **software trainee**, **business development manager**, and **sales market executive**, respectively. We find that words representing personality, appearance, communication skills and basic computer proficiency 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.

We construct a **net score** for each word that occurs at least ten times in the $13,735$ M and F jobs; this is the difference in the relevance of a word to the female vs male class. The use of net scores allows us to identify words that **distinctively** contribute to one class (female) relative to the other (male). We refer to words with a positive median net score as **female gendered words** and those with a negative median net score as **male gendered words** (see details in Appendix A.5). We assign female and male gendered words to meaningful categories related to desired hard and soft-skills, personality traits, and job flexibility related words.

We sum net scores across all words in a job ad's title and description for a given category to

obtain a category specific net score for each job ad. Job ads with a positive (negative) net score in a particular category either contain **more words** of the category that are relevant for the female (male) vs male (female) class or contain words of the category that have a **higher relevance** for the female (male) vs male (female) class. We examine how category specific net scores in job ads are associated with posted wages and female applicant shares to identify the type of words that contribute to the gender wage gap in applications.[17]

4.1 Gendered words

We list a maximum of twenty words for each category $C \in$\{hard-skills, soft-skills, personality, flexibility\} that have the highest median net scores or contribute the most to an employer's female or male requests in Table 4. The results are striking and show that many words that one would typically associate with male and female job roles indeed show up on the list.

Within the category of **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 the category of **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 job text that requests 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,

[17] Deming and Kahn (2018) classify skills-related key words in US job ads for professional workers into ten general skill categories in order to examine their correlation with external measures of pay and firm performance. In our work, apart from using skills-related words, we also examine words related to personality traits and job flexibility which may provide further insights on *gender differences* in applicant behavior.

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 the possibility of work from *home* or *home* based work are distinctively associated with jobs requesting a female (column (III), Panel B).[18] On the other hand, *night/rotational shifts*, working on *weekends*, possible *relocation* and *travel* (*petrol/fuel*) are distinctively associated with male requests (column (IV), Panel B).

4.2 Empirical methodology and results

To examine the association between category specific net scores in a job ad and the log posted wage we estimate the following Mincer regressions separately for F, N and M jobs:

$$\ln W_{ijst} = \rho^W + \sum_C \delta^C (NS_{C,ijst} \times \mathbb{1}[NS_{C,ijst} > 0]) + \sum_C \upsilon^C (-NS_{C,ijst} \times \mathbb{1}[NS_{C,ijst} < 0]) + \tau^W X_{ijst} + \gamma_{j\times s} + \phi_t + \zeta_{ijst} \tag{4.1}$$

where $\ln W_{ijst}$ 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. The explanatory variables of interest are the positive and negative values of standardized category specific net scores. By using equation (4.1) we allow positive values of the category specific net score ($(NS_{C,ijst} \times \mathbb{1}[NS_{C,ijst} > 0])$ for $C \in$ {hard-skills, soft-skills, personality, flexibility, others} or NS_C^+ for brevity) to have a different (linear) effect on the log wage in comparison to negative values of the net score ($(-NS_{C,ijst} \times \mathbb{1}[NS_{C,ijst} < 0])$ or NS_C^-).[19] A larger NS_C^+ indicates that the job ad text contains either more female gendered words or more highly relevant female gendered words related to category C while a larger NS_C^- indicates that the job ad text contains either more male gendered words or more highly relevant male gendered

[18] 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.

[19] We also estimate a flexible specification where we use quartics in category specific net scores as the explanatory variables of interest rather than positive and negative values of the scores, and find that our results are largely robust (Appendix C.4).

words related to C. Coefficients on NS_C^+ and NS_C^- (δ^C and υ^C for each C) give the log points change in the posted wage within a job ad for a standard deviation increase and decrease in the category specific net scores, everything else equal. X_{ijst} is a set of dummy variables for education and experience requirements in a job ad. Our preferred specifications include occupation and state fixed effects ($\gamma_{j\times s}$) as well as month-year fixed effects (ϕ_t). We cluster standard errors by occupation and state.

Table 5 reports results from estimation of equation (4.1). Using within occupation and state variation in jobs without an explicit gender preference (N jobs), we find that an increase in $NS^+_{hard-skills}$ is associated with a decrease in the posted wage while an increase in $NS^+_{personality}$ is associated with an increase in the posted wage (column (IV)). On the other hand, increases in $NS^-_{hard-skills}$, $NS^-_{soft-skills}$, $NS^-_{personality}$, and $NS^-_{flexibility}$ are all associated with an increase in the posted wage. A standard deviation increase in $NS^-_{flexibility}$ is associated with the highest increase in wages (= 1.8 log points) while a similar increase in $NS^+_{hard-skills}$ is associated with the largest decline in the posted wage (= 1.4 log points) in N jobs. In other words, posted wages increase the most in N jobs when the ad includes more or highly relevant flexibility related words that contribute to male vs female requests, everything else equal; these are words which indicate that a job is **less flexible**. At the same time posted wages decrease the most in N jobs when the ad includes hard-skills related words that contribute to female vs male requests, everything else equal. We find similar patterns in jobs with an explicit gender preference (F and M jobs), where an increase in $NS^-_{flexibility}$ is associated with higher posted wages while an increase in $NS^+_{hard-skills}$ is associated with lower posted wages.[20]

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

$$Y^S_{ijst} = \rho^S + \sum_C \eta^C (NS_{C,ijst} \times \mathbb{1}[NS_{C,ijst} > 0]) + \sum_C \iota^C (-NS_{C,ijst} \times \mathbb{1}[NS_{C,ijst} < 0]) + \tau^S X_{ijst} + \gamma_{j\times s} + \phi_t + \varsigma_{ijst} \tag{4.2}$$

where Y^S_{ijst} is the share of female applicants to job ad i. This specification is similar to equation

[20] In additional analysis, we find that the negative association of $NS^+_{hard-skills}$ 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.

(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. Coefficients on NS_C^+ and NS_C^- (η^C and ι^C for each C) give the percentage point change in the female applicant share for a standard deviation increase and decrease in the category specific net scores, everything else equal.

Table 6 reports the results from estimation of equation (4.2). Using within occupation and state variation in jobs without an explicit gender preference (N jobs) we find that a standard deviation increase in $NS^+_{hard-skills}$ and $NS^+_{soft-skills}$ are associated with an increase in the fraction of female applicants by 0.4 ppt (= 1.3% of mean applicant share to N jobs) and 0.2 ppt (= 0.6% of the mean) respectively (column (IV)). On the other hand, an increase in $NS^-_{flexibility}$ is associated with a reduced female applicant share by 0.6 ppt (= 1.9% of mean). While an increase in $NS^-_{flexibility}$ continues to be associated with a lower female applicant share in jobs with an explicit gender preference (F and M jobs), an increase in $NS^+_{hard-skills}$ is not associated with higher female applicant share in these jobs.

These results indicate the importance of flexibility and hard skills related gendered words in a job ad in explaining why women apply to lower wage jobs than comparable men. While gendered words indicating less flexibility are associated with higher posted wages in a job ad, fewer women apply to such jobs. At the same time, while words indicating hard skills which are predictive of female vs male requests are associated with lower wages in a job ad, we find that more women apply to such jobs.

5 Robustness checks

We check the robustness of our results to using a specification at the application rather than job ad level which allows us to include controls for applicant characteristics. We find that our results are robust; see Appendix C.1. We also find that our results are robust to an alternative method of categorizing job ads to occupations (Appendix C.2). Our results are also largely robust to using (firm × state) or (firm × occupation × state) fixed effects rather than (occupation × state) fixed effects (Appendix C.3). Finally, we find that the results in Section 4.2 are largely robust to an alternative specification that includes quartics in category specific net scores rather than positive and negative values of the net scores (Appendix C.4).

6 Words correlated with a high fraction of female applicants

In Section 4.1 we provided a list of **female** and **male gendered words** (or words that are predictive of an explicit female vs male preference by an employer) and then examined their effect on female applicant share. In this Section we examine all words in job ads, not just gendered words, which are associated with a higher fraction of female applicants. We arrive at these words using variation within a given occupation and location, to ensure comparability with estimates in Table 6, column (IV). Consistent with our findings in Section 4.2, we find a high correlation across the two lists in the categories of desired **hard-skills** (Spearman's rank correlation or $\rho = 0.23$) and job **flexibility** ($\rho = 0.50$) but a low or negative correlation in the categories of desired **soft-skills** ($\rho = 0.03$) and **personality** traits ($\rho = -0.12$). This indicates a higher correspondence between employer stereotypes and applicant responses for hard-skills and flexibility than for soft-skills or personality. We discuss the details below.

We use jobs without a gender preference (N jobs) to first estimate applicant share variation within a given occupation-location. 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.[21] 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.[22] The model gives a coefficient for each word which we interpret as a marginal effect of the presence of the word on the (residual) female applicant share. Words with a positive effect, or which are associated with an increased female applicant share, are included in the female list while those with a negative effect are included in the male list. These are further classified into each category $C \in$ {hard-skills, soft-skills, personality, flexibility}. Table 7 reports the top 20 words in the female and male list for each category, with the coefficient for the word in parentheses.[23]

Within the category of **hard-skills** (column (I), Panel A), words related to beauty service, accounting, and *architectural* skills continue to appear among words associated with a larger share

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

[22] Ridge regression prevents over-fitting from 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 words instead of a sparse number of features. Secondly, ridge regression gives 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.

[23] We only keep words which have a coefficient exceeding one percentage point in the table.

of female applicants. In addition we find words related to legal professions, software and database management, automation, and content creation in this list. Within this category, words that are associated with the highest fraction of male candidates continue to be dominated by those 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).[24]

Within the **soft-skills** category, the female applicant share increases 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* continue to be associated with a larger share of male applicants (column (IV), Panel A).

Within the category of **personality** traits (columns (I)–(II), Panel B), there are several deviations from the list of gendered words. The 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*. In 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 4). Indeeed, we find that *punctual*, *smile* and *pleasant*, which are all female gendered words within the category of **personality** traits, are actually associated with a **reduced** share of female applicants to a job ad.

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 *weekdays* which increase the (residual) female share of applicants by approximately 2.5 ppt. At the same time words reflecting job characteristics involving *night shift* and *travel* decrease the (residual) female applicant share by 10 and 5 ppt respectively, which are relatively large effects.

7 Conclusion

We find that young, skilled women in the urban Indian labor market apply to lower wage jobs than comparable men. This can partly be explained by employers' gender preferences in job ads, since employers exhibit a female preference in low wage jobs and women tend to direct their applications

[24]Note that these words for hard skills can be associated with certain occupations. However, we have purged the effect of occupation in our analyses by taking the residual female applicant share as the dependent variable. These results then show that among the job ads for the occupation of 'beautician', those that mention 'facial' or 'makeup' are more likely to attract a higher share of female applicants.

toward these jobs. We find that employers' gender requests can explain as much as 7% of the gender wage gap in applications, while gender associations in job ad text together with gender requests can explain as much as 17% of this gap. We find that it is gendered words in job ads which are related to hard skills and job flexibility that play an important role. This is because a higher fraction of women apply to job ads which include hard-skills related words predictive of an employer's relative female preference which have low returns. At the same time a lower fraction of women apply to jobs with low-flexibility related words predictive of an employer's relative male preference which have high returns. The gender wage gap in applications we study is important—in a recent paper Fluchtmann et al. (2022) show that "differences in applied-for jobs are able to explain 86 percent of the residual gender gap in the typical wage level of the jobs males and females hold and 73 percent of the residual gender gap in realized starting wages."

The job-seekers we examine consist primarily of young Indian workers who are just entering the labor market after completing their university degree. Several papers document persistent effects of initial labor market conditions, such as a recession, at the time when young workers enter the labor market on their long-term labor market outcomes (Oyer, 2006; Kahn, 2010; Oreopoulos et al., 2012; Rothstein, 2020). Gender differences at an early career stage for the job-seekers we look at are also likely to have important cumulative consequences for future labor market returns, and result in persistent gender wage gaps.

References

Afridi, F., M. Bishnu, and K. Mahajan (2019): "What Determines Women's Labor Supply? The Role of Home Productivity and Social Norms," .

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

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

Blau, F. D. and L. M. Kahn (2017): "The gender wage gap: Extent, trends, explanations," *Journal of Economic Literature*, 55, 789–865.

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.

Brenčič, V. (2012): "Wage posting: Evidence from job ads," *Canadian Journal of Economics*, 45(4), 1529–59.

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.

Burn, I., D. Firoozi, D. Ladd, and D. Neumark (2021): "Machine learning and perceived age stereotypes in job ads: Evidence from an experiment," 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.

Card, D., F. Colella, and R. Lalive (2021): "Gender Preferences in Job Vacancies and Workplace Gender Diversity," Tech. rep., National Bureau of Economic Research.

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.

Dal Bó, E., F. Finan, and M. Rossi (2013): "Strengthening state capabilities: The role of financial incentives in the call to public service," *Quarterly Journal of Economics*, 128(3), 1169–218.

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.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044.

Fluchtmann, J., A. Glenny, N. Harmon, and J. Maibom (2022): "The Gender Application Gap: Do men and women apply for the same jobs?" Unpublished manuscript.

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.

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.

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.

Kuhn, P. J. and K. Shen (2021): "What Happens When Employers Can No Longer Discriminate in Job Ads?" Tech. rep., National Bureau of Economic Research.

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).

Menzio, G., I. Telyukova, and L. Visschers (2016): "Directed search over the life cycle. Special issue in honor of Dale Mortensen," *Review of Economic Dynamics*, 19, 38–62.

Michelacci, C. and J. Suarez (2006): "Incomplete Wage Posting," *Journal of Political Economy*, 114(6), 1098–123.

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.

Olivetti, C. and B. Petrongolo (2016): "The Evolution of Gender Gaps in Industrialized Countries," *Annual Review of Economics*, 8, 405–434.

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.

Rothstein, J. (2020): "The Lost Generation? Labor Market Outcomes for Post Great Recession Entrants," Tech. rep., National Bureau of Economic Research.

Shi, S. (2002): "A directed search model of inequality with heterogeneous skills and skill-biased technology," *Review of Economic Studies*, 69(2), 467–91.

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.

Tables & Figures

Table 1: Wages

Sample:	F jobs		N jobs		M jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
Implicit *femaleness* (F_p)	–0.185***	–0.202***	–0.379***	–0.264***	–0.320***	–0.192***
	(0.052)	(0.039)	(0.023)	(0.017)	(0.069)	(0.069)
Implicit *maleness* (M_p)	–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)
Graduate 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)
Graduate 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)
Postgraduate 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)
Postgraduate 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
$F_p = M_p$, 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 in Appendix A.1.1. All columns report the effective number of observations after incorporating (occupation × state) fixed effects which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table 2: Applications

Dependent variable:	total applications			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female preference (F_e)	–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 (M_e)	–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)
Graduate 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)
Graduate 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)
Postgraduate 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)
Postgraduate 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 in Appendix A.1.1. Columns (II)-(III) and (V)-(VI) report the effective number of observations after incorporating (occupation × state) fixed effects which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table 3: Wage Decomposition

	(I)	(II)	(III)	(IV)
	Baseline wage gap	Explained	Residual	Description
Panel A: All jobs				
Model 1	0.0349	0.0156	0.0193	Explained by differential applications by gender across
	(.0011)	(0.0014)		job location and occupation
Model 2	0.0349	0.0180	0.0169	Explained by differential applications by gender across
	(.0011)	(0.0014)		F_e and M_e as well as job location and occupation
Model 3	0.0349	0.0215	0.0134	Explained by differential applications by gender across
	(.0011)	(.0014)		F_e and M_e interacted with quartics in F_p and M_p as well
				as job location and occupation
Panel B: N jobs				
Model 1	0.0294	0.013	0.0165	Explained by differential applications by gender across
	(.0011)	(0.0014)		job location and occupation
Model 2	0.0294	0.0161	0.0133	Explained by differential applications by gender across
	(.0011)	(0.0014)		quartics in F_p and M_p as well as job location and occupation

Notes: The baseline wage gap in column (1) refers to the log of the difference between average male and female application wage when male and female applicants have the same characteristics (education, age and state of residence). Each model successively adds more job attributes to understand how much of the baseline wage gap is explained by the various job attributes or characteristics. Column (2) shows the part of the wage gap explained by differential applications by men and women to jobs with different attributes. Column (3) shows the residual wage gap for each model. Job attributes or characteristics are listed in column (4). Robust standard errors in parenthesis.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions in Appendix A.1.1.

Table 4: Gendered words

(I)	(II)	(III)	(IV)
Panel A			
Hard-skills		**Soft-skills**	
Female	Male	Female	Male
autocad	hardware	telugu	arabic
facial	wpm	fluent	supervise
pedicure	rcm	malayalam	pitch
manicure	regulation	talk	negotiate
ppt	qc	counsel	verbally
tally	manual	speak	marathi
computer	mysql	gujarati	persuade
cake	scan	edit	punctuation
auto	machine	verbal	write
coral	sql	bengali	french
hashtag	audit	hindi	liaise
zoho	troubleshoot	crm	motivate
word	receivable	accommodate	read
ms	rf	oral	communicate
ledger	trouble	convince	advise
expense	visual	english	ar
manuscript	demat	etiquette	grammar
makeup	instagram	coordinate	rapport
keyword	outward	story	relationship
architectural	campaign	engage	color
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: Up to 20 words with the highest positive median category specific net score (NS_C for $C \in$ {hard-skills, soft-skills, personality, flexibility}) are listed in columns (I) and (III); these words contribute the most towards female vs male requests in a particular category. Up to 20 words which have the lowest negative median category specific net score are listed in columns (II) and (IV); these words contribute the most towards male vs female requests. Words are sorted in decreasing order of importance or magnitude of the median category specific net scores. Appendix A.5 provides details on how the category specific net scores are constructed.
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 on the portal.

Table 5: Net scores and wages

Sample:	*F* Jobs		*N* Jobs		*M* Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
$NS^{+}_{hard-skills}$	–0.044***	–0.025***	–0.031***	–0.014***	–0.022***	–0.021***
	(0.006)	(0.005)	(0.003)	(0.002)	(0.008)	(0.008)
$NS^{+}_{soft-skills}$	–0.009	–0.009*	–0.000	–0.001	–0.003	–0.004
	(0.006)	(0.004)	(0.002)	(0.002)	(0.006)	(0.006)
$NS^{+}_{personality}$	0.011*	0.005	0.018***	0.005***	0.019***	–0.001
	(0.005)	(0.005)	(0.003)	(0.002)	(0.006)	(0.005)
$NS^{+}_{flexibility}$	0.009	0.003	0.006***	0.003	–0.000	0.002
	(0.008)	(0.008)	(0.002)	(0.002)	(0.006)	(0.005)
NS^{+}_{others}	–0.016	–0.019***	–0.055***	–0.027***	0.007	0.006
	(0.008)	(0.006)	(0.004)	(0.002)	(0.016)	(0.017)
$NS^{-}_{hard-skills}$	–0.025*	–0.014	0.008**	0.006***	–0.019***	0.005
	(0.012)	(0.012)	(0.004)	(0.002)	(0.006)	(0.007)
$NS^{-}_{soft-skills}$	–0.006	–0.003	0.017***	0.011***	0.011	0.010
	(0.007)	(0.005)	(0.002)	(0.002)	(0.008)	(0.010)
$NS^{-}_{personality}$	–0.001	0.003	0.008***	0.005***	–0.003	–0.002
	(0.007)	(0.006)	(0.002)	(0.002)	(0.006)	(0.005)
$NS^{-}_{flexibility}$	0.044***	0.027***	0.027***	0.018***	0.013*	0.010*
	(0.013)	(0.007)	(0.003)	(0.002)	(0.007)	(0.005)
NS^{-}_{others}	0.012	–0.000	0.006	–0.005	0.013*	–0.003
	(0.015)	(0.015)	(0.005)	(0.003)	(0.006)	(0.004)
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. Coefficients on positive and negative values of the category specific net scores are reported; see Appendix A.5 for details on how the category specific net scores are constructed and Section 4 for the regression specification. All regressions control for the set of education and experience requirement categories given in 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 in Appendix A.1.1. All columns report the effective number of observations after incorporating (occupation × state) fixed effects which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table 6: Net scores and the share of female applications

Sample:	*F* Jobs		*N* Jobs		*M* Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
$NS^{+}_{hard-skills}$	0.004	0.002	0.011***	0.004***	0.000	–0.001
	(0.005)	(0.003)	(0.002)	(0.001)	(0.004)	(0.003)
$NS^{+}_{soft-skills}$	–0.002	–0.003	0.006***	0.002**	0.006	0.002
	(0.004)	(0.002)	(0.002)	(0.001)	(0.003)	(0.003)
$NS^{+}_{personality}$	0.001	0.001	0.001	0.001	0.008	0.002
	(0.004)	(0.002)	(0.002)	(0.001)	(0.004)	(0.003)
$NS^{+}_{flexibility}$	–0.002	–0.000	0.001	0.001	0.004	0.001
	(0.004)	(0.004)	(0.001)	(0.001)	(0.003)	(0.001)
NS^{+}_{others}	0.008	–0.002	0.017***	0.007***	0.027**	0.018***
	(0.004)	(0.002)	(0.003)	(0.001)	(0.011)	(0.007)
$NS^{-}_{hard-skills}$	–0.036***	–0.013**	0.005***	–0.001	0.011***	0.003
	(0.009)	(0.005)	(0.002)	(0.001)	(0.002)	(0.001)
$NS^{-}_{soft-skills}$	–0.006	0.001	0.002	0.001	–0.001	0.001
	(0.006)	(0.003)	(0.001)	(0.001)	(0.005)	(0.004)
$NS^{-}_{personality}$	0.003	0.000	0.001	–0.001	0.002	–0.005**
	(0.005)	(0.003)	(0.001)	(0.001)	(0.002)	(0.002)
$NS^{-}_{flexibility}$	–0.022***	–0.014***	–0.007***	–0.006***	0.004	–0.007**
	(0.004)	(0.004)	(0.002)	(0.001)	(0.003)	(0.003)
NS^{-}_{others}	–0.052***	–0.027**	–0.032***	–0.011***	–0.020***	–0.005***
	(0.014)	(0.011)	(0.004)	(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

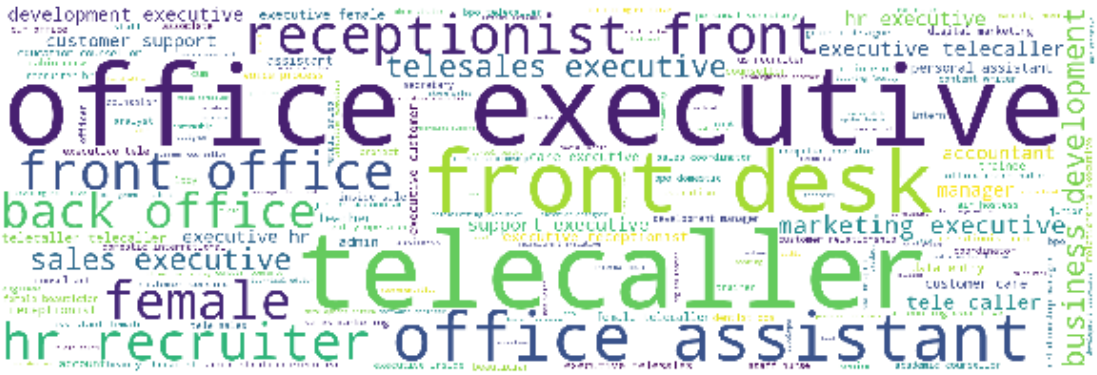
Notes: The dependent variable is the fraction of female applicants to a job ad. Coefficients on positive and negative values of the category specific net scores are reported; see Appendix A.5 for details on how the category specific net scores are constructed and Section 4 for the regression specification. All regressions control for the set of education and experience requirement categories given in a job ad as well as being 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 in Appendix A.1.1. All columns report the effective number of observations after incorporating (occupation × state) fixed effects which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table 7: Words associated with the share of female applicants

(I)	(II)	(III)	(IV)
Panel A			
Hard-skills		**Soft-skills**	
Female	Male	Female	Male
makeup (.106)	python (-.115)	write (.057)	collaborate (-.048)
legal (.076)	desktop (-.061)	bengali (.055)	ar (-.04)
facial (.066)	robotic (-.055)	guide (.053)	telugu (-.039)
architectural (.062)	quantitative (-.047)	counsel (.052)	speak (-.026)
rf (.061)	install (-.043)	rapport (.037)	supervise (-.023)
manuscript (.057)	machine (-.039)	relationship (.036)	verbal (-.021)
compute (.051)	server (-.038)	english (.035)	read (-.02)
court (.048)	plc (-.036)	story (.03)	edit (-.017)
cnc (.045)	guest (-.036)	french (.028)	negotiate (-.017)
content (.044)	statement (-.034)	crm (.025)	marathi (-.016)
proofread (.044)	configuration (-.033)	coordinate (.022)	articulate (-.015)
draft (.04)	repair (-.032)	feedback (.021)	persuade (-.015)
database (.038)	adobe (-.032)	verbally (.02)	neutral (-.013)
software (.038)	es (-.031)	influence (.018)	engage (-.013)
risk (.036)	network (-.031)	conversation (.016)	pitch (-.012)
cake (.034)	knowledgeable (-.03)	convince (.014)	clientele (-.011)
demonstration (.033)	erp (-.03)	communicate (.012)	malayalam (-.011)
animation (.032)	ui (-.03)	liaise (.012)	etiquette (-.01)
automation (.031)	collate (-.028)	color (.009)	motivate (-.009)
regulation (.031)	seo (-.027)	interpersonal (.009)	arabic (-.009)
Panel B			
Personality/Appearance		**Flexibility**	
Female	Male	Female	Male
personality (.053)	punctual (-.034)	skype (.026)	night (-.103)
appearance (.046)	smile (-.032)	weekday (.02)	travel (-.049)
ethic (.042)	adapt (-.028)	outstation (.015)	petrol (-.041)
mile (.042)	tone (-.026)		fuel (-.019)
resourceful (.04)	dedicate (-.024)		rotational (-.016)
initiative (.039)	keen (-.024)		relocate (-.013)
motivation (.039)	pleasant (-.021)		shift (-.012)
determination (.031)	neat (-.021)		
proactively (.031)	chest (-.019)		
zeal (.027)	entrepreneurial (-.019)		
responsive (.027)	adaptability (-.019)		
proactive (.026)	confident (-.018)		
creative (.026)	vigilant (-.017)		
passionate (.022)	enthusiasm (-.017)		
rejection (.021)	hardworke (-.017)		
thinker (.021)	height (-.017)		
attitude (.02)	initiate (-.017)		
persuasive (.019)	learner (-.016)		
professionalism (.018)	empathy (-.015)		
creatively (.016)	dedication (-.013)		

Notes: Parentheses show the marginal effect on female applicant share for a word. Up to 20 words in each category $C \in$ {hard-skills, soft-skills, personality, flexibility} that increase the female applicant share the most are listed in columns (I) and (III) while those that decrease the female applicant share the most are listed in columns (II) and (IV). We only retain words where the absolute marginal effect exceeds one percentage point. Words are sorted in decreasing order of absolute marginal effects within each gender-category.
Abbreviations - rf (radio frequency), cnc (computerized numerical control), plc(programmable logic controller), es(engineering science), erp (enterprise resource planning), ui(user interface), seo (Search Engine Optimization), ar(augmented reality), crm (customer relationship management).
Source: Data from the population of N jobs and applicants to these jobs on the portal, subject to the restrictions in Appendix A.1.1.

Figure 1: Word clouds of job titles



(a) Female preference (F jobs)

(b) Male preference (M jobs)

(c) No gender preference (N jobs)

Notes: Word clouds are constructed using words contained in job titles for F, N and M job ads.

Source: Data from the population of all job ads on the portal, subject to the restrictions in Appendix A.1.1.

Figure 2: Heat map visualization of words in F and M job ads

i. **SOFTWARE TRAINEE**: lady faculty for following subjects - basic of computer having complete knowledge of ms office. friendly with internet. advance english with grammar. personality development classes having good comunication skills. basic & accounting with taly & gst

ii. **BUSINESS DEVELOPMENT MANAGER**: language:- bengali (fluently speak), english (read, write & fluently speak), hindi (fluently speak) grooming must (looking like air hostess) job role:- manager, hr, student, counselling, employee handling, eod report sharing (total office management) bond applicable for this employee qualification (preferable) :- minimum graduate, mba in marketing, master in psychology. only female candidates applicable. (good looking with smart candidates) computer knowledge:- power point, mail communication, excel, presentation skills. age :-18-30 height:-5'6, weight:- proportionate to height

iii. **SALES AND MARKETING EXECUTIVE:** we are hiring a smart, intelligent and good looking female candidate for the below role: preference for candidates who have sales experience in the aviation sector or you have experience in selling to tour operators, hotels & corporate clients. If you have completed cabin crew training, will be an added advantage | candidate must have good communication skills in english & malayalam and if you can speak other regional languages it will be an added advantage | must be smart and good looking | able to handle high profile clients | new business development & manage existing clients with their day to day flight requirements | managing customer relationships | supporting the head of sales | in addition to salary, you will be entitled to incentives on achieving set targets.

(a) Female preference (F jobs)

i. **SOFTWARE TRAINEE**: qualification: b.e/ b.tech/b.sc/bca mca msc freshers – 2018 & 2019 passed out requirement: candidates from it/ computer science background are preferred. Excellent verbal and written communication skills should have basic knowledge on it technologies quick learners should be able to work in rotational shifts only male candidates are preferred

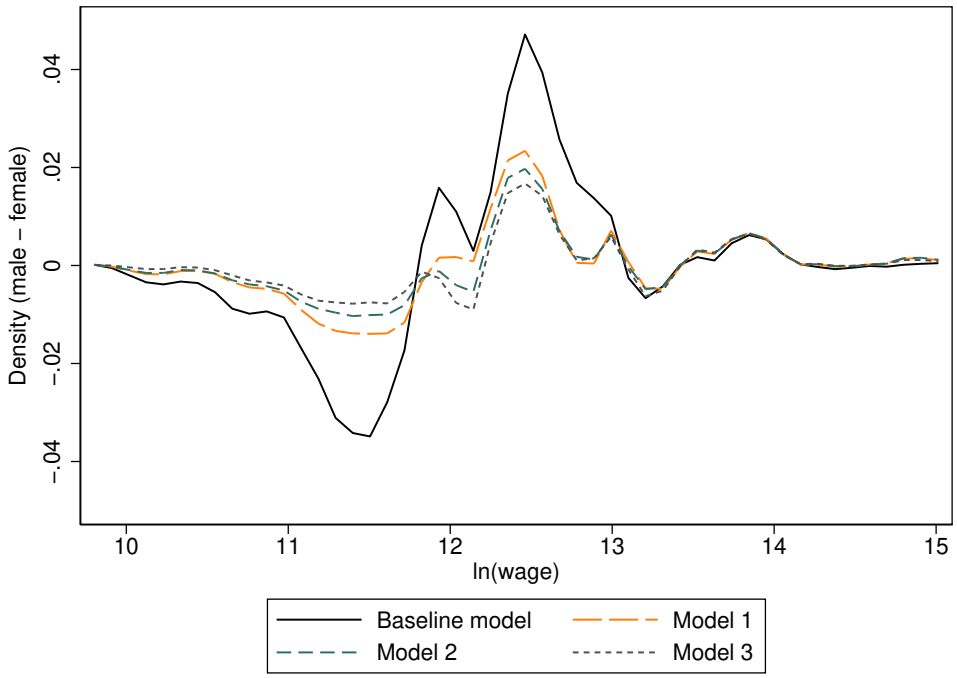
ii. **BUSINESS DEVELOPMENT MANAGER**: we are looking for energetic candidates for the post of bdm who has experience in b2b sales and has good communication skills, only boys with two-wheelers. Salary will be 4-6 lakhs p.a. jd – you have to set up and deliver sales presentations, demo on a daily basis, to identify potential clients and implementing innovative business strategies.

iii. **SALES / MARKETING EXECUTIVE**: fixed salary + incentive up to 25k job role: calling + field work education: any degree/diploma experience: fresher and experience designation: marketing manager salary: salary up to 25k shift: general shift gender: male (two wheeler mandatory) language: tamil job location: chennai

(b) Male preference (M jobs)

Notes: Panels (a) and (b) show correctly classified job ads with an explicit female (F jobs) and male preference (M jobs). Words highlighted in red reflect female associations while those in blue reflect male associations, as returned by LIME. Color intensity reflects the strength of the attached gender association with darker shades indicating a stronger association.

Figure 3: Distribution of the gender wage gap in applications: Actual vs Counterfactual wage gaps



Notes: Distributions are the kernel density estimates for distribution of differences in job applications by male and female applicants using the mid-point of the posted wage range in job ads. The baseline wage gap plots the application wage density differences between male and female applicants who have the same characteristics (education, age and state of residence). Model 1 plots the application wage density differences for male and female applicants who have the same characteristics and application behavior with regards to occupation and the state in which a job is located; Model 2 additionally accounts for gender requests in job ads while Model 3 additionally accounts for gender requests interacted with quartics in F_p and M_p.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions in Appendix A.1.1.

A Appendix: Additional details on data and constructed variables

A.1 Job ads

A.1.1 Sample restrictions

There were 196,821 jobs advertised on the portal with a last date of application between 24^{th} July 2018 and 25^{th} February 2020. We use the following steps to reach our final estimation sample:

1. We drop job ads with a location outside India. We also drop ads having an application window of less than a day (so that there is sufficient time for job-seekers to make an application) and more than 4 months/120 days (removing job ads posted well before July 2018). These restrictions reduce the sample to 188,857 job ads.

2. We combine duplicate job ads posted within a month of the original ad which reduces the sample of job ads further to 175,126 unique ads.[1]

3. Since we are interested in how men or women apply to different jobs, we drop job ads that have no male or female applicants. This restriction reduces the sample further to 171,960 ads.

4. Since we use education and experience requirements as controls in our regressions we drop any job ads that do not explicitly mention an education and experience requirement. This reduces the sample slightly to 171,940 ads.

5. In order to include (occupation $\times$ state) fixed effects in our regressions we restrict the sample to job ads that specify cities within a single Indian state as the location of the job. This reduces the sample to 158,249 ads.

6. In order to include (occupation $\times$ state) fixed effects in our regressions we further restrict the sample to those job ads for which we are able to obtain an occupation classification based on the method described in Appendix A.3. This leaves us with a final estimation sample of 157,888 job ads.

[1] Duplicate job ads are defined as those which have identical 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.

The largest reduction in the estimation sample comes from restriction 5. We check the robustness of our results to this sample restriction, and find that our results are robust to including job ads which advertise for positions across multiple Indian states and to using (occupation $\times$ state) fixed effects where multiple state jobs are all given the same (artificial) location or state. We omit these results for brevity but they are available on request.

A.1.2 Explicit gender preferences

The job portal does not have a separate field allowing employers to directly state the preferred gender in a job ad. However, employers indicate their explicit gender preference in the job title or description of a job ad 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. Some job ads include words related to both genders. We categorize such ads as having no explicit gender preference, together with ads that do not include words related to either gender.

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. 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. Next, 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 F, N or M. The annotators agreed on the classification for 90% of these 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.

A.1.3 Descriptive statistics

A small proportion of jobs advertised on the portal specify 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.

Consistent with the literature on gender targeting in job ads, we find that ads specifying a gender preference are also more likely to specify other preferences, such as those related to age or beauty. We derive the presence of age and beauty preferences from the text of the job ad (see Appendix B.1 for details), and find that M jobs are more likely to specify an age preference (with these jobs tending to specify a higher minimum and maximum required age than F or N jobs) while F jobs are most likely to specify a beauty requirement.

We take the mid-point of the wage range as our measure of the posted wage; the mean of this posted wage for jobs in our sample is just above INR $213,000$ per year. N jobs have higher mean posted wage than F and M jobs while 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 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 not 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 applicant characteristics on the portal, we use Cohen's kappa.[2] 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.

There are about 41 applications per job ad, on average. The average number of applications to F jobs is less than half of this, at about 17, while the average number of applications to M jobs is

[2]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%.

about 31. This suggests that explicit gender preferences are associated with a substantially reduced number of applications, particularly by job seekers of the opposite gender to the preferred one.

A.2 Job seekers

We also use data on the 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 applicants. The smaller number of female applicants 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 applicants make a similar number of job applications as male applicants. Most job seekers on the portal (86%) have an undergraduate or postgraduate degree though women are more likely to have a postgraduate degree. Applicants are relatively young with an average age of 24 years and about 76% have less than a year of experience. Female applicants are slightly younger than men and are less experienced. Despite having better education qualifications, women (unconditionally) apply to job ads with similar posted wages as men.

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).[3] Advertised wages on the portal are higher than the PLFS sample by about Rs. 14,000 per annum. However, wages in the PLFS sample could also be higher than those advertised on the portal because the PLFS sample has older and more experienced workers. To make the PLFS sample comparable to the age group catered to by the portal we only keep adults aged 18–32 years (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 than the PLFS sample. Thus, the portal caters to young and inexperienced, but more educated and skilled workers.

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

Appendix Figure A.1 further confirms these patterns. The wage distributions for urban workers using the PLFS sample 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 (Appendix Figures A.1(a) and A.1(c)). However, if we restrict the PLFS sample to employed urban workers aged 18–32 with at least an undergraduate degree, we find that gender wage disparities are comparable to disparities in posted wages across F and M jobs on the portal (Appendix Figures A.1(b) and A.1(c)).

A.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. The final number of topics or occupation categories obtained using this method for our sample of job ads is 483.

To implement GSDMM, we use the following pre-processing steps on the job title text: (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.[4] This leaves us with $D = 28,957$ documents and $V = 3,127$ unique words.

Next, we implement the GSDMM algorithm. It 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 automatically infers the appropriate number of clusters. In each subsequent iteration, it 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 the following probability:

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

where $\vec{z}$ is the cluster label of each document, m_z is the number of documents in cluster z, n_z is the number of words in cluster z and n^w_z represents the number of occurrences of word w in cluster z. $\neg d$ denotes that cluster label of document d is removed from $\vec{z}$. D refers to the total number of documents in the corpus, N_d is the number of words in document d and V is the total number of 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.[5]

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

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

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

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 becomes very small after that number. We tried up to 100 iterations and found that at approximately 75 iterations both these criteria are met. Appendix Figure A.2 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. Lastly, the initial number of clusters (K) are chosen to be approximately equal to the number of clusters obtained in the n-gram based classification.[6]

Our empirical results are also robust to an alternative n-gram based clustering of job ads to occupation categories based on distinctive unigrams, bigrams, and trigrams in job titles as used in the existing literature (Marinescu and Wolthoff, 2020; Banfi and Villena-Roldan, 2019).[7] To implement the n-gram based 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. This gives us a total of 747 occupation categories.

We prefer GSDMM to the n-gram based 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 together through

[6]There is no direct way to assess objectively whether short text topic model or n-gram based 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 n-gram based 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 n-gram 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.

[7]We discuss estimation results using the alternative categorization in Appendix C.2.

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 n-gram based 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 n-gram based classification because the word n-grams contained in them occur with a low frequency across the corpus.

A.4 Implicit *femaleness* and *maleness*

Text contained in a job ad which is predictive of an explicit female preference may also be associated with more female applicants—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 machine learning to infer F_p and M_p for each job ad based on the text that appears in a job ad's title and description. Specifically, we train a Multinomial Logistic Regression (LR) classifier where the output class can take three values depending on the employer making an explicit request for women, men, or no gender request. We use the complete set of 196,857 job ads to increase data points for the classification model. We use balanced class weights since the classes are imbalanced due to a relatively smaller fraction of jobs having an explicit gender request.

We follow standard pre-processing steps. 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 preference as mentioned in Section A.1.2. If we were to retain these words, our algorithm's accuracy would be artificially inflated by classifying jobs largely on the basis of words that we originally used to code employers' gender preferences. We also filter out stop words (such as "the", "are", "and") which are uninformative in

representing the text using the Stopwords corpus of the Natural Language Toolkit (NLTK) version 3.5.[8] 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*.

We convert each processed document to its bag-of-words (BOW) representation using term frequency-inverse document frequency (TF-IDF) vectors which we use as inputs to the model. In a 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. 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:

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

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.

We use stratified 10-folds cross-validation wherein we split the data into 10 parts while preserving the class proportions in each split. 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

[8]NLTK is a python package used for NLP. For more details see https://www.nltk.org/.

training data, but will not generalize well. We 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.[9]

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 set. We find that F_p and M_p constructed using the method above capture gender requests well, with correlations of 0.38 and 0.44 with binary variables indicating explicit female and male requests. Appendix Figure A.3 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 also train a Bernoulli Naive Bayes (NB) classifier using text contained in job titles only, as done in Kuhn et al. (2020). We find that NB (with job title text only) does not perform as well as LR (with text contained in a job's title and description) in our context. Correlations of F_p and M_p with explicit employer requests for women and men are smaller using NB, at 0.23 and 0.22 respectively. Even when we use the full job text with NB (as opposed to using only job titles), the correlations of F_p and M_p with male and female requests improve only marginally to 0.24 and 0.26. This indicates that NB does not benefit from additional information in full job text as opposed to text contained in job titles only. A possible reason could be that NB uses only **word occurrence** rather than **word count** vectors, and therefore, might be less suitable for longer text. On the other hand, LR is better able to exploit additional information in a longer text which includes the job description as well as title for each job ad. For instance, going from only job title to full job text with LR increases the correlation of F_p with explicit female requests from 0.25 to 0.38; and of M_p with male requests from 0.33 to 0.44. Ng and Jordan (2002) discuss that LR has a lower asymptotic error, and is expected to outperform NB when the number of training examples is large even though NB converges to its (higher) asymptotic error with fewer observations. In our data comprising over 160,000 job ads, we find that LR significantly outperforms NB—particularly with complete job text.

[9] 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.

A.5 Category specific net scores

We use the Local Interpretable Model-agnostic Explanations (LIME) algorithm proposed by Ribeiro et al. (2016) to examine words in job ad text which contribute 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.[10] LIME has been used to explain predictions made by machine learning models in many applications in the biomedical domain, music content analysis, computer vision, and 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 gendered words.

We first map classification scores returned by the Multinomial Logistic Regression (LR) classifier into the input space by applying the LIME algorithm on test set documents. This allows us to assign contextual relevance scores $R^G_{i,w}$ to every word w in each job ad i which indicates the importance of that word to class $G \in \{F, N, M\}$.[11] We restrict our analyses to words that occur at least ten times in the 13,735 M and F jobs; 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 classify the 3,113 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 **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

[10] 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.

[11] 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 and 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.

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 a net score for each of the 3,113 words w and for each job ad i by category $C \in$ {hard-skills, soft-skills, personality, flexibility, others}. To do this, we use the relevance scores R_w^G for each word w towards the F and M class. The net score for a word w is just the difference between it's relevance scores for the F and M class or $NS_w = R_w^F - R_w^M$. To construct a net score for each job ad i we sum the relevance scores for words in job ad i towards the F class which are also assigned to a given category C, $S_{i,F}^C = \sum_{(w \in i) \wedge (w \in C)} R_w^F$. Similarly we sum the relevance scores for words in job ad i towards the M class which are also assigned to a given category C, $S_{i,M}^C = \sum_{(w \in i) \wedge (w \in C)} R_w^M$. We then take the difference between the two sums for each job ad i to arrive at a net score towards the F class ($NS_{C,i} = S_{i,F}^C - S_{i,M}^C$) in each category.[12] Taking this difference allows us to examine how employers distinctively associate words of a particular category in a job ad with women vs men. A positive (negative) net score for a job ad in a category indicates either that the ad contains *more words* that contribute towards a gender request for a female (male) vs a male (female) or that the words in the ad have a *higher relevance* for the female (male) vs the male (female) class.

We report summary statistics for positive and negative values of the category specific net scores in Appendix Table A.4 separately for F, N, and M jobs. Positive values of the net score have the highest mean in F jobs; for instance, $NS_{hard-skills}^+$ gets an average score of 0.25, 0.18 and 0.14 in F, N, and M jobs respectively. However, negative values of the net score do not always have the highest mean in M jobs. For instance, $NS_{hard-skills}^-$ has the highest average score in N jobs ($= 0.23$), and then in M jobs ($= 0.18$). Nevertheless, negative values of the net scores are consistently higher in M jobs than F jobs. We standardize positive and negative values of the category specific net scores for use in regression analysis for ease of interpretation.

[12] We take the difference since 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.

Table A.1: Descriptive statistics, job ads

	Prefer female	No pref.	Prefer male	Total
Education requirements:				
Other (education not specified)	0.006	0.004	0.004	0.004
None (illiterate)	0.018	0.014	0.042	0.015
Secondary education	0.113	0.099	0.322	0.108
Senior secondary education	0.318	0.263	0.259	0.265
Diploma	0.075	0.090	0.077	0.089
Graduate degree, STEM	0.034	0.089	0.054	0.086
Graduate degree, non-STEM	0.425	0.424	0.237	0.417
Postgraduate degree, STEM	0.003	0.007	0.000	0.006
Postgraduate degree, non-STEM	0.006	0.007	0.002	0.006
Experience requirements:				
$0-1$ years	0.688	0.663	0.687	0.665
$1-2$ years	0.215	0.177	0.202	0.179
>2 years	0.096	0.160	0.111	0.155
Other job requirements:				
Age requirement present	0.073	0.083	0.187	0.086
Minimum age requirement present	0.059	0.075	0.173	0.078
Maximum age requirement present	0.066	0.078	0.168	0.080
Beauty requirement present	0.118	0.057	0.060	0.059
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 Indian Rupees. Wages and experience are the mid-point of the range specified in the job ad.
Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions in Appendix A.1.1.

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
Graduate degree, STEM	0.535	0.545	0.541
Graduate degree, non-STEM	0.155	0.135	0.142
Postgraduate degree, STEM	0.122	0.067	0.087
Postgraduate degree, non-STEM	0.122	0.080	0.095
Experience:			
$0-1$ years	0.799	0.736	0.758
$1-2$ years	0.069	0.079	0.075
>2 years	0.132	0.185	0.166
Age:			
Age at registration	23.460	23.863	23.720
Applied wage:			
Mean annual wage	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 four categories to correspond to the job ads sample.
Source: The applicant sample includes those who applied to at least one job in the job ads sample (subject to the restrictions in Appendix A.1.1), 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.

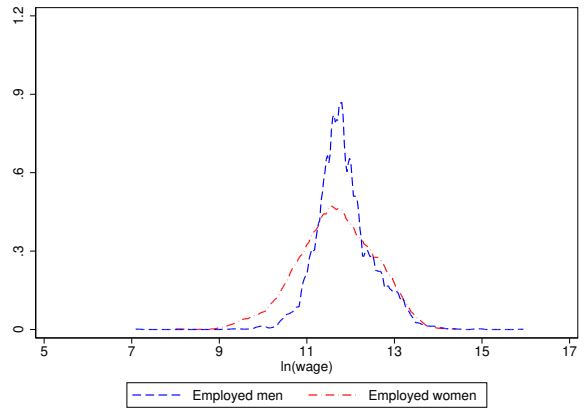
Table A.4: Descriptive statistics, net scores

	F Jobs	N Jobs	M Jobs	All jobs
$NS^{+}_{hard-skills}$	0.253	0.181	0.136	0.182
$NS^{-}_{hard-skills}$	0.097	0.225	0.180	0.218
$NS^{+}_{soft-skills}$	0.281	0.145	0.142	0.150
$NS^{-}_{soft-skills}$	0.051	0.064	0.044	0.063
$NS^{+}_{personality}$	0.128	0.077	0.068	0.078
$NS^{-}_{personality}$	0.049	0.055	0.051	0.055
$NS^{+}_{flexibility}$	0.009	0.007	0.007	0.007
$NS^{-}_{flexibility}$	0.120	0.100	0.211	0.105
NS^{+}_{others}	2.866	0.772	0.251	0.837
NS^{-}_{others}	0.187	0.651	4.303	0.761
Observations	6791	158946	6009	171746

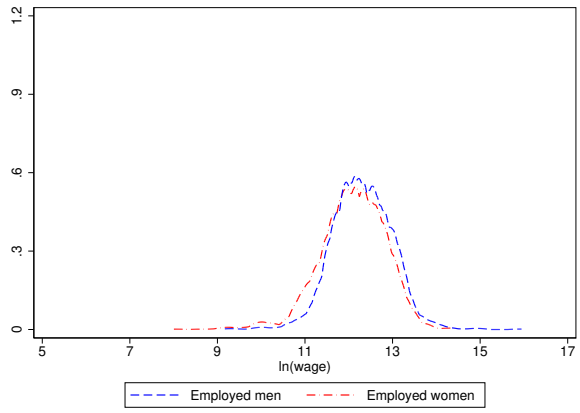
Notes: Each cell gives the average (non-standardized) magnitude of positive and negative values taken by category specific net scores in the respective sub-sample of job ads; see Appendix A.5 for details on how the category specific net scores are constructed.

Source: Data from the population of all job ads and applicants on the portal, subject to the restrictions in Appendix A.1.1.

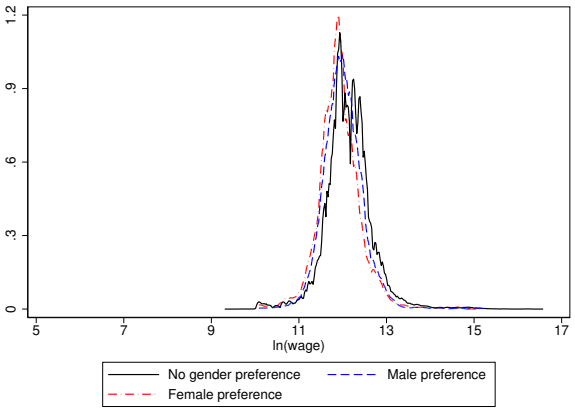
Figure A.1: Wage distributions



(a) Wage distributions by gender, PLFS



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

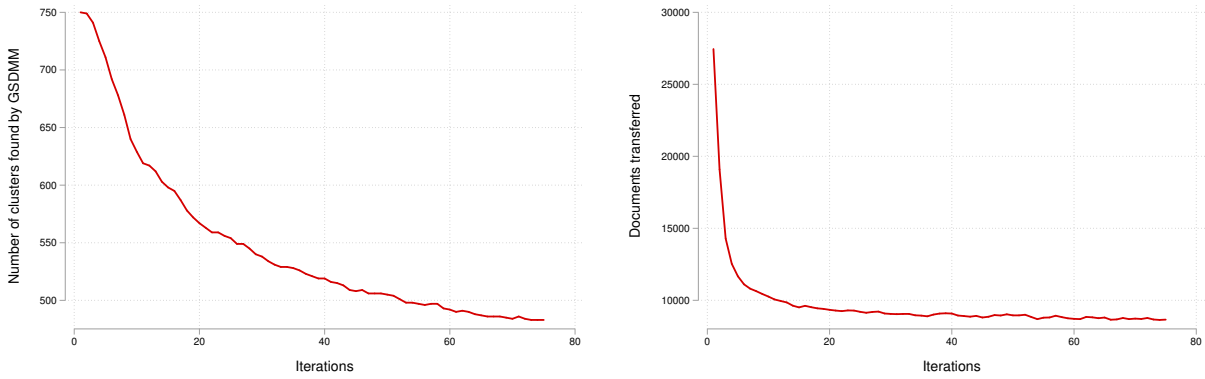


(c) Wage distributions by gender preference, job portal

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 on the portal, subject to the restrictions in Appendix A.1.1.

Figure A.2: GSDMM Iterations and Clusters

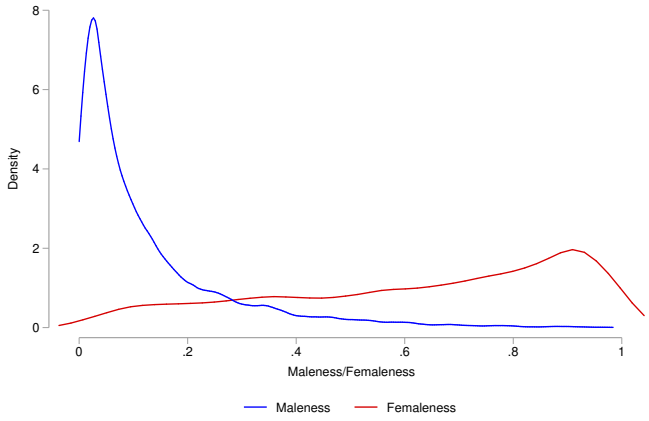


(a) Number of clusters

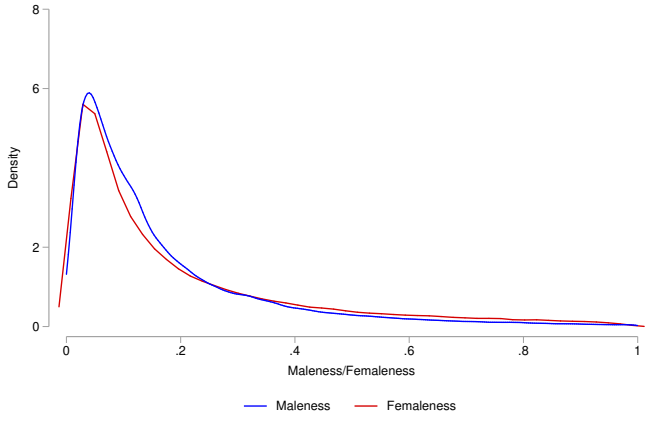
(b) Number of transferred documents

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

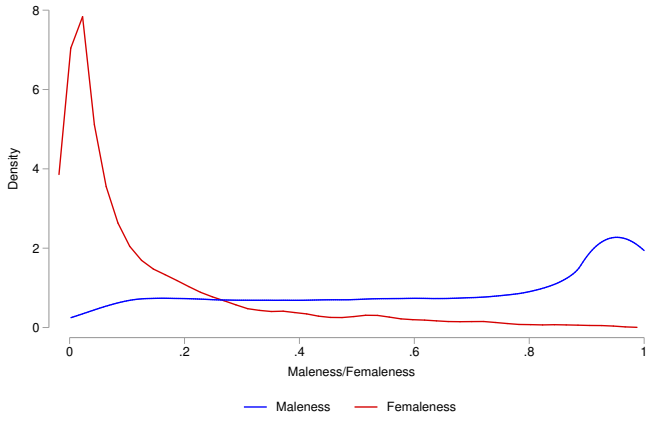
Figure A.3: Distribution of F_p and M_p in F, N and M job ads



(a) F jobs



(b) N jobs



(c) M jobs

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

Source: Data from the population of all job ads on the portal, subject to the restrictions in Appendix A.1.1.

B Appendix: Additional results

B.1 Gender requests in job ads and negative skill targeting

When examining gender requests in job ads we also study their association with age and beauty requirements in job ad text. 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 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]

To examine characteristics of jobs in which employers exhibit explicit gender preferences we estimate variations of the following regressions:

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

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 while 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

[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.

variables including dummy variables indicating education requirements, experience requirements, the presence of age and beauty requirements, and log posted wage. 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 Appendix A.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), Appendix 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 $\times$ 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 is associated with an increased probability of a job having an explicit gender preference (columns (II) and (III)).

Columns (IV)–(VI) in Appendix Table B.1 give results from estimation of equation (B.1) when the outcome of interest is male preference in a job ad or 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 is associated with 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

[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 $\times$ 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; these results are available on request.

[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.

an increased preference for men.

B.2 Applicant and match quality

In additional estimations we examine the effect of explicit gender preferences on applicant and match quality by using specifications similar to equation (3.3), but with applicant and match quality as the dependent variables of interest. Results are reported in Appendix Table B.2. We use two measures of applicant quality—completed years of schooling and the percentage marks (out of 100) obtained by a candidate in secondary school (matriculation) examination.[6] To the extent that gender requests are likely to be made in low skill jobs, applicant quality can be lower. On the other hand, gender requests may decrease or increase applicant quality by deterring otherwise highly or less qualified candidates of the non-preferred gender. We find that explicit male requests are associated with a decline in applicant quality; however, we find that explicit female requests are not associated with a reduction in applicant quality as measured by completed years of education and are actually associated with a statistically significant increase in applicant quality as measured by matriculation marks, although the effect size is economically small (columns (I) and (IV)). These associations are driven by a higher fraction of female applicants applying to jobs with an explicit female preference (on average women on the portal are more educated and have higher matriculation scores than men). Once we control for the share of female applicants, an explicit female preference has a larger negative impact on applicant quality than an explicit male preference (columns (II) and (V)). However, these effects continue to be economically very small. At the same time, we find that applicant quality improves with higher posted wages, consistent with the theory and evidence in Dal Bó et al. (2013) and Marinescu and Wolthoff (2020) (columns (III) and (VI)). We also find the effect of explicit gender requests on match quality (in terms of the share of applicants complying with the job ad's minimum education and experience requirements) to be economically small (columns (VII) and (VIII)).

B.3 The gender wage gap in applications

We examine the gender wage gap in applications by estimating the following regressions at the job application (rather than job ad) level:

[6] We do not use experience as a measure of quality since the portal mostly caters to inexperienced graduates.

$$ln\ \mathrm{W}_{ijstc} = \alpha + \delta Female_c + \beta_1 X_c + \beta_2 X_{ijstc} + \gamma_{j\times s} + \phi_t + \epsilon_{ijstc} \tag{B.2}$$

where $ln\ \mathrm{W}_{ijstc}$ is the log posted wage in job ad i advertising for a job of occupation j in state s and month-year t to which job-seeker c makes an application. $Female_c$ is a dummy variable that takes the value 1 if the applicant is female and 0 otherwise. X_c includes a set of dummy variables for the applicant's education and state of residence, as well as a quadratic in applicant's age. The coefficient of interest is δ which gives the gender wage gap in applications or the percentage difference in the application wage between female and male job-seekers after controlling for their observable characteristics. X_{ijstc} refers to job ad attributes such as the presence of gender requests, implicit gender associations in the job ad text and firm specific factors.[7] Our regressions include occupation and state fixed effects ($\gamma_{j\times s}$) as well as month-year fixed effects (ϕ_t). We cluster standard errors by occupation and state. These regressions are estimated by inversely weighting applicants by the number of applications made by them to ensure that each applicant gets an equal weight in the estimation.

Appendix Table B.3 reports estimation results for equation (B.2). Column (I) shows the gender wage gap in applications after controlling for applicant characteristics while column (II) further includes occupation and state fixed effects. We find that, on average, women apply to jobs that pay 3.7% less than comparable men. This gap reduces to 1.9% once we include occupation and state fixed effects. Column (III) further includes controls for the presence of female or male requests in a job ad; we find that this reduces the gender wage gap further to 1.6%. In column (IV) we also add controls for quartics in the implicit *femaleness* and *maleness* of a job ad as well as their interaction with explicit gender requests; the gender wage gap now falls to 1.3%. Finally, we control for firm fixed effects in column (V); the gender wage gap in applications almost disappears and only 0.2% remains.

Appendix Table B.4 shows the gender wage gap in job ads that do not have an explicit gender request (N jobs). First, as may be seen in column (I) the gender wage gap in applications for N jobs is 3% which is less than the gap for all jobs at 3.7%; this indicates the importance of gender requests

[7] Controlling for a job ad's education and experience requirements does not change the results. This is due to the high compliance by applicants to these requirements which are already controlled for in applicant's education and age. These results are available on request but are omitted for brevity.

in explaining why women send applications to lower wage jobs than comparable men. Column (II) shows that the gender wage gap falls to 1.6% after controlling for (occupation × state) fixed effects and column (III) shows that it falls to 1.2% after further controlling for quartics in implicit *femaleness* and *maleness* of a job ad. Finally, controlling for firm fixed effects almost eliminates the gender gap in column (IV).

B.4 Further results on the female applicant share in different kinds of job ads

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.[8] 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).

Appendix Figure B.1(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%.

Appendix Figure B.1(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

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

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.[9] Similarly, as M_p increases (or as we move along jobs with an increasingly male job description **within** the same 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.[10]

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.[11] Importantly, this persists even within a given occupation in a state, though the magnitudes decline.

[9] 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 (results available on request).

[10] 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.

[11] 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. 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. These results are available on request.

Table B.1: Gender requests

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)
Graduate 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)
Graduate 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)
Postgraduate 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)
Postgraduate 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 in Appendix A.1.1. Columns (II)-(III) and (V)-(VI) report the effective number of observations after incorporating (occupation × state) fixed effects, which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table B.2: Applicant and match quality

Dependent variable:	Years of education			Matriculation score			% satisfying educ. req.	% satisfying exp. req.
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
Female preference	–0.001	–0.132***	–0.128***	0.438***	–0.558***	–0.391***	–0.003***	–0.011
	(0.015)	(0.017)	(0.018)	(0.129)	(0.113)	(0.102)	(0.001)	(0.006)
Male preference	–0.130***	–0.046	–0.045	–1.004***	–0.370	–0.285	–0.002***	0.005
	(0.032)	(0.029)	(0.030)	(0.187)	(0.194)	(0.191)	(0.001)	(0.013)
% female applicants		0.847***	0.894***		6.397***	6.433***		
		(0.036)	(0.041)		(0.245)	(0.251)		
ln(wage)			0.048***			0.666***		
			(0.008)			(0.062)		
Fixed Effects	month, occ × state	month, occ × state	month, occ × state	month, occ × state	month, occ × state	month, occ × state	month, occ × state	month, occ × state
N	156168	156168	136414	155335	155335	135697	154097	154097

Notes: The dependent variable in columns (I)-(III) is the average years of education of all applicants to a job ad. The dependent variable in columns (IV)-(VI) is the average matriculation score of all applicants to a job ad. The dependent variable in column (VII) is the fraction of applicants who satisfy an ad's education requirement and in column (VIII) who satisfy an ad's experience requirement. All regressions are weighted by the total number of applications made to a job ad, and include a set of education and experience requirement controls. 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 in Appendix A.1.1. Years of education are reported in the applicant sample for almost all applicants while matriculation marks are observed for 95% of the applicants. The lower proportion of observed matriculation marks is due to some candidates not reporting these or when reporting using a CGPA scale with no information available on the conversion to percentage for such scores. All columns report the effective number of observations after incorporating (occupation × state) fixed effects which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table B.3: Gender wage gap in job applications (all jobs)

	(I)	(II)	(III)	(IV)	(V)
Female	–0.037***	–0.019***	–0.016***	–0.013***	–0.002***
	(0.008)	(0.003)	(0.003)	(0.003)	(0.001)
Fixed Effects	month,	month, occ × state	month, occ × state	month, occ × state	month, occ × state, firm
N	5327232	5327232	5327232	5327232	5325203

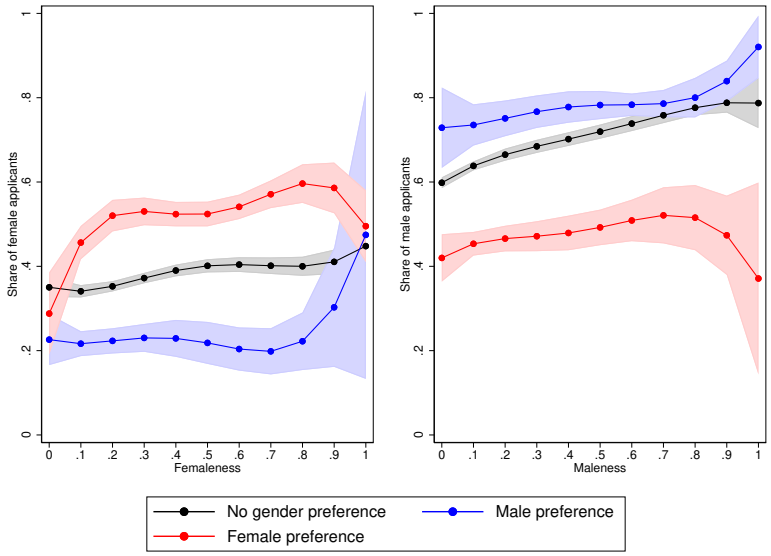
Notes: Regressions are at the application level and the dependent variable is the log of the mid-point of the wage range in the job ad to which an applicant applied. All regressions control for education, a quadratic in age and location (state of residence) of the applicant. Column (III) additionally controls for explicit gender requests in job ads while Columns (IV)-(V) additionally control for interactions of gender requests with quartics in implicit *femaleness* and *maleness*. 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 in Appendix A.1.1. Each applicant is weighted by the inverse of the total number of job applications made by her/him. All columns report the effective number of observations after incorporating fixed effects.

Table B.4: Gender wage gap in job applications (*jobs*)

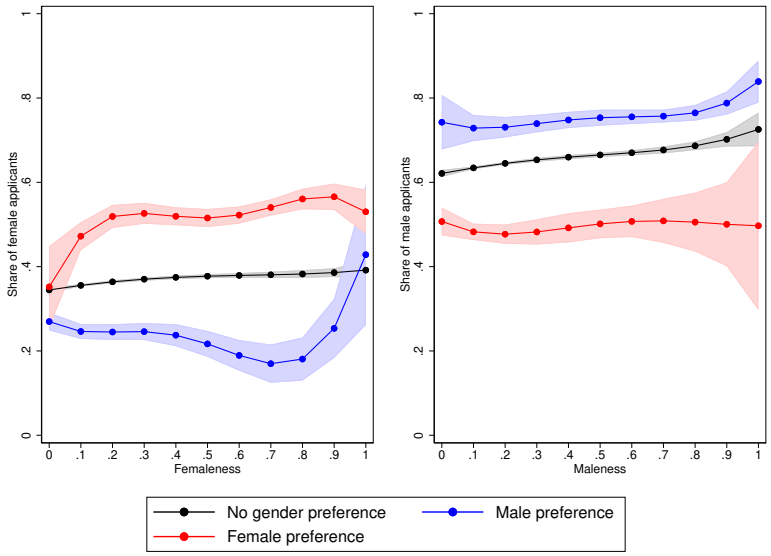
	(I)	(II)	(III)	(IV)
Female	–0.031***	–0.016***	–0.012***	–0.002***
	(0.008)	(0.003)	(0.003)	(0.001)
Fixed Effects	month,	month, occ × state	month, occ × state	month, occ × state, firm
N	5066227	5066227	5066227	5064425

Notes: Regressions are at the application level and the dependent variable is the log of the mid-point of the wage range in the job ad to which an applicant applied. All regressions control for education, a quadratic in age and location (state of residence) of the applicant. Columns (III)-(IV) also control for quartics in implicit *femaleness* and *maleness*. 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 N job ads and applicants to these jobs on the portal, subject to the restrictions in Appendix A.1.1. Each applicant is weighted by the inverse of the total number of job applications made by her/him. All columns report the effective number of observations after incorporating fixed effects.

Figure B.1: 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 Appendix A.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. All 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 in Appendix A.1.1.

C Appendix: Robustness checks

We examine the robustness of our results to several modifications, as discussed below.

C.1 Including controls for applicant characteristics

We 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 1 if an applicant to a job ad is female and 0 if it is a male. Using these regressions we are able to control for applicant characteristics such as 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 C.1). 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 positive and negative values of category specific net scores on the probability that a female applies to a job and find that our results on hard-skills and job flexibility related words for N jobs persist (Appendix Table C.2).

C.2 Using an alternative method of constructing occupation categories

We carry out all estimations using a more dis-aggregate n-gram based occupational classification (with 747 occupation categories) derived from the job title of an ad as described in Appendix A.3; we find that our results are robust. In wage regressions that use the sample of N jobs, we find that the decrease in posted 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 C.3). 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 C.4). 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 posted wage and an increase in female applicant share with higher $NS^{+}_{hard-skills}$ as well as an

increase in the posted wage and a decrease in female applicant share with higher $NS^{-}_{flexibility}$ for N jobs (columns (I) and (IV), Appendix Table C.5).

C.3 Including firm fixed effects

We 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.[1] We continue to find that our results are largely robust. We still find that higher F_p has a larger negative effect on the log posted 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 0.152 with (firm × occupation × state) fixed effects (columns (II) and (III), Appendix Table C.3). 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 C.4). Lastly, our results on the positive (negative) effect of $NS^{-}_{flexibility}$ ($NS^{+}_{hard-skills}$) on wages and the negative (positive) effect of these variables on female applicant share for N jobs are largely robust (columns (II)–(III) and (V)–(VI), Appendix Table C.5). We see a significant decrease in the female applicant share with higher $NS^{-}_{flexibility}$ in job ads posted by the same firm for the same occupation. However, the coefficients on $NS^{+}_{hard-skills}$ are now insignificant, albeit still positive.

C.4 Using an alternative specification with quartics in net scores

We also check the robustness of the results in Section 4.2 to using an alternative specification where we use quartics in the category specific net scores (NS^{k}_{C} for $k = 1, .., 4$ and each C) rather than positive and negative values of these scores (NS^{+}_{C} and NS^{-}_{C} for each C).

Instead of equation (4.1) we estimate:

$$\ln W_{ijst} = \kappa^{W} + \sum_{C} \sum_{k=1}^{4} \nu^{Ck} NS^{k}_{C,ijst} + \omega^{W} X_{ijst} + \gamma_{j\times s} + \phi_{t} + \zeta_{ijst} \tag{C.1}$$

where $\ln W_{ijst}$ is the log of the posted wage in job ad i advertising for a job of occupation j in state

[1] In Appendix Tables C.3–C.5, 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.

s and month-year t; apart from the way in which category specific net scores are specified, equation (C.1) is identical to equation (4.1). Estimation results are reported in Appendix Table C.6 while Appendix Figure C.1 plots the predicted wage as the category specific net scores vary using the estimated results for N jobs. As may be seen, the results continue to show a negative association of $NS_{hard-skills}$ and $NS_{flexibility}$ with the log posted wage.

Similarly, instead of equation (4.2) we estimate:

$$Y^S_{ijst} = \kappa^S + \sum_C \sum_{k=1}^{4} \chi^{Ck} NS^k_{C,ijst} + \omega^S X_{ijst} + \gamma_{j\times s} + \phi_t + \varsigma_{ijst} \tag{C.2}$$

where Y^S_{ijst} is the share of female applicants to job ad i. As before, apart from the way in which category specific net scores are specified, equation (C.2) is identical to equation (4.2). Estimation results are reported in Appendix Table C.7 while Appendix Figure C.2 shows the predicted share of female applicants as the category specific net scores vary using the estimation results for N jobs. While $NS_{hard-skills}$ and $NS^2_{hard-skills}$ are associated with a higher female applicant share in N jobs, the coefficient on $NS_{flexibility}$ in these jobs is large and positive but just misses statistical significance at the 5% level (p-value= 0.068). The positive association of $NS_{hard-skills}$ and $NS_{flexibility}$ with the predicted share of female applicants can also be seen in Appendix Figure C.2.

Table C.1: Gender requests and female applicants

	(I)	(II)	(III)
Female preference (F_e)	0.204***	0.167***	0.166***
	(0.012)	(0.006)	(0.006)
Male preference (M_e)	–0.117***	–0.089***	–0.092***
	(0.007)	(0.006)	(0.005)
Education requirements:			
Senior secondary	0.020***	0.009***	0.009***
	(0.004)	(0.002)	(0.002)
Diploma	–0.040***	–0.006	–0.005
	(0.007)	(0.004)	(0.004)
Graduate degree, STEM	0.010	0.009**	0.007
	(0.010)	(0.004)	(0.005)
Graduate degree, non-STEM	0.049***	0.019***	0.020***
	(0.005)	(0.003)	(0.003)
Postgraduate degree, STEM	0.068***	0.040***	0.048***
	(0.009)	(0.009)	(0.012)
Postgraduate degree, non-STEM	0.076***	0.044***	0.053***
	(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 the value 1 if an applicant to a job ad is female and is 0 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 in Appendix A.1.1. All columns report the effective number of observations after incorporating (occupation × state) fixed effects which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table C.2: Net scores and female applicants

Sample:	*F* Jobs		*N* Jobs		*M* Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
$NS^{+}_{hard-skills}$	0.004	0.006**	0.010***	0.004***	0.001	0.006
	(0.004)	(0.003)	(0.002)	(0.001)	(0.003)	(0.003)
$NS^{+}_{soft-skills}$	0.001	–0.001	0.005***	0.002***	0.005	–0.001
	(0.003)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
$NS^{+}_{personality}$	0.001	–0.002	0.002	0.001	0.005	0.003
	(0.003)	(0.002)	(0.001)	(0.001)	(0.003)	(0.002)
$NS^{+}_{flexibility}$	–0.001	0.000	0.001	0.000	–0.000	0.001
	(0.003)	(0.002)	(0.001)	(0.000)	(0.003)	(0.001)
NS^{+}_{others}	0.007**	–0.001	0.016***	0.007***	0.026***	0.014***
	(0.003)	(0.002)	(0.002)	(0.001)	(0.006)	(0.005)
$NS^{-}_{hard-skills}$	–0.048***	–0.019***	0.004*	–0.001	0.010***	0.002
	(0.009)	(0.006)	(0.002)	(0.001)	(0.002)	(0.003)
$NS^{-}_{soft-skills}$	–0.005	0.002	0.001	–0.001	–0.002	0.001
	(0.006)	(0.003)	(0.001)	(0.001)	(0.004)	(0.003)
$NS^{-}_{personality}$	0.000	–0.000	0.000	–0.001	0.004	–0.003
	(0.006)	(0.003)	(0.001)	(0.001)	(0.003)	(0.002)
$NS^{-}_{flexibility}$	–0.019***	–0.008***	–0.007***	–0.007***	0.002	–0.007**
	(0.003)	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)
NS^{-}_{others}	–0.052***	–0.031***	–0.021***	–0.010***	–0.010***	–0.006***
	(0.012)	(0.007)	(0.003)	(0.001)	(0.001)	(0.002)
Fixed Effects	month	month, occ × state	month	month, occ × state	month	month, occ × state
N	112876	112876	6115802	6115802	173188	173188

Notes: The dependent variable is a dummy variable that takes the value 1 if an applicant to a job ad is female and is 0 otherwise. Coefficients on positive and negative values of the category specific net scores are reported; see Appendix A.5 for details on how the category specific net scores are constructed. All regressions control for a set of education and experience requirement categories given in a job ad. All regressions also 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 in Appendix A.1.1. All columns report the effective number of observations after incorporating (occupation × state) fixed effects which exclude job ads for which there is no variation in the dependent variable within an (occupation × state) cell.

Table C.3: Wages

	(I)	(II)	(III)
Implicit *femaleness* (F_p)	–0.225***	–0.283***	–0.127***
	(0.013)	(0.019)	(0.018)
Implicit *maleness* (M_p)	–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)
Graduate degree, STEM	0.145***	0.143***	0.107***
	(0.011)	(0.028)	(0.019)
Graduate 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 N job ads on the portal which advertise a wage, subject to the restrictions in Appendix A.1.1. 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 (alternative occupation × state), (firm × state) or (firm × occupation × state) cell, depending on the fixed effects used.

Table C.4: Applications

Dependent variable:	total applications			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Female preference (F_e)	–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 (M_e)	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)
Graduate 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)
Graduate 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 in Appendix A.1.1. 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 (alternative occupation × state), (firm × state) or (firm × occupation × state) cell, depending on the fixed effects used.

Table C.5: Net scores, advertised wages and the share of female applicants

Dependent variable:	log of advertised wage			share of female applications		
	(I)	(II)	(III)	(IV)	(V)	(VI)
$NS^{+}_{hard-skills}$	–0.011***	–0.010***	–0.006*	0.002***	0.009***	0.002
	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
$NS^{+}_{soft-skills}$	–0.002	–0.005***	–0.001	0.001*	0.004***	0.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
$NS^{+}_{personality}$	0.004***	–0.000	–0.003	0.001	0.003**	0.001
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
$NS^{+}_{flexibility}$	0.001	–0.002	0.002	0.000	–0.001	0.000
	(0.002)	(0.003)	(0.002)	(0.000)	(0.001)	(0.001)
NS^{+}_{others}	–0.020***	–0.039***	–0.015***	0.004***	0.015***	0.003*
	(0.002)	(0.003)	(0.003)	(0.001)	(0.002)	(0.001)
$NS^{-}_{hard-skills}$	0.006***	0.012***	0.002	–0.000	–0.001	–0.001
	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
$NS^{-}_{soft-skills}$	0.009***	0.004	–0.001	0.001	–0.001	0.000
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
$NS^{-}_{personality}$	0.005***	–0.000	–0.008***	0.000	–0.002	–0.001
	(0.002)	(0.002)	(0.002)	(0.000)	(0.001)	(0.001)
$NS^{-}_{flexibility}$	0.016***	0.009***	0.007***	–0.005***	–0.009***	–0.004***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
NS^{-}_{others}	0.000	–0.000	–0.011***	–0.006***	–0.024***	–0.007***
	(0.003)	(0.003)	(0.004)	(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	121931	74729	42059	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. Coefficients on positive and negative values of the category specific net scores are reported; see Appendix A.5 for details on how the category specific net scores are constructed. All regressions control for a set of education and experience requirement categories given in a job ad. 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 N job ads on the portal which advertise a wage, subject to the restrictions in Appendix A.1.1. 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 C.6: Net scores and wages

Sample:	*F* Jobs		*N* Jobs		*M* Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
$NS_{hard-skills}$	–0.019	–0.013	–0.035***	–0.018***	–0.010	–0.020*
	(0.012)	(0.012)	(0.004)	(0.003)	(0.013)	(0.010)
$NS^2_{hard-skills}$	–0.022***	–0.014***	–0.008***	–0.002	–0.011***	–0.006***
	(0.005)	(0.005)	(0.002)	(0.001)	(0.002)	(0.002)
$NS^3_{hard-skills}$	–0.000	–0.000	0.001*	0.000	0.000	–0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
$NS^4_{hard-skills}$	0.000*	0.000*	0.000*	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS_{soft-skills}$	–0.001	–0.002	–0.011***	–0.009***	–0.004	–0.012
	(0.008)	(0.007)	(0.003)	(0.002)	(0.012)	(0.011)
$NS^2_{soft-skills}$	–0.003	–0.002	0.003***	0.001	0.002	0.002
	(0.003)	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)
$NS^3_{soft-skills}$	0.000	0.000	–0.000	–0.000	–0.001	–0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
$NS^4_{soft-skills}$	0.000	–0.000	–0.000	–0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS_{personality}$	0.002	0.002	0.007	–0.002	0.023*	0.009
	(0.011)	(0.009)	(0.004)	(0.002)	(0.011)	(0.010)
$NS^2_{personality}$	0.001	0.001	0.003***	0.001	0.001	–0.003
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
$NS^3_{personality}$	0.000	0.000	0.000	0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS^4_{personality}$	–0.000	–0.000	–0.000***	–0.000***	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS_{flexibility}$	–0.065***	–0.050***	–0.049***	–0.028***	–0.021	–0.018
	(0.016)	(0.018)	(0.006)	(0.005)	(0.019)	(0.016)
$NS^2_{flexibility}$	0.005	–0.003	–0.000	0.000	–0.003	–0.003
	(0.007)	(0.008)	(0.003)	(0.003)	(0.006)	(0.005)
$NS^3_{flexibility}$	0.002	0.000	0.001	0.001	–0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
$NS^4_{flexibility}$	0.000	0.000	0.000*	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
NS_{others}	–0.043***	–0.029*	–0.062***	–0.026***	–0.009	0.008
	(0.014)	(0.014)	(0.006)	(0.004)	(0.015)	(0.013)
NS^2_{others}	–0.000	–0.008	–0.021***	–0.015***	0.012	–0.002
	(0.007)	(0.006)	(0.002)	(0.002)	(0.008)	(0.008)
NS^3_{others}	0.003	0.002	0.002***	0.001***	0.003	–0.000
	(0.002)	(0.001)	(0.000)	(0.000)	(0.003)	(0.002)
NS^4_{others}	–0.000	0.000	0.001***	0.000***	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
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 education and experience requirements in 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.

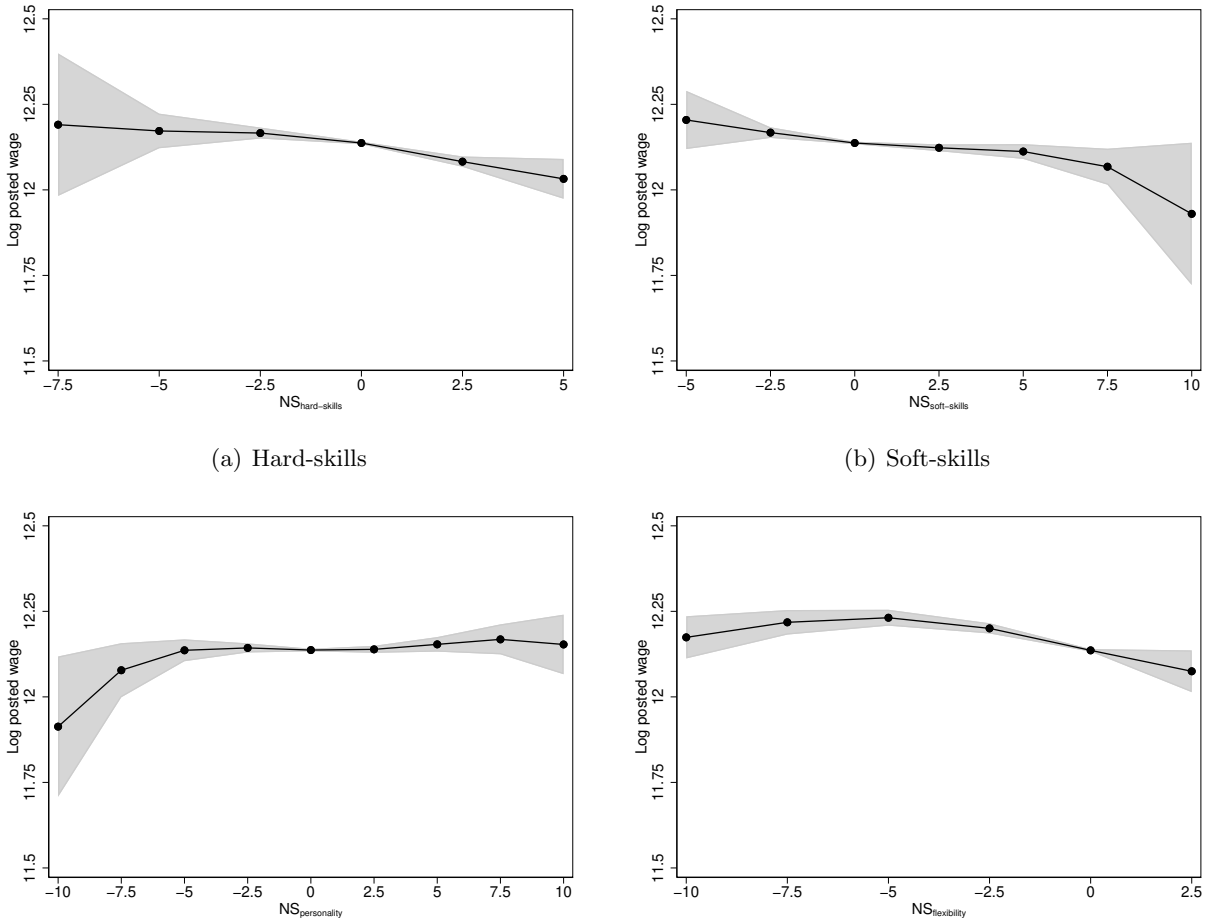
*Source:*Data from the population of all job ads and applicants on the portal, subject to the restrictions in Appendix A.1. All columns report the effective number of observations after incorporating (occupation × state) fixed effects.

Table C.7: Net scores and the share of female applications

Sample:	*F* Jobs		*N* Jobs		*M* Jobs	
	(I)	(II)	(III)	(IV)	(V)	(VI)
$NS_{hard-skills}$	0.036***	0.014**	0.001	0.004***	–0.024***	–0.014**
	(0.008)	(0.006)	(0.002)	(0.001)	(0.007)	(0.006)
$NS^2_{hard-skills}$	–0.009**	–0.003	0.005***	0.001*	–0.001	–0.000
	(0.004)	(0.003)	(0.001)	(0.000)	(0.001)	(0.001)
$NS^3_{hard-skills}$	–0.002*	–0.001	0.000***	–0.000	0.001	0.001
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
$NS^4_{hard-skills}$	0.000***	0.000**	–0.000***	–0.000	0.000	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS_{soft-skills}$	0.008	–0.002	0.005**	0.002	0.014**	0.004
	(0.007)	(0.004)	(0.002)	(0.001)	(0.006)	(0.005)
$NS^2_{soft-skills}$	–0.005	–0.001	0.001	0.000	–0.002	–0.001
	(0.002)	(0.001)	(0.001)	(0.000)	(0.002)	(0.001)
$NS^3_{soft-skills}$	–0.000	–0.000	–0.000	–0.000	–0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
$NS^4_{soft-skills}$	0.000***	0.000	0.000	0.000	0.000	–0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS_{personality}$	0.003	0.003	0.001	0.002**	0.006	0.008**
	(0.006)	(0.004)	(0.002)	(0.001)	(0.005)	(0.003)
$NS^2_{personality}$	0.001	0.000	0.000	–0.000	0.002	–0.001
	(0.002)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
$NS^3_{personality}$	–0.000	–0.000	–0.000	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS^4_{personality}$	0.000	0.000	–0.000	0.000	–0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$NS_{flexibility}$	0.037***	0.019	0.007	0.007	–0.012	0.009
	(0.013)	(0.010)	(0.004)	(0.004)	(0.009)	(0.005)
$NS^2_{flexibility}$	–0.004	–0.000	0.000	–0.000	–0.004	–0.000
	(0.005)	(0.004)	(0.002)	(0.001)	(0.005)	(0.003)
$NS^3_{flexibility}$	–0.001	–0.000	0.000	–0.000	–0.001	–0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
$NS^4_{flexibility}$	–0.000	–0.000	0.000	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
NS_{others}	0.056***	0.017*	0.038***	0.013***	0.042***	0.015***
	(0.012)	(0.009)	(0.005)	(0.002)	(0.008)	(0.005)
NS^2_{others}	–0.008	–0.007*	–0.006*	–0.001	–0.001	0.001
	(0.005)	(0.004)	(0.003)	(0.001)	(0.006)	(0.003)
NS^3_{others}	–0.003**	0.000	0.000	0.000	–0.001	–0.000
	(0.001)	(0.001)	(0.000)	(0.000)	(0.002)	(0.001)
NS^4_{others}	0.000	0.000	0.000	0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
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 education and experience requirements in a job ad and are weighted by the total number of applications made to the 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 in Appendix A.1. All columns report the effective number of observations after incorporating (occupation × state) fixed effects.

Figure C.1: Net scores and predicted wages



(a) Hard-skills

(b) Soft-skills

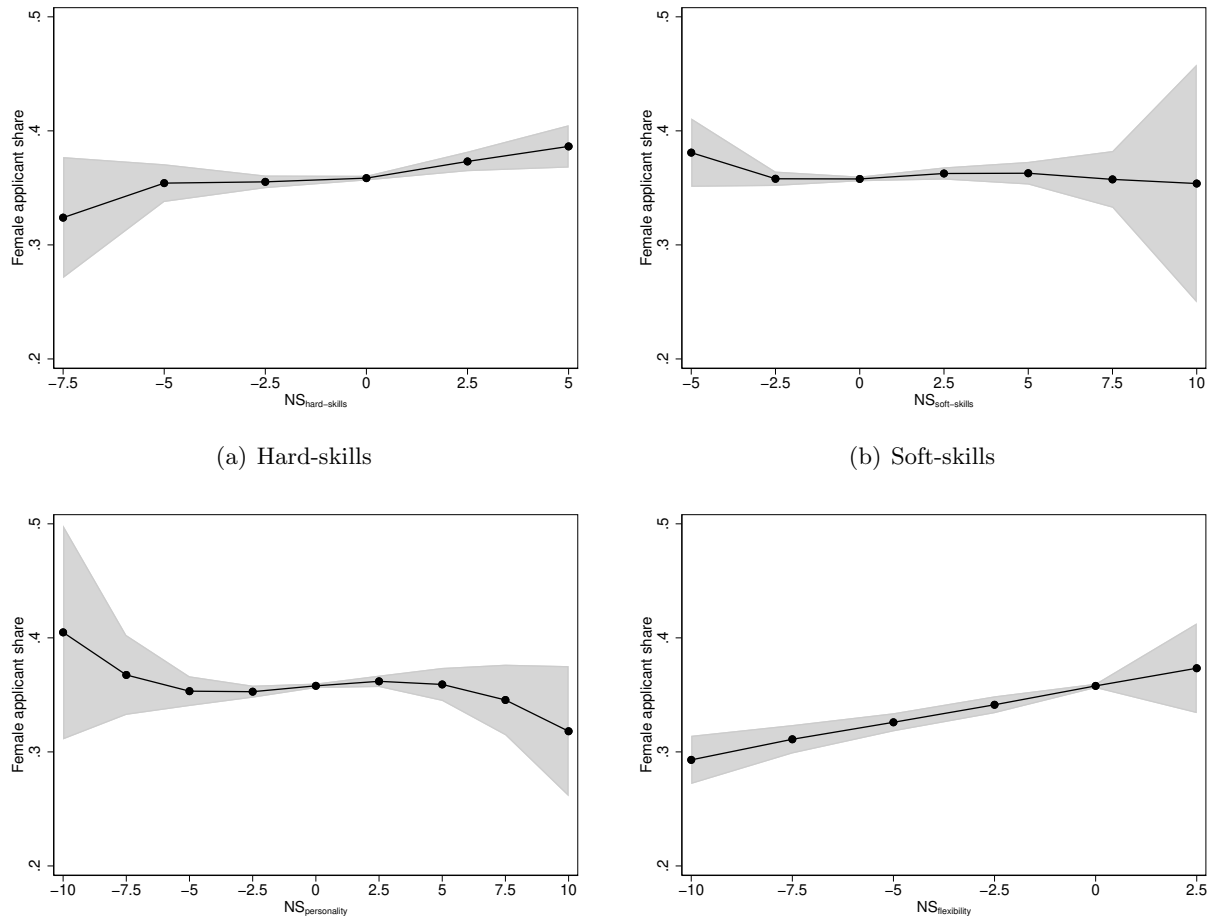
(c) Personality

(d) Flexibility

Notes: Shaded areas give the 95% confidence intervals around predicted values. Each sub-figure shows the predicted log posted wage as the category specific net score on the x-axis varies using the regression specification given in equation (C.1); the values of all variables apart from the category specific net score on the x-axis are held constant at their mean values.

Source: Data from the population of N job ads on the portal, subject to the restrictions in Appendix A.1.1.

Figure C.2: Net scores and the predicted share of female applicants



(a) Hard-skills

(b) Soft-skills

(c) Personality

(d) Flexibility

Notes: Shaded areas give the 95% confidence intervals around predicted values. Each sub-figure shows the predicted fraction of female applicants to a job ad as the category specific net score on the x-axis varies using the regression specification given in equation (C.2); the values of all variables apart from the category specific net score on the x-axis are held constant at their mean values.

Source: Data from the population of N job ads on the portal, subject to the restrictions in Appendix A.1.1.

Supplementary References

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.

Dal Bó, E., F. Finan, and M. Rossi (2013): "Strengthening state capabilities: The role of financial incentives in the call to public service," *Quarterly Journal of Economics*, 128(3), 1169–218.

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.

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.

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.

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.