Chaturvedi, Sugat; Mahajan, Kanika; Siddique, Zahra

Working Paper

Using Domain-Specific Word Embeddings to Examine the Demand for Skills

IZA Discussion Papers, No. 16593

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Chaturvedi, Sugat; Mahajan, Kanika; Siddique, Zahra (2023) : Using Domain-Specific Word Embeddings to Examine the Demand for Skills, IZA Discussion Papers, No. 16593, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/282720

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16593

Using Domain-Specific Word Embeddings to Examine the Demand for Skills

Sugat Chaturvedi
Kanika Mahajan
Zahra Siddique

NOVEMBER 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16593

Using Domain-Specific Word Embeddings to Examine the Demand for Skills

Sugat Chaturvedi
Ahmedabad University

Kanika Mahajan
Ashoka University

Zahra Siddique
University of Bristol and IZA

NOVEMBER 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Using Domain-Specific Word Embeddings to Examine the Demand for Skills*

We study the demand for skills by using text analysis methods on job descriptions in a large volume of ads posted on an online Indian job portal. We make use of domain-specific unlabeled data to obtain word vector representations (i.e., word embeddings) and discuss how these can be leveraged for labor market research. We start by carrying out a data-driven categorization of required skill words and construct gender associations of different skill categories using word embeddings. Next, we examine how different required skill categories correlate with log posted wages as well as explore how skills demand varies with firm size. We find that female skills are associated with lower posted wages, potentially contributing to observed gender wage gaps. We also find that large firms require a more extensive range of skills, implying that complementarity between female and male skills is greater among these firms.

JEL Classification: J16, J23, J31, J63, J71, L2

Keywords: text analysis, online job ads, gender, skills demand, machine learning

Corresponding author:
Zahra Siddique
University of Bristol
Senate House
Tyndall Avenue
Bristol BS8 1TH
United Kingdom

E-mail: zahra.siddique@bristol.ac.uk

* We are grateful to the editor Solomon W. Polachek at Research in Labor Economics and two anonymous referees for their constructive and helpful comments as well as Medha Chatterjee for excellent research assistance. Kanika Mahajan acknowledges financial support from the CEDA-BMGF Grant at Ashoka University.

Using Domain-Specific Word Embeddings to Examine the Demand for Skills*

Sugat Chaturvedi† Ahmedabad University

Kanika Mahajan‡ Ashoka University

Zahra Siddique§ University of Bristol

November 2023

Abstract

We study the demand for skills by using text analysis methods on job descriptions in a large volume of ads posted on an online Indian job portal. We make use of domain-specific unlabeled data to obtain word vector representations (i.e., word embeddings) and discuss how these can be leveraged for labor market research. We start by carrying out a data-driven categorization of required skill words and construct gender associations of different skill categories using word embeddings. Next, we examine how different required skill categories correlate with log posted wages as well as explore how skills demand varies with firm size. We find that female skills are associated with lower posted wages, potentially contributing to observed gender wage gaps. We also find that large firms require a more extensive range of skills, implying that complementarity between female and male skills is greater among these firms.

Keywords: Text analysis; online job ads; gender; skills demand; machine learning

JEL Codes: J16; J23; J31; J63; J71; L25

*We are grateful to the editor Solomon W. Polachek at Research in Labor Economics and two anonymous referees for their constructive and helpful comments as well as Medha Chatterjee for excellent research assistance. Kanika Mahajan acknowledges financial support from the CEDA-BMGF Grant at Ashoka University.

†Amrut Mody School of Management, Ahmedabad University, Central Campus, Navrangpura, Ahmedabad, 380009, Gujarat, India. Email: sugat.chaturvedi@ahduni.edu.in

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

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

1 Introduction

How does the demand for skills vary over time and across space? Is there a gender element to skills demand i.e., do employers associate some skills with men and others with women? Do larger firms demand a greater variety and different types of skills? These questions are important since they allow a better understanding of how labor markets function and how they respond to different kinds of shocks (such as the Covid pandemic). Our work demonstrates how these questions can be investigated using text analysis and machine learning methods. In developing countries such as India, the setting in which we carry out our analysis, administrative data on labor demand is often not available so we use data from 332,044 job ads posted on an online job portal by employers between 2020 and 2022. These ads recruit for high skill jobs in the services sector which are likely to be located in large Indian cities such as Bangalore.

We focus on the skill requirements in online job ads. We study underlying gender associations for these skills by using word embeddings (described in Section 3.1); these allow a representation of words in job ads as vectors which incorporate information on local co-occurrence patterns, so that similar words have proximate vectors. We use the HDBSCAN algorithm (described in Section 3.2) to categorise skill requirements into thirty seven aggregate categories. We examine and compare domain-specific word embeddings with word embeddings trained on Common Crawl on the web and Wikipedia. Next, we use results from domain-specific word embeddings in standard Mincer regressions to examine the association of log offered wages with different kinds of required skills.[1] We find that job ads which require female skills tend to offer lower wages; this has implications for observed gender wage gaps in the labor market which are relatively large in the Indian setting. We also merge job ads data with firm data, for firms which post the ads. This allows us to examine how demand for different kinds of skills varies with firm size. We find that complementarity between female and male skills is higher among larger firms.

Our comparison of domain-specific word embeddings with word embeddings trained on Common Crawl on the web and Wikipedia demonstrates the benefits of using the former.[2] As opposed to pre-trained word embeddings trained on Common Crawl on the web and

[1] These regression estimates give correlations and not causal estimates since we do not claim to control for all possible confounding factors which could impact wages; we discuss this in more detail in Section 3.4.

[2] It may not always be possible to construct domain specific word embeddings; for instance, it could be the case that the corpus is not sufficiently large to learn reliable vectors. In such instances, it may still be possible to use a related but different corpus which is sufficiently large so that one can get pre-trained word embeddings, and then use these in one's application. This is also referred to as transfer learning, and is the strategy adopted by Hansen *et al.* (2021).

Wikipedia, domain-specific word embeddings capture the meaning and semantic relationships for the specific corpora on which they are trained—in our case, online job descriptions. This is important since job ads may have their specific vocabulary so the language in job descriptions may be substantially different in comparison to other domains. For example, the word *Python* refers to the programming language in online job descriptions but refers to both the genus of snakes and the programming language in the Common Crawl data.[3] We provide a trained model for use on our sample of online Indian job ads which are in the English language; however, the methods we employ are language-agnostic and may also be used for non-English job ads.[4]

We also provide applications of how a data-driven categorization of skills and gender associations of specific skills using domain-specific word embeddings can be used to examine questions of interest in labor economics. For instance, earnings inequality across different demographic groups (e.g., men and women) and across different local labor markets is an important concern for policymakers everywhere. The Indian Periodic Labour Force Survey (PLFS) 2020-21 shows that, after accounting for differences in observable demographic characteristics, urban women age 15–59 earn 21% less than urban men even when working within the same occupation and location. Given the large and persistent gender inequalities in labor market earnings, understanding and reducing these inequalities is a very important concern for Indian policymakers.

We examine the role played by labor demand in such inequalities by making use of wage data contained in online job ads and estimating Mincer regressions to examine how different kinds of required skills correlate with log posted wages. We find that female skills tend to be associated with low log wages (e.g. *Front office support*, *Counselling*, *Recruitment*, *Language skills*).[5] In contrast, we find that male skills tend to be associated with high log wages (mostly related to Information Technology, e.g. *Machine Learning Solution/Technical architect*, *Software development and coding*, etc.). In wage regressions that do not control for occupation but include state fixed effects and other controls, we find that more female skills in a job ad or an increase in net female association by one unit is associated with a **reduction** in the posted wage within the ad by 2.2%. This is attenuated to 0.4% once we control for 300 dis-aggregate occupation fixed effects but remains statistically significant at

[3]Therefore, the most similar words to *Python* in the pre-trained `fastText` model include *snake* and *reptile* along with libraries from python programming language, while in the case of job ad-specific `fastText` only Python and associated programming languages are among the most similar words. Similarly, the usage of the words *plant* or *hibernate* in online job descriptions might differ from their general usage.

[4]The model and code are available at https://github.com/sugatc/job-skills.

[5]A caveat is that we can only observe wages for about a third of the job ads in our sample, and job ads with non-missing wages are relatively low skill jobs.

the 10% level. We continue to find a negative relationship which is significant at the 10% level after controlling for even more dis-aggregate occupation fixed effects, with up to 400 categories.

Our second application is an exploration of the relationship between firm size and skills demand. The set of skills that a firm requires, and therefore its productivity, depends on its scale of production.[6] In a recent paper, Adenbaum (2022) extend the task based approach by Acemoglu & Autor (2011) and Ocampo (2022) to show that larger firms are more likely to hire specialized workers and, hence, hire across a range of occupations and a dispersed set of skills. We test this hypothesis using text data contained in online job ads after merging this with data on firm size. We find that larger firms post ads requesting a larger number of skills; a one percent increase in firm size (as proxied by more postings and vacancies by a firm) increases the number of skills demanded by $0.32 - 0.43\%$.

We also examine whether the demand for **both** male and female skills increases with firm size. Recent evidence suggests that women in India are more likely to be employed in larger firms (Chakraborty & Mahajan, 2023). This is also consistent with prior research which shows a positive correlation between firm size and the proportion of female employees in the U.S. (Mitra, 2003). This might be due to higher demand for female workers by larger firms arising from differences in skill requirements by firm size. Chakraborty & Mahajan (2023) show that larger firms pay additional benefits like maternity and paid leave which are valued by female employees. They propose that one reason for offering these benefits and incurring additional costs by larger firms could be that the skills associated with women are complementary to those associated with men for larger firms.[7] If male and female skills are more complementary in larger firms then the joint demand for both skills should increase as firm size increases. We test this hypothesis using several different measures of firm size and consistently find that larger firms are more likely to request both male and female skills in the job ads they post online; a one percent increase in the number of postings and vacancies for a firm increases the probability that it demands both male and female associated skills by $0.35 - 0.5\%$. We take this as an indication that these two different kinds of skills are likely to be complementary for larger firms. We find a similar positive elasticity when using paid up capital as a measure of firm size.

Our work contributes to a growing literature that uses job portal data to study the labor

[6]For instance, a large firm specializing in Information Technology may also demand specialized Human Resource personnel whereas a small firm would either outsource this task or have someone who does not necessarily specialize in this area take up the role in a partial capacity.

[7]Aquilina *et al.* (2006) show that there exists a negative relationship between firm size and the elasticity of substitution between different factors of production. While their study uses labor and capital as the factors, their theory can be plausibly extended to different types of labor.

market (Kuhn & Shen, 2013; Hershbein & Kahn, 2018; Deming & Kahn, 2018; Marinescu & Wolthoff, 2020). It illustrates the use of text analysis and machine learning methods to study economic questions (Gentzkow *et al.*, 2019; Ash & Hansen, 2023 provide reviews). Specifically, it contributes to the literature that examines the demand for skills (Beaudry, 2016; Deming, 2017; Deming & Kahn, 2018; Hansen *et al.*, 2021 are a few recent papers). While much of the economics literature uses a dictionary or keywords based approach for categorization of skills into pre-defined categories, word embeddings can be used to provide a context-sensitive categorization.[8]

Word embeddings can also be used to capture gender associations in text corpora that might reflect cultural stereotypes (Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2017; Ash *et al.*, forthcoming). Chaturvedi *et al.* (2021) show that more women apply to job ads when the ad uses words predictive of a female preference by the employer. Several other papers also provide evidence that women's application decisions are affected by the wording in job ads (Abraham & Stein, 2020; Kuhn *et al.*, 2020). In this paper we find that when employers use required skill words with a higher female association in ads, the average posted wage tends to be lower. This can potentially contribute to a gender wage gap at the application stage of hiring if women are more likely to apply to jobs which demand such skills. Our analyses, thus, also contributes to the large literature that examines gender wage gaps (Olivetti & Petrongolo, 2016; Blau & Kahn, 2017 provide reviews).

We also directly extend the recent theoretical work by Adenbaum (2022) which shows that larger firms demand more varieties of skills. We also provide empirical evidence for the theoretical foundations in Chakraborty & Mahajan (2023) who propose a task based explanation that leads to greater relative demand for women in larger firms. It shows that higher relative productivity of women in these new tasks leads to an increase in firm's demand for female labor. We provide direct empirical evidence for these theoretical observations by exploiting textual information within job ads.

In the next Section we describe the job portal data we use. Section 3 starts by providing details of the methods we employ, including domain-specific word embeddings and skills categories. We discuss our regression specifications for examining different aspects of skills demand in the second half of Section 3. In Section 4 we discuss our results, while Section 5 concludes.

[8] Ao *et al.* (2023) use wage regressions to show that a skills categorization based on unsupervised machine learning can explain a higher fraction of wage variation across jobs than dictionary based methods.

2 Data

We use data from 332,044 English language job ads posted on the National Career Services (NCS) portal which is managed by the Ministry of Labor & Employment of the Government of India.[9] The NCS portal was launched on July 20, 2015 with the goal of connecting over 950 employment exchanges in India and to provide a free alternative to large Indian online job portals run by companies in the private sector. The portal emphasizes improving employment opportunities for youth and female workers, and caters to applicants from diverse backgrounds.[10] It also provides career counselling, vocational guidance, and skills training as well as conducting job fairs.

Our estimation sample consists of job ads that were active on the portal between July 29, 2020 and November 13, 2022.[11] Figure 1 shows variation in the number of ads posted on the NCS portal over time and across Indian districts.[12] As may be seen, the number of posted job ads has increased over time. To obtain the number of ads for each district, we first drop job ads for which job location information is missing or uninformative (e.g. referred to as "All India"); these form 6.33% of all ads. Of the remaining 311,010 job ads, 18.08% list multiple districts as the job location. Since this is a substantial fraction of overall ads, we do not drop these. Instead, for ads that list multiple districts as the job location, we consider the ad to be equally split across each of the listed districts while counting each ad as one. Over 20% of ads list Bangalore as the job location (accounting for the highest number of ads), followed by Mumbai, Pune, Chennai, and Central Delhi. Appendix Figure A.1 gives the geographic variation of ads across Indian districts and shows that ads posted on the NCS portal are located in districts across all regions of India.[13]

Appendix Figure A.2 compares posted job ads (Panel (a)) with the regional distribution of employment in urban areas within the working age population (15-59 years) from the Periodic Labor Force Survey or PLFS 2020-21 (Panel (b)). Most ads on the NCS portal are located in the Indian state of Karnataka, which accounts for over one-fifth of the ads, followed by Maharashtra, Tamil Nadu, and West Bengal. In contrast, employment shares across Indian states from the PLFS are more evenly distributed with Maharashtra accounting

[9]The portal website is https://www.ncs.gov.in/.

[10]Since its inception and as of September 26, 2022 over 10.9 million vacancies have been advertised on the portal. As of December 11, 2022 there were 27.6 million job seekers registered on the portal of which 11.5 million were rural youth in the age group 15–29 years.

[11]These are ads having a last date of application on or after July 29 2020, which are also posted before November 13, 2022. The two dates correspond to when we started and ended scraping job ads data from the portal.

[12]We show the twenty five Indian districts with the highest number of job ads for ease of readability.

[13]Of the 766 districts in India, 650 had online job ads posted on the NCS portal.

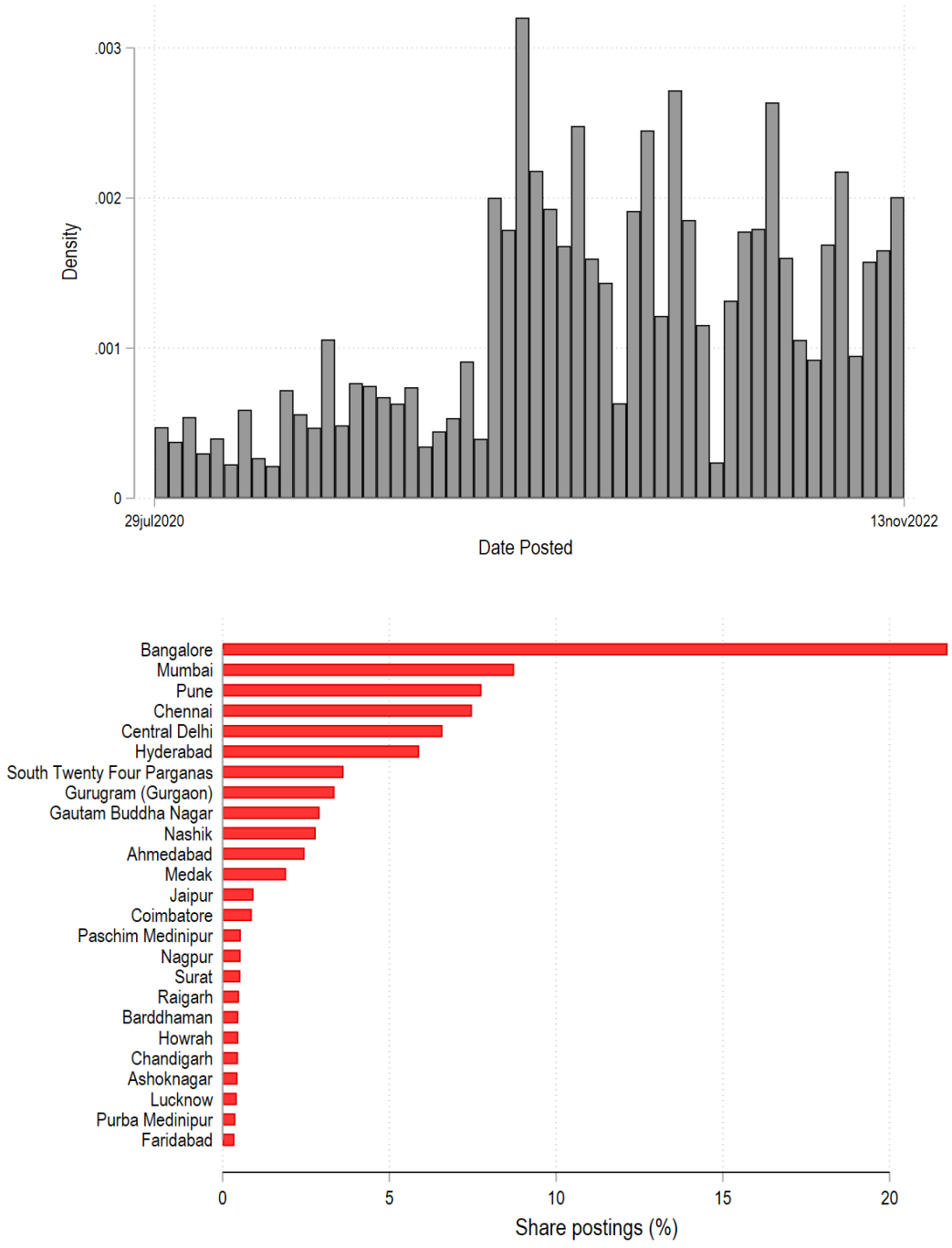


Figure 1: Temporal and geographic distribution of online job ads posted on the National Career Services (NCS) portal in India between July 2020 and November 2022. The lower panel reports the top 25 districts by number of postings.

for 8.8% of urban employment followed by Tamil Nadu (7.6%), Uttar Pradesh (7.1%), West Bengal (6.3%), and Karnataka (5.2%). This is consistent with the NCS portal hosting jobs which are disproportionately located in large Indian cities such as Mumbai, Delhi, Kolkata, Chennai, and Bangalore.

We make use of data on the job title and description for each job ad posted on the portal, the number of openings (or vacancies), the type of organization putting up the ad (e.g. private, government, NGO etc.), their sector (e.g. information technology, finance, manufacturing etc.), the functional role for which a person is being sought (e.g. accountant, customer care, HR etc.), education and experience requirements, key skill requirements, type of job contract (e.g. full time, part time, or internship), job location, name of employer, and offered salary range.[14]

Appendix Table A.1 gives descriptive statistics for all job ads posted on the NCS portal. On average, a job ad advertises for 11.7 vacancies, although there is considerable variation in the number of vacancies across job ads as indicated by the high standard error. The average experience requirement is ≈ 4.5 years, although there is again considerable variation in this across job ads. Education requirements are specified for $\approx 57\%$ of job ads; within these, almost all ads require candidates to have completed schooling. This indicates that ads posted on the NCS portal are for relatively high skill jobs with high education requirements. In contrast, just 40% percent of the workforce in urban India has completed schooling (PLFS 2020-21).

Job ads provide information on wages for just over one-third of all postings. In cases where a salary range is provided, we use the mid value of the range. The average annual salary using these ads is 0.278 million Indian Rupees (INR). In comparison, the average wage among urban salaried individuals is 0.19 million INR in the nationally representative PLFS 2020-21. This also shows that ads posted on the NCS portal are for relatively high skill jobs which offer higher salaries than those earned by comparable nationally representative samples of employed workers.

We examine how posted wages correlate with skills using job ads that include a posted wage or a range for the posted wage. We exclude job ads that are outliers, i.e., for which the wage is above the 99^{th} percentile or below the 1^{st} percentile. Of the $332,044$ job ads we have data on, $119,740$ post a wage ($= 36\%$ of all job ads), and of these $81,468$ include required key skills ($= 68\%$ of job ads that post a wage).[15] For regressions that use wages

[14]There are 129,759 unique job titles, 221,086 unique job descriptions (255,240 unique title-description combinations; some of them also include gender preferences), and 139,738 unique skills combination requirements.

[15]Among all job ads, $277,700$ contain key skills information ($\approx 85\%$ of all ads).

as the dependent variable of interest, we further restrict our sample to job ads that specify locations within the same state since we control for state fixed effects in these regressions. This leaves us with an estimation sample of 62,958 job ads (= 19% of all job ads).

Appendix Table A.2 gives descriptive statistics for the estimation sample of 62,958 job ads. This sample has higher mean vacancies (= 13.62) and required years of experience (= 1.49) compared to the overall sample, as well as lower posted wages (0.25 million INR). Almost 90% of job ads in this sample specify an education requirement. This sample is also a selected sample of low skill jobs. Appendix Table A.3 shows correlates of job ads that do not specify a wage; it shows that job ads that require a higher education and experience are **less** likely to post a wage. This is in line with the existing empirical literature using job portal data from the US, UK and Slovenia (Brenčič, 2012) as well as Chile (Banfi & Villena-Roldan, 2019). It is also consistent with a model where hidden wages may be used as a signal to high skill applicants that the employer is open to ex-post bargaining (Michelacci & Suarez, 2006).

As may be seen in Appendix Table A.1, the sectoral distribution of posted job ads on the NCS portal is highly skewed towards the services sector (at 93.3%). There are almost no jobs in the agriculture sector, while construction and manufacturing comprise 0.5% and 5.9% of all job ads posted on the portal. In contrast, urban India has 5.9%, 11%, 20%, and 63% of it's workforce in agriculture, construction, manufacturing, and services (PLFS 2020-21). In addition, 43% of firms posting ads on the NCS portal are private enterprises. Most posted job ads (= 81%) are for full time work while 6.5% are for part time work, and 12.5% are internships.

To investigate the association of firm characteristics with skills demand, we match firm names in the job portal data with the Ministry of Corporate Affairs (MCA) database in India which contains firm details including principal business activity, industry, year of registration, and paid-up capital of a firm. Using the merged data we are able to measure the size of a firm using paid-up capital. We only use the set of firms that match exactly on firm name in the MCA database. There are 50,155 firms that post job ads on the portal, and we are able to match 16,672 of these with the MCA database (= 33% of all firms). Of these 16,512 have information on paid-up capital, and 12,000 post at least one job ad which includes required key skills. Of the 2.17 million firms in the MCA database, 62% are currently active compared to 95% in our matched sample. Within these, 5.1% and 7.3% firm are public enterprises in the MCA database and the matched sample, respectively. The mean paid-up capital in the sample of matched firms is an order of magnitude higher at 492 million INR (s.d. 14.5 billion) compared to 3.4 million INR (s.d. 2.4 billion) for all active firms in the MCA database—indicating that larger firms are more likely to post ads on the NCS

portal. Additionally, matched sample firms are more likely to be involved in computer-related activities (30.9% vs. 9.2% overall).[16]

Appendix Table A.4 shows descriptive statistics for the sample of 12,000 firms (= 24% of all firms) which we matched with the MCA database and which include required key skills. In our analysis, we use three different measures of firm size. These include the number of job postings advertised by the firm, the number of vacancies advertised and paid-up capital. The first two measures capture firm size since bigger firms are likely to hire more workers and hence have more postings and vacancies. They are also likely to have higher paid-up capital. The median postings and vacancies advertised by a firm are 2 and 4, respectively. The median paid-up capital of firms is 0.1 million INR. 73% of firms are private while 26% fall in the category of others. There are very few (or hardly any) government organizations or NGOs in our estimation sample.

The use of the NCS portal data only allows us to uncover patterns in skills demand when firms use the portal to recruit workers; we cannot study patterns in skills demand where firms recruit workers through informal channels or using headhunters. Our comparisons with nationally representative samples of employed workers from the PLFS in this Section also indicate that our sample primarily consists of high skill jobs in the services sector where jobs are located in large Indian cities. Within this sample we can only observe wages for relatively low skill jobs. For these reasons, our estimation results are best extrapolated to similar kinds of jobs as contained in our sample.

3 Methods

3.1 Word embeddings and gender associations in skills demand

We use computational linguistics methods to obtain continuous vector representations of all words and phrases in our sample of job ads. Specifically, we calibrate a neural network model that can identify linguistic information such as analogies or semantic meaning based on the usage of words in job titles and descriptions, without relying on any external data source. For this, we use an open-source library called `fastText` proposed by Bojanowski *et al.* (2017) which improves upon the popularly used `word2vec` by leveraging sub-word (or character-level) information to take the morphological structure of words into account. This

[16] Comparing across regions, a larger percentage of firms in the matched sample are registered in Karnataka (14.3% vs 6.7% overall), Telangana (8.6% vs 5.6%), Maharashtra (22% vs 19.4%), Tamil Nadu (8.9% vs 6.5%), and Delhi (17.8% vs 16.2%). Conversely, firms registered in West Bengal (3.9% vs 9.8%), Uttar Pradesh (4.9% vs 7%), Kerala (1.3% vs 3%), and Bihar (0.9% vs 2.2%) are underrepresented in the matched data.

allows us to obtain vector representations even for skills that do not appear in the training data—for example, due to spelling inconsistencies or rarely occurring word forms.[17] The `fastText` model makes use of the following components:

- **Skipgram model**: The continuous skipgram model, introduced by Mikolov *et al.* (2013), trains a neural network to predict the presence or absence of context (or nearby) words given a target word. This problem is formulated in terms of independent binary classification tasks in which the training corpus is represented as a sequence of target words w_t, where $t \in \{1, ..., T\}$ denotes the word position. For a given word w_t, all the context words w_c, $c \in C_t$ are taken as positive examples while the negatives $n \in N_{t,c}$ are sampled at random from the dictionary of words having a vocabulary of size W. The model minimizes the sum of negative log-likelihood below:

$$\sum_{t=1}^{T} \left[\sum_{c \in C_t} log(1 + e^{-u_{w_t} \cdot v_{w_c}}) + \sum_{n \in N_{t,c}} log(1 + e^{u_{w_t} \cdot v_n}) \right]$$

  u_{w_t} represents the input vector for the target word and $v_{w_c} \in \mathbb{R}^d$ represents the output vector for the context word. For example, given the sentence "analyze business data using Python and advanced analytics modules" and the target word "Python", a skipgram model tries to predict the presence of all nearby words such as "business" or "analytics". This produces a vector representation for each word such that words that are often used in a similar context are located closer to each other in the vector space.

- **Subword information**: Subword information for a word w is incorporated by assigning vector representation z_g to each constituent character n-gram $g \in G_w \subset \{1, ..., G\}$, where G_w includes all substrings in the word between a minimum and maximum length and the entire word, and G is the size of the n-gram dictionary. Each word is then represented as the sum of vectors of its constituent n-grams. In practice, with subword information, the skipgram model works better than the continuous-bag-of-words model which simply uses the **sum of vectors** of all the nearby words to predict a target word.

We follow standard pre-processing steps before training the model. We start by combining the job title and description for each job ad and lowercase our data. Next, we only keep alphabets, full-stops, commas, space, hyphens, &, and the + sign (this is included in the names of many software packages). Finally, we remove duplicate job ads (i.e. duplicates in

[17] Edwards *et al.* (2020) demonstrate that `fastText` trained on domain-specific data can outperform the pre-trained language model BERT on classification tasks while having a lower computational cost.

skills, job title, and description) so that we train the model on a diverse set of examples. The total corpus size is 17.85 million words and includes 143,726 distinct words. Appendix Figure A.3 gives the distribution of relative frequencies, i.e. frequency per million words ($fpmw$) against the word frequency rank on a logarithmic scale. The word "and" is the most frequent word accounting for 4.30% of all word occurrences followed by "the" (2.66%). The words "female" and "male" account for around 0.11% of word occurrences each. Consistent with Zipf's law, there is a strong correlation of 0.94 between $fpmw$ and $rank^{-1}$.

We use default parameters to train a 300-dimensional model and loop over the data 10 times.[18] To compare domain-specific word embeddings with word embeddings that are not domain specific, we also use the 300-dimensional pre-trained English `fastText` word vectors trained by Grave *et al.* (2018) on Common Crawl on the web and Wikipedia.[19] For both domain-specific word embeddings and word embeddings that are not domain specific, we compute the cosine similarity of a skill word i mentioned in job ads with the words "female" or $CS(i = skill, female)$, as well as "male" or $CS(i = skill, male)$.[20] To reliably compute a gender association measure for different required skills, we only compute this for skills that are not rare, i.e. occur more than 10 times in our data.

Our specific choice of words "female" and "male" to identify gender associations stems from how employers indicate their gender preferences in job ads. The words "female" and "male" are mentioned in 6.93% and 5.94% of job descriptions and are often used to indicate explicit gender preferences. This may be seen in Figures A.4 and A.5 which re-produce job ads in which employers explicitly state their female and male gender preferences. Figure A.4 shows a job ad which requires a "Customer Relationship Executive" where the job description clearly indicates a female preference with the word "female" appearing in the job description. The key required skill in this ad is *telecaller* with a relatively high net female association of 8.26. Figure A.5 shows a job ad which requires "HI Tech Lenses" where the job description indicates a male preference with the word "male" appearing in the job description. The key

[18]These default parameters include dropping words that occur less than 5 times in the training data. The context windows size c is 5. We use negative sampling loss with a learning rate of 0.05. For each positive word, 5 negatives are randomly sampled with probability proportional to the square-root of their frequency. The length of character n-grams is set from 3 to 6 and a sampling threshold of 10^{-4} is taken to discard more frequent words such as "and", 'the", and "too" from the context window (see Mikolov *et al.*, 2013).

[19]Available at https://fasttext.cc/docs/en/crawl-vectors.html.

[20]Formally, cosine similarity between two word vectors is defined as:

$$CS(\vec{x}, \vec{y}) = cos(\theta) = \frac{\vec{x}.\vec{y}}{||\vec{x}||||\vec{y}||} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

where $\vec{x}$ and $\vec{y}$ are non-zero vectors, θ is the associated angle, and $||.||$ is the 2-norm; $CS(\vec{x}, \vec{y})$ varies between -1 and $+1$.

required skills include *Industrial Engineer Executive* and *Associate Operations Engineer* with a relatively low net female association of −5.02.

In contrast to "female" and "male", gender attribute words such as woman, girl, lady, feminine, man, boy, gent, guy, and masculine and their plural forms collectively occur in less than 1% of job titles and descriptions. In addition, these words may be used in different contexts. Thus, their presence is not necessarily indicative of employers' gender preferences. For example, the word feminine occurs in the context of feminine hygiene products, woman or women often occur in the context of women's issues or products linked to health and fashion, and girl or girls commonly occur in the context of a girl child. Similarly, the word man also occurs as part of the phrase "man power". Consequently, there is low cosine similarity between the words female and woman (= 20.73%), or between male and man (= 13.26%). Therefore, as reflected by their usage, these words are not synonymous with female and male in the job ads corpus. Table A.5 separately reports their frequency and cosine similarity with the reference words female and male respectively.

Seminal work by Bolukbasi *et al.* (2016) quantifies gender bias using word embeddings by comparing word vectors to pairs of gender specific words and projecting them on a gender dimension using Principal Component Analysis (PCA). This is motivated by their argument that "the fact that nurse is close to woman is not in itself necessarily biased (it is also somewhat close to man—all are humans), but the fact that these distances are unequal suggests bias" (page 2, paragraph 3). They confirm that such biases using word embeddings are indeed aligned with gender stereotypes based on evaluation by crowd workers from the U.S. on Amazon's Mechanical Turk; for example, they find that doctor is closer to "man" than to "woman". Similarly, in the context of job ads, skills related to *software architecture*, *Artificial Intelligence* and *Machine Learning*, and coding skills such as *Python* or *Java* have low cosine similarity with both "female" and "male" but are relatively closer to "male" than "female". On the other hand, non-technical skills such as *sales*, *cooking*, *language skills*, and *project management* have high similarity with both "female" and "male". Furthermore, *content writer* has higher cosine similarity with both "female" (= 25.46%) and "male" (= 20.95%) compared to *content writing* (cosine similarity of 20.38% and 17.98% with "female" and "male" respectively). This is because the word *writer* refers to a person, and therefore, is more similar to both "male" and "female" than *writing*. We see similar patterns when the skills mention *engineer* (e.g. *quality assurance* vs. *quality assurance engineer*) or when the skill word is *counsellor* (cosine similarity of 30.73% with "female" and 20.47% with "male") rather than *counselling* (25.77% and 18.50% cosine similarity with "female" and "male" respectively).

Gonen & Goldberg (2019) demonstrate that the projection of words on the gender dimension does not fully capture gender bias. Therefore, Caliskan *et al.* (2017) use a cosine similarity-based measure of differential gender associations of 50 occupation words in pre-trained GloVe embeddings and show that this measure correlates strongly with the share of women employed in these occupations based on data from the United States Bureau of Labor Statistics, 2015. Similarly, Garg *et al.* (2018) simply compute distance of occupation words (e.g. lawyer) from words that represent women (e.g. female) minus distance of occupation words from words representing men (e.g. men) to compute embedding bias. They find that changes in this measure during each decade from 1910 to 1990 correlate strongly with changes in the relative share of women in occupations in the United States.

We follow a similar approach and construct a measure of **relative** gender association with women vs men for specific required skills by computing the difference in cosine similarity of a required skill with the words "female" and "male" or $CS(i = skill, female) - CS(i = skill, male)$.[21] We also construct a measure of aggregate relative gender association for job ad j by taking the mean of differences in cosine similarity across all skill words i in job ad j (or $CS_j^{diff} = \frac{\sum_{i \in j} CS(i=skill,female) - CS(i=skill,male)}{\sum_{i \in j} \mathbb{1}[i=skill]}$).[22]

3.2 Skills categorization

Rather than using pre-defined skill categories, we follow a data-driven approach towards skills classification. We first create a dictionary of skills while restricting the set of skills to those mentioned at least 10 times in the corpus of job ads. We then obtain a 300-dimensional vector representation of each skill using `fastText` and map it on a 2-dimensional space using Uniform Manifold Approximation and Projection (UMAP)—a widely used manifold learning or non-linear dimension reduction method that tries to maintain the underlying topological structure of the data (McInnes *et al.*, 2018). This dimension reduction step not only allows us to visualize the listed skills, but also improves the HDBSCAN clustering algorithm which

[21] We also compute this measure by taking the difference of mean cosine similarity of each skill with all the feminine (female, females, woman, women, girl, girls, lady, ladies, and feminine) and masculine (male, males, man, men, boy, boys, gent, gents, guy, guys, and masculine) attribute words. We find a moderately low but positive correlation (= 37%) between relative gender associations of skills using the two measures. We see the largest absolute change in the relative gender association for "receptionist" skills from 0.07 when only using the words "female" and "male" to −0.04 when using all feminine and masculine gender attribute words.

[22] This measure cannot be constructed for job ads that do not mention any required skills or only mention rare skills. We also separately aggregate male and female gender associations of skills within each job ad. For this, we first normalize the number of skills mentioned in each job ad to one. To compute aggregate female (male) associations at job ad level, we take the weighted sum of $CS(i = skill, female)$ ($CS(i = skill, male)$) for female (male) skills, i.e. for which $CS(i = skill, female) - CS(i = skill, male) > 0$ ($CS(i = skill, female) - CS(i = skill, male) < 0$).

works better on low-dimensional data.

UMAP involves the following two steps:

1. **Graph Construction**:

   - In the first stage, weighted directed graphs (or simplicial sets) are constructed for each skill x and its k-nearest neighbours represented as $y \in X_k$ such that the nodes correspond to skill vectors and the edge represents smoothed similarity $v_{y|x}$ between them, roughly interpreted as the probability that the edge exists:

   $$v_{y|x} = exp[-max(0, d(x, y) - \rho_x)/\sigma_x]$$

   where $d(x, y)$ is the distance (e.g. Euclidean or Cosine) between x and y, ρ_x is the distance to the nearest neighbour of x, and σ_x is a normalizing factor.[23]

   - It is then symmetrized into a fuzzy topological structure (i.e. a simplicial complex) or an undirected weighted graph by taking a fuzzy union. The weight on the undirected graph is obtained by taking the probabilistic t-conorm and can be interpreted as the probability that at least one of the edges between x and y exists:

   $$v_{x,y} = v_{y|x} + v_{x|y} - v_{y|x} \cdot v_{x|y}$$

2. **Finding a low-dimensional representation**: This is then represented on a low-dimensional space by searching for a projection having the closest possible equivalent fuzzy topological structure by minimising the edgewise cross-entropy between the two representations. Given the weight $w_{x,y} = (1 + a||x - y||^{2b})^{-1}$ of edge connecting nodes x and y in the low-dimensional space, where a and b are some constants, cross entropy is:

   $$\sum_{x \neq y} v_{x,y} log \frac{v_{x,y}}{w_{x,y}} + (1 - v_{x,y}) log \frac{1 - v_{x,y}}{1 - w_{x,y}}$$

The key hyperparameter in UMAP is k or the number of nearest neighbours. Since a small value of k incorporates only a small number of neighbours, it captures the local details well while a larger value works better at depicting the overall structure more accurately. We take $k = 20$ to depict some of the global structure while retaining finer local details, though the visualization remains similar when we take values of k up to 200. In addition, we set the

[23] The value σ_x is chosen such that $\sum_{y \in X_k} exp(\frac{-max(0, d(x,y) - \rho_x)}{\sigma_x}) = log_2(k)$. See McInnes *et al.* (2018) for the mathematical underpinnings. The implementation takes $v_{y|x} = 0$ if $y \notin X_k$ for computational reasons.

minimum possible distance between points to 0 which packs them together more densely in the low-dimensional space and is useful for clustering.

Compared to the popularly used manifold learning technique t-SNE, UMAP has several advantages. First, UMAP captures the global structure of the data better while preserving local distances which also makes it better suited to clustering. Moreover, it scales more efficiently with a linear time complexity $O(N)$ as opposed to O(N log(N)) for t-SNE. Finally, it also supports cosine distances which are commonly used to compute distances between word vectors. In contrast to linear dimensionality reduction algorithms such as PCA whose dimensions indicate the direction of greatest variance in the underlying data, and like other manifold learning techniques such as t-SNE or Isomap, UMAP dimensions don't have a specific interpretation. However, UMAP can account for non-linear structure in the data and allows us to create a skills map such that similar skills are placed close to each other while also incorporating the global structure to some extent.

We then use a hierarchical, density-based clustering algorithm HDBSCAN of Campello *et al.* (2013). This algorithm groups together regions of high density into distinct clusters separated by low-density regions which it eliminates as noise. There are two major advantages of HDBSCAN for us. First, it doesn't require a pre-specified number of skill categories. Second, it discards some of the skills as not falling into any of the categories, and is therefore conservative in cluster assignment. The algorithm uses the following steps, which we describe in detail:

1. **Transform the space**: To make the single-linkage clustering algorithm robust to noise, points in low density neighbourhoods, i.e. those further from their k^{th} nearest neighbours, are spread out by defining a metric called the mutual reachability distance:

$$d_{mreach-k}(x, y) = max\{core_k(x), core_k(y), d(x, y)\}$$

   $d(x, y)$ denotes some measure of distance between two distinct points x and y; $core_k(x)$ or core distance is the distance between x and its k^{th} nearest neighbour. The data is then represented as a weighted graph with edge weights of $d_{mreach-k}(x, y)$.

2. **Build the minimum spanning tree**: A minimum spanning tree of the graph—the subset of edges that connect all vertices without any cycles and minimum possible total edge weight—is constructed using Prim's algorithm. Initializing at a random vertex, we sequentially add the lowest weight edge connecting a vertex not yet in the graph.

3. **Build cluster hierarchy**: Starting with every data point as an individual cluster, the

minimum spanning tree is converted into a hierarchical tree by sorting the edges in increasing order of their weights, and iteratively merging clusters connected by an edge.

4. **Condense the cluster tree**: The algorithm traverses the hierarchy by gradually decreasing the distance threshold. If any cluster created at a split has fewer points than the "minimum cluster size"—a user-defined parameter, then the larger cluster retains the identity of parent and we record how many points drop out of the cluster at each threshold. Otherwise, if both the clusters are at least as large as the minimum cluster size, then they are allowed to persist. Ultimately, this results in a smaller tree with information on how the size of each cluster decreases at various distance thresholds.

5. **Obtain flat partition of clusters**: For $\lambda = [d_{mreach-k}(x, y)]^{-1}$, let λ_{birth} and λ_{death} denote the values when a specific cluster is born from a larger cluster and when it splits into smaller clusters respectively. We also denote $\lambda_p \in (\lambda_{birth}, \lambda_{death}]$ as the λ value at which a given point p falls out of the cluster. The cluster stability is then defined as:

$$\sum_{p \in cluster} (\lambda_p - \lambda_{birth})$$

   The stability criterion is applied to maximize the overall stability of **selected** clusters. Starting from individual data points (or leaf nodes) as selected clusters, if the sum of stability of child clusters is greater than a cluster's stability, then the cluster stability is set to be the sum of child stabilities. On the other hand, if the cluster's stability is more than the sum of stability of its child clusters, then this cluster is selected and all descendents are unselected. This process continues until we reach the top of the tree (root node) and the currently selected set of clusters represent the final clustering.

Centroid-based or parametric clustering algorithms such as k-means assume clusters have a convex shape and perform poorly when this assumption is violated. Similarly, agglomerative clustering—a hierarchical clustering algorithm—tends toward spherical clusters especially when using Ward linkage and is susceptible to noise when using single-linkage. This is overcome by the non-hierarchical density-based DBSCAN algorithm. However, it uses a global density threshold and does not handle variable density clusters well. In contrast, HDBSCAN uses the stability criterion and can properly characterize the data even when clusters have different densities. Moreover, the main hyperparameter for HDBSCAN is the intuitive "minimum cluster size" parameter described above instead of the number of clusters for k-means and agglomerative clustering or cluster resolution in DBSCAN.[24]

[24]We set the "minimum cluster size" parameter so that a skill cluster has at least 20 distinct skills, k is set to

We arrive at thirty seven skill categories using this method. Appendix Table A.1 shows that, on average, there are 1.47 required skill categories across posted job ads on the NCS portal. In the sample of job ads with a non-missing wage, the number of required skill categories per job ad falls slightly to 1.39 for each posted job ad, on average. We also estimate gender associations for each of the thirty seven skill categories l to obtain a systematic understanding of skills that employers associate relatively more with women vs men. For this we simply take the weighted (by frequency of occurrence) mean of $[CS(i = skill, female) - CS(i = skill, male)]$ for all $(i \in j) \wedge (i \in l)$. Within the $62,958$ job ads that report a wage, key skills, and post a job location within a single Indian state, we are able to arrive at gender association of skills for $59,659$ job ads.

Appendix Table A.2 shows that the average female association is 1.08 and average male association is 1.14 in job ads that post a wage. This shows that, on average, job ads are more likely to require male associated skills than female associated skills. We also construct the net female association of a job ad by taking a difference between female and male association score for the skills required in the ad. On average, the net female association score in skills is negative at -0.056 for job ads that include wage information. This again shows that, on average, skills associated with men dominate over skills associated with women in the posted job ads. It is also important to note that the net female association score is more negative $(= -0.328)$ in the sample of job ads that include ads with missing wages (Appendix Table A.1); this indicates that high skill jobs which exclude wages tend to require skills associated with men rather than women. Appendix Figure A.6 shows the density function of net female association in all job ads, and in job ads with non-missing wages. This also shows that net female association of required skills is higher in ads with non-missing wages than in all ads.

At the firm level, Appendix Table A.4 shows that, on average, 0.75 skills associated with females are required by a firm across all its job ads while 2.03 skills associated with males are required. In total, 2.78 skills are demanded by a firm, on average. Around 32% firms on the NCS portal demand both a male and a female associated skill.

3.3 Occupation classification

We partition job ads into occupations based on job titles and descriptions. This allows us to include occupation fixed effects in the analyses. To do this, we first concatenate job title and description for a given job ad and obtain sentence embeddings, i.e. a single 300-dimensional

1 so that fewer skills are discarded as noise and the clustering is not too conservative. We further set a parameter ϵ to 0.2 so that clusters below this threshold Euclidean distance in the UMAP space are not split further. We use the implementation in McInnes *et al.* (2017) (for details, see `https://hdbscan.readthedocs.io/`).

vector representation of each job ad. This is obtained by dividing each word vector by its $L2$ norm and then taking the average over all the words in the job ad. Therefore, for a given job title and description combination with n words (including the end of sentence or the newline character), the sentence vector s is created as follows:

$$s = \frac{1}{n}\sum_{i=1}^{n}\frac{x^i}{\sqrt{\sum_{k=1}^{300}(x_k^i)^2}}$$

where x^i refers to the vector for word i in the job description while k indexes the dimensions of the vector.[25] We then apply the k-means clustering algorithm to cluster job ads into 300 occupation categories using these embeddings. This is a simple algorithm and scales well to large samples. It partitions the data into k disjoint clusters by first choosing cluster centroids to minimize within-cluster sum-of-squares (or inertia):

$$\sum_{i=1}^{N}\min_{\mu_k}||x_i - \mu_k||^2$$

where μ_k corresponds to centroid closest to the job ad having vector x_i. It then assigns each point (or job ad) to the cluster represented by the closest centroid. The number of clusters or occupation categories is chosen based on two considerations. First, we use a heuristic known as the **elbow method**. We plot how the inertia changes as we increase the number of clusters from 10 to 1,000 (in increments of 10) in Appendix Figure A.7. We then identify the point at which the rate of decline in inertia with respect to the number of clusters becomes approximately flat, i.e. there is no substantial decrease in inertia on increasing the number of clusters further. Second, a visual examination of cluster labels also suggests that having 300 occupation groups provides a good balance of separating job ads belonging to different occupations while grouping together similar ads.

3.4 Skills demand

We examine skills demand across several dimensions:

Regional patterns Since ads contain information on job location across thirty seven states/union territories of India, we can find the share of each skill in a given location. In order to examine regional patterns we drop job ads specifying the location of the job as "All India"; such ads form 4.7% of all job ads in our sample. For job ads that specify multiple

[25]The averaging process excludes word vectors having an $L2$ norm of 0.

locations we assign the required skill to all these locations after normalizing skill counts such that each job ad is assigned the same weight. To estimate geographic variation in the relative demand for different skills we use Balassa's Revealed Comparative Advantage (RCA) index proposed by Balassa (1965). This index measures the share of a given skill category in a location relative to its overall share in job ads. Specifically, for location d and skill category l, this index is computed as follows:

$$RCA_{d,l} = \frac{N_{d,l}/N_d}{\sum_p N_{p,l}/\sum_p N_p}$$

where $N_{d,l}$ is the number of job ads in district d that mention skill l, N_d is the total number of job ads in district d, $\sum_p N_{p,l}$ is total number of job ads across all districts p that mention skill l and $\sum_p N_p$ is the total number of job ads across all districts p.

Skills demand and posted wages We estimate the correlates of different skill categories with log wages and examine how this varies with the gender association of a required skill category. As discussed earlier, on average the gender wage gap is 21% in urban India for workers age 15-59. Even among the younger cohort of individuals age 20-35 who are at least school educated, and closer to the relatively skilled labour market targeted by the portal, women continue to earn 10% less than men within the same occupation and location. Our proposed analyses, thus, allows us to investigate whether skills correlated with high posted wages have a lower or greater net female association, and the potential role played by skills demand in gender wage gaps.

We use job ad level data on the skills demanded by a firm and the posted annual wage. It is important to note that there are likely to be omitted variables which have an impact on posted wages which we are unable to control for in our regressions. It is also possible that high paying job ads add more skill requirements, or there is reverse causality between wages and required skills. Without a completely comprehensive set of controls or some exogenous variation which makes use of either relevant instruments or a natural experiment, we caution the reader to interpret the subsequent regression estimates as associations rather than causal effects.

We use the following regression specification to estimate the association of different skill categories with the log wage:

$$ln(wage_{jos}) = \beta_0 + \sum_{l=1}^{37} \beta_{1l} Scat_{l,jos} + \beta_2 \mathbb{X}_{jos} + \delta_{os} + \varepsilon_{jos} \quad (1)$$

$ln(wage_{jos})$ is the log posted wage in job ad j of occupation o located in state s. $Scat_{l,jos}$ is a dummy variable that takes the value one when a given skill category l is requested in job ad j and zero otherwise. $\mathbb{X}_{jos}$ includes controls for the requirements in job ad j i.e. required minimum education qualification and a quadratic in required experience. δ_{os} are (occupation $\times$ state) fixed effects. We estimate and report robust standard errors. The coefficient for each skill category (β_{1l}) indicates the association of that category with the log posted wage.

Next, we examine how posted wages vary by the net female association of required skills in a job ad (CS_j^{diff}) by estimating the following regressions:

$$ln(wage_{jos}) = \alpha_0 + \alpha_1 CS_{jos}^{diff} + \alpha_2 \mathbb{X}_{jos} + \delta_{os} + \varepsilon_{jos} \tag{2}$$

CS_{jos}^{diff} is the mean difference between female and male cosine similarity for all skills required in job ad j. As before we control for education and experience requirements in a job ad, and use variation in posted wages within an occupation and state by including $o \times s$ fixed effects. We estimate and report robust standard errors.

Skill demand and firm size We construct the following measures of overall and by gender skill demand at the firm level.

1. The number of skill categories demanded in all job ads j that are posted by firm f, which varies between 1 and 37:

$$ScatN_f = \sum_{l=1}^{37} \mathbb{1}\left[\left(\sum_{j:firm(j)=f} \sum_{i \in j} \mathbb{1}[i=l]\right) > 0\right]$$

2. The number of skills demanded by firm f which have a male association:

$$ScatNM_f = \sum_{l=1}^{37} \mathbb{1}\left[\left(\sum_{j:firm(j)=f} \sum_{i \in j} \mathbb{1}[i=l]\right) > 0\right] \times \mathbb{1}[CS(i=l, female) - CS(i=l, male) < 0]$$

3. The number of skills demanded by firm f which have a female association:

$$ScatNF_f = \sum_{l=1}^{37} \mathbb{1}\left[\left(\sum_{j:firm(j)=f}\sum_{i\in j}\mathbb{1}[i=l]\right) > 0\right] \times \mathbb{1}[CS(i=l, female) - CS(i=l, male) > 0]$$

4. An indicator variable that takes the value one if firm f demands both male and female skills, and zero otherwise:

$$ScatFM_f = \mathbb{1}[ScatNF_f > 0] \times \mathbb{1}[ScatNM_f > 0]$$

Using the skill demand measures (1–4 above), we test whether larger firms are more likely to demand specialized workers across a range of occupations and a dispersed set of skills, and whether this varies by gender association of skills. We also test whether joint demand for *both* male and female associated skills increases with firm size. To test these hypotheses, we estimate the following regressions:

$$Y_{fks} = \gamma + \gamma_1 ln(Firm\ Size_{fks}) + \gamma_2 X_f + \delta_k + \delta_s + \epsilon_{fks} \tag{3}$$

where the dependent variable $Y_{fks} \in \{ScatN_{fks}, ScatNM_{fks}, ScatNF_{fks}, ScatFM_{fks}\}$ is the number of skills, number of skills with a female association, number of skills with a male association, and the presence of both female and male skills demanded by firm f operating in business activity k and registered in state s.[26] We use three measures of firm size: the number of postings by a firm over the entire period, the number of vacancies by a firm and paid-up capital of the firm. Larger firms are likely to hire more workers and hence have more postings as well as vacancies. They are also likely to have higher paid-up capital, as discussed earlier. We use a logarithm in firm size for two reasons. First, firm size measures are skewed due to long right tails. In such cases, a log transformation is a standard statistical transformation to remove skewness from the predictor. Second, it eases the interpretation of γ_1 as the change in the dependent variable when firm size increases by one percent since we use various measures of firm size.[27] X_f include controls for industry of operation of the firm (at National Industrial

[26]There are 50 principal business activities classified in the MCA database largely based on industrial categories of the firms.

[27]We also check robustness of our results to using a linear specification in firm size. Here, instead of taking a logarithm we winsorize the firms size variables at 99 percentile and find that our results continue to hold; these results are available on request.

Classification 3-digit level), organization type (Private - partnership, Private - limited liability company, State government, Central government, NGO, and Others), and year of registration of a firm to measure its age. Standard errors are robust.

Unlike wages posted by firms in job ads, the dependent variables in equation (3) can contain zeros. We use a linear model rather than taking the inverse hyperbolic sine transformation ($arcsinh(Y)$) or taking the log after adding a small value to the dependent variable ($\log(Y+1)$) due to the pitfalls associated with these approaches when the outcome includes zeros, as highlighted by Chen & Roth (2023). We also show the robustness of our linear estimates to using a Poisson model for the first three measures of skill demand since these are count data.

When estimating the regressions above we use variables constructed using text algorithms as either a covariate or dependent variable of interest. Our empirical analysis treats these variables as any other numerical variable, but there are potential uncertainties in the measurement of these variables as well as possible correlations of the constructed variables with other variables within these regressions that should be kept in mind when interpreting results. We refer the reader to Section 4 of Ash & Hansen (2023) (and references cited therein) for a detailed discussion of these issues.

4 Results

4.1 Pre-trained vs domain-specific word embeddings

We find a positive but low correlation (= 11%) between estimates of gender association with skills when using a pre-trained model vs using one trained on our job descriptions corpus.

Using a model trained on our job descriptions corpus, we find the skills *counselling*, *human resource management and recruitment*, *teaching* (including *home tuition*), *front-desk* and *receptionist*, *tele-calling*, and *basic computer skills* are associated with women. On the other hand, working in *factories* and *warehouses*, *mechanical* and *electrical engineering*, *quality inspection*, *machine learning*, and knowledge of *programming languages* is associated more with men by employers, as reflected in the word embeddings. This is consistent with the findings in Chaturvedi *et al.* (2021) who uncover words predictive of explicit gender requests in job ads by employers using data from a different job portal and employing a different method.

In contrast, many of the associations outlined above are absent in the pre-trained model. For example, in the pre-trained model skills related to *programming*, *analysis of big data*, *inspection*, *microfinance*, and *supervisory* roles are associated more with women. This could

reflect online discussions aimed at improving the participation of women in fields where they are underrepresented (such as the Women in Big Data initiative), or microfinance programs aimed at female beneficiaries, and online forums supporting female supervisors.[28] Since the model trained on our job descriptions corpus is not familiar with such ideas and online discussions, by virtue of having **seen** only the job ads data, these biases do not feed into it allowing it to capture demand side gender associations reflected in job descriptions only. This model also captures the local Indian context more accurately where the gender associations of skills and the intensity of these associations might be distinctive from other settings.

4.2 Gender associations

Our clustering approach gives us thirty seven intuitive and varied skill categories. These include skill groups related to *computer programming*, *writing and designing*, *consulting*, *sales*, *interpersonal skills*, and *language* skills among others.

Figure 2 shows the four most frequent skills associated with each of these categories. The marker size indicates the frequency of the skill while the color represents each skill category. We see that programming languages such as *java*, *python*, and *c++* appear in the top right while *language* and *writing* skills appear in the top center. Thus, our methods allow us to create a data-driven skills map such that similar required skills appear in close proximity to each other.

Figure 3 shows associations of the different skill categories with men and women in online job ads. The skills in red represent higher association with women or higher $CS(i = l, female) - CS(i = l, male) > 0$ while those in blue represents higher association with men or lower $CS(i = l, female) - CS(i = l, male) < 0$. The Figure shows that employers associate *accounting*, *career counselling*, and *human resources/recruitment* skills more with women than men. On the other hand, *software* and *hardware engineering* are more strongly associated with men than women.

Table 1 lists the top ten required skill categories that are most most strongly associated with women and men (or have the highest and lowest values of $[CS(i = l, female) - CS(i = l, male)]$). It shows that *career counselling*, *recruitment*, *teaching*, *writing*, *language skills*, and *front office and customer support* skills are most strongly associated with women vs men. On the other hand, *quality control*, *software*, *data analytics*, and *cooking* skills are most strongly associated with men vs women.[29]

[28] It could also be that an inspector is implicitly assumed to be a male, but specific reference to the term "female inspector" is made to emphasize a *non-obvious* characteristic.

[29] This is also consistent with culinary industry being male-dominated globally. For example, according to

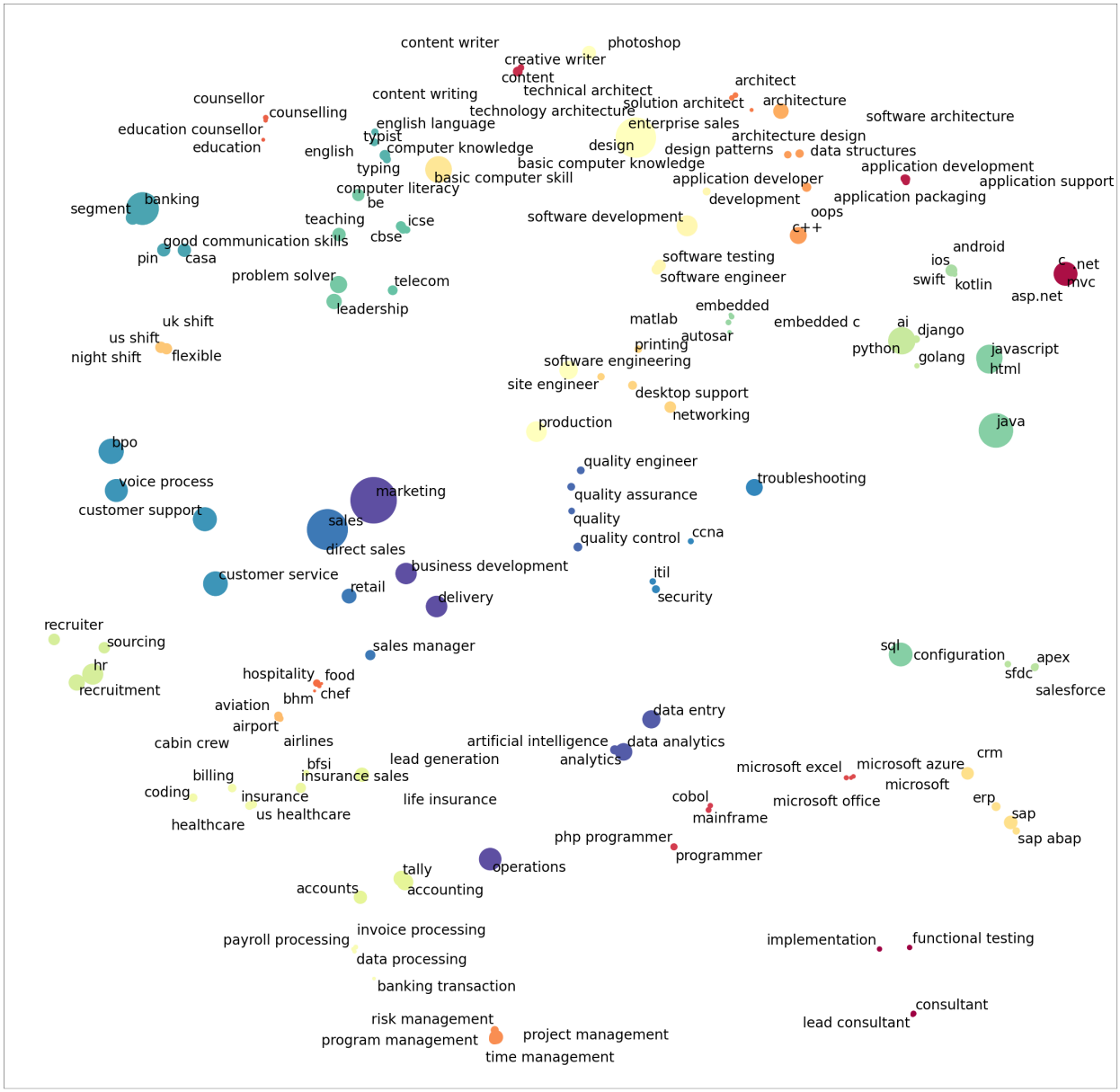


Figure 2: Skills map. Vector representation of skills projected on a two-dimensional space using UMAP using online job ads posted on the NCS portal in India between July 2020 and November 2022. The top 4 most frequent skills within each skill group (obtained using HDBSCAN clustering method) are included. Though the axes don't have a specific meaning (see Section 3.2), the map places similar skills close to each other in the local neighbourhood, while also depicting the global skill similarity structure. Marker size indicates the frequency with which a skill appears in job ads, while the color represents the skill group.

data from Office for National Statistics in the United Kingdom, only around one-fourth of chefs are women. This could be because working as a chef requires working in high-pressure environment and may be physically strenuous. Therefore, employers might consider it to be relatively better suited to men.

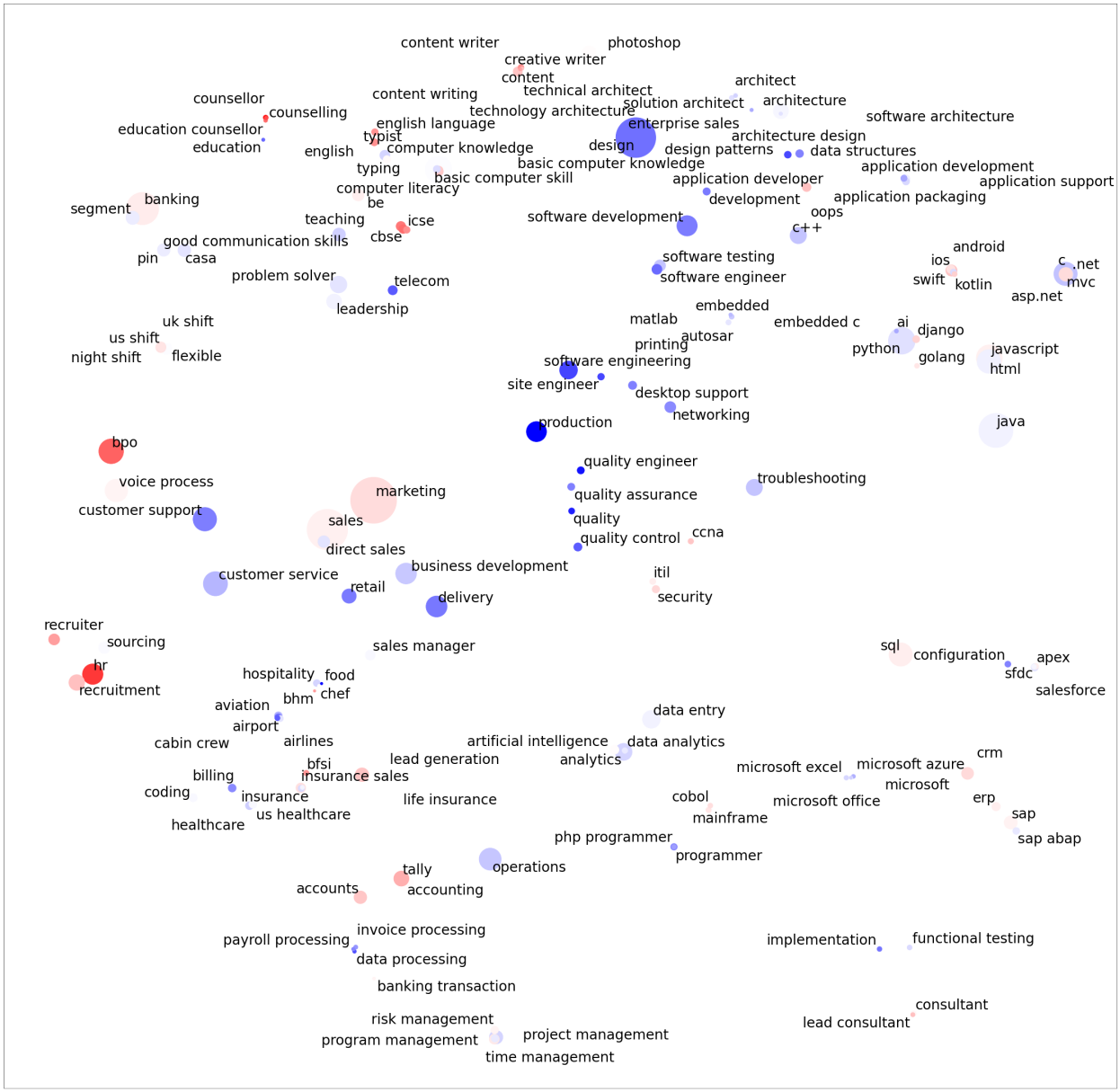


Figure 3: Skills map with gender associations. Vector representation of skills projected on a two-dimensional space using UMAP in online job ads posted on the NCS portal in India between July 2020 and November 2022. The top 4 most frequent skills within each skill group (obtained using HDBSCAN clustering method) are included. Though the axes don't have a specific meaning (see Section 3.2), the map places similar skills close to each other locally while also depicting the global skill similarity structure. Marker size indicates the frequency with which a skill appears in job ads while color intensity represents the degree of gender association. Red indicates a higher association with women while blue indicates a higher association with men.

Table 1: Skill categories with the strongest gender associations[a].

Skill Group ($= l$)	CS_l^{diff}	Frequency	Distinct Skills
Panel A: Women			
Career Counselling	6.67	1,340	29
Recruitment	4.03	19,047	128
Teaching	3.48	5,079	53
Writing	2.55	2,324	50
Language Skills	2.16	5,162	65
Front Office and Customer Support	1.33	118,082	686
Consulting	1.31	2,445	56
Accounting	1.08	18,413	159
Application Development Software	0.92	3,645	34
Basic Computer Knowledge	0.63	11,034	27
Panel B: Men			
Quality Control and Assurance	-4.27	3,341	63
Software Development and Testing	-3.39	11,657	106
Computer Hardware and Network Engineer	-3.31	10,954	126
Data and Payroll Processing	-3.29	938	25
Embedded Systems	-2.24	3,829	99
Aviation: Ticketing and Cabin Crew	-2.02	2,663	28
Cooking and Hospitality	-1.49	1,120	28
Microsoft Office	-1.10	1,077	28
Web Development (coding)	-1.07	13,345	52
Analytics	-1.03	8,641	68

[a] Aggregate gender bias for each skill category l in online job ads posted on the NCS portal in India between July 2020 and November 2022. The aggregate gender bias CS_l^{diff} is the mean difference in cosine similarity of skill-related phrases in each skill category with the words "female" and "male", weighted by their frequency.

4.3 Regional variation

To gain a systematic understanding of the distribution of skills, we compute ubiquity of skills and diversification of regions in terms of skills demand using the metrics in Hidalgo & Hausmann (2009). Ubiquity measures the number of districts that specialize in a given skill group or category while diversity measures the number of skill groups that each district specializes in (i.e. RCA$>$ 1).

We find that *sales & management*, *front office & customer support*, *language skills*, *basic computer knowledge*, *accounting*, *banking*, and *teaching* are the most ubiquitously demanded skills. On the other hand, knowledge of *customer relationship management software*, *computer systems architecture*, *data collection & management*, *embedded systems*, and *application support*

are the least ubiquitous.[30]

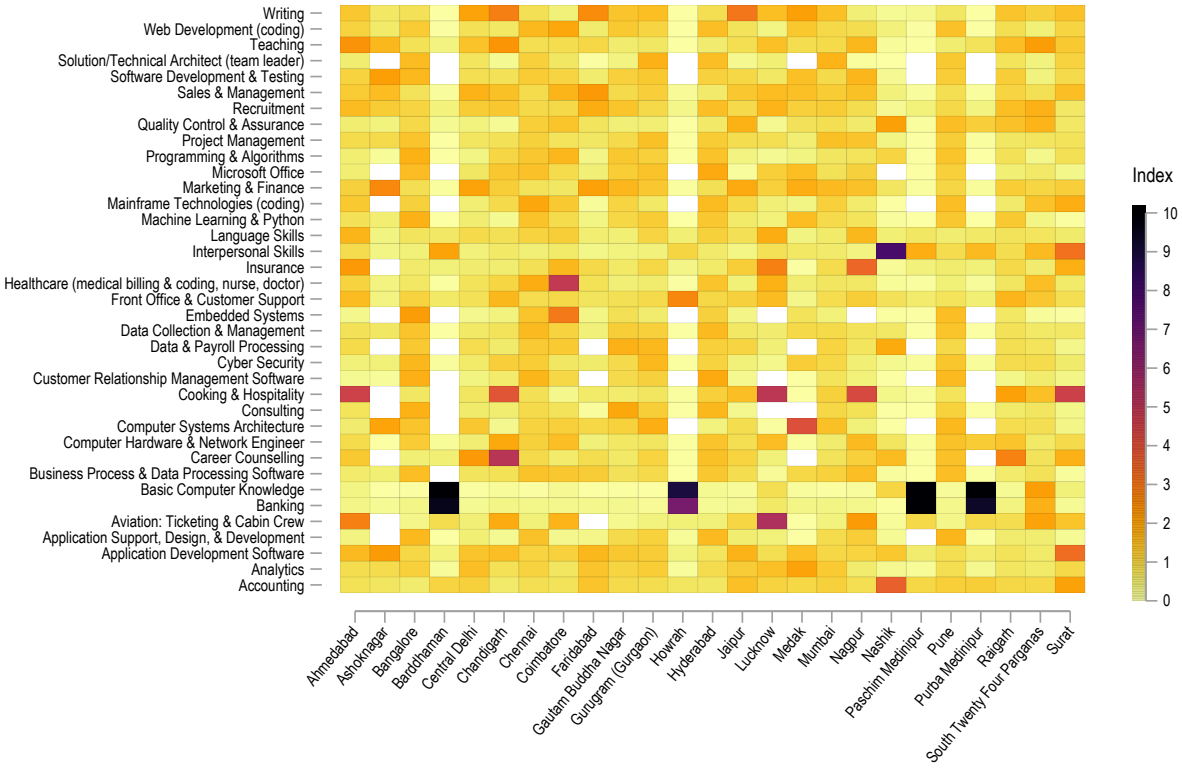


Figure 4: Skills demand across Indian districts. A heatplot of Balassa's Revealed Comparative Advantage (RCA) index, indicating relative skills demand, is given for the top 25 Indian districts (by number of postings). The heatplot is constructed using data on online job ads posted on the NCS portal in India between July 2020 and November 2022.

Figure 4 shows specialization of the top twenty five districts (by job postings) in each required skill category. Hyderabad, Chennai, Pune, Gurugram, and Bangalore are the most diverse districts specializing in a large number of skill categories. The least diverse districts are Paschim Medinipur, Barddhaman, Howrah, and Purba Medinipur, all of which are located in West Bengal and specialize in banking skills and basic computer knowledge. This is consistent with West Bengal having a disproportionately high demand for *finance*, *insurance, and accounting services*, and *administration/back office activities* in our data. In addition,

[30]Appendix Figure A.8 shows a heatplot of relative skills demand across different states and union territories in online job ads. Unsurprisingly, we find that Goa—a major tourist destination, has relatively high demand for *cooking* and *hospitality* related skills. We find that Telangana, Tamil Nadu, and Karnataka are the most diverse states while Lakshadweep, Dadra & Nagar Haveli, and Ladakh are the least diverse. On the other hand requirements for *language*, *sales & management*, and *teaching* skills are ubiquitous across states while the more technical skills related to *application development*, *embedded systems*, and *machine learning* have the least ubiquity in demand.

Ashoknagar (specializing in *marketing and finance* and *software development*), Faridabad (specializing in *writing* and *sales and management*), and Nashik (specializing in *interpersonal skills* and *accounting* due to a well-developed IT and communication sector) also specialize in relatively few skills.

4.4 Female labor force participation

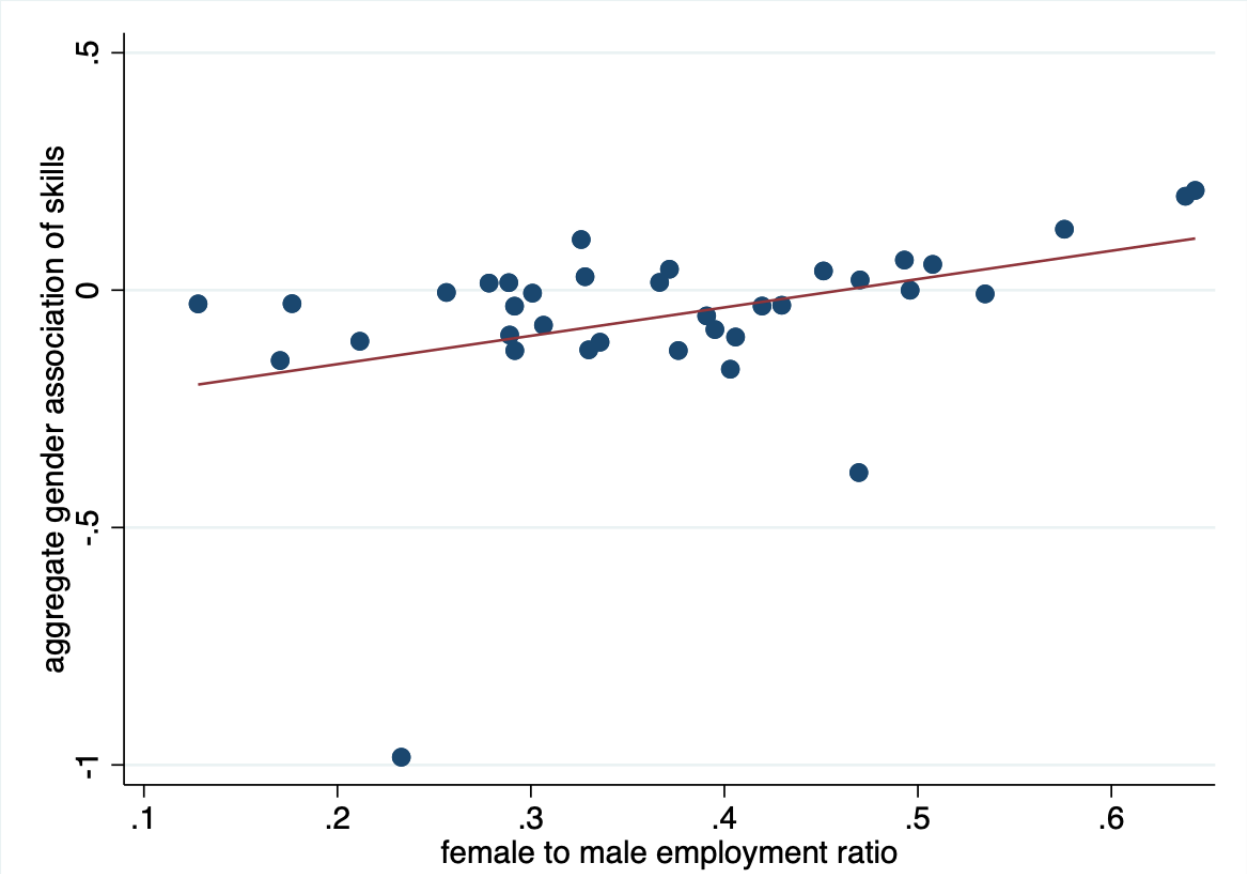


Figure 5: Gender associations in skills demand vs female-to-male employment. The y-axis shows the aggregate gender associations in skills demand in a state using NCS portal data. Higher values indicate a greater demand for skills that are associated with women vs men in a state. The x-axis shows the proportion of women employed relative to the proportion of men employed in a state. This percentage is calculated for women and men aged 15-59 with completed schooling who reside in urban India. We restrict our analyses to urban and school-educated individuals since the NCS portal largely caters to this sub-population. We use the Periodic Labor Force data for 2020–21 to calculate employment proportions; this is the most recent nationally representative data available for employment in India.

The urban Indian labor market is characterised by a low female Work Force Participation Rate (WFPR) at around 25% (Afridi *et al.*, 2019). However, there is considerable variation

in female WFPR across different regions or states of India. This could be related to variation in the demand for skills.

We investigate whether observed gender associations in skills demand correlate with the gap between female and male employment rates across Indian states. We first create a measure of aggregate gender association with skills demand in a state. We use the weighted mean of gender association in skills demand within each state such that every job ad is given the same weight. Therefore, for job ad j in state s having overall gender association in skills $CS^{diff}_{j,s}$ and mentioning location in n_j states including the state s, this measure is computed as follows:

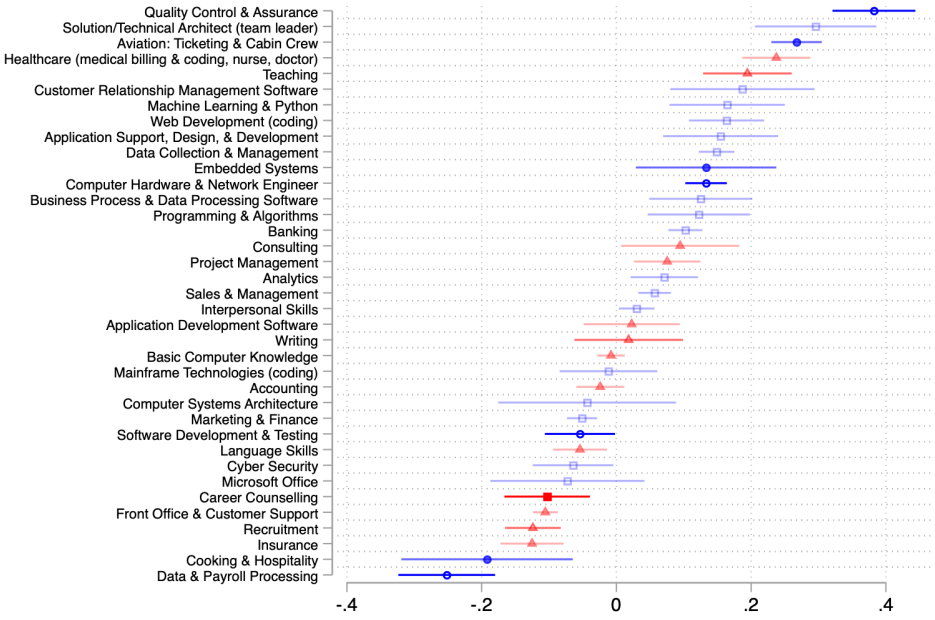
$$CS^{diff}_{s} = \frac{\sum_{j,s} CS^{diff}_{j,s}/n_j}{\sum_{j,s} 1/n_j}$$

An increasing value of the measure shows that, on average, job ads in the state require skills that are associated relatively more with women than men. Next, we estimate the ratio of female to male employment among individuals of working age (15–59 years) who have completed schooling and reside in urban areas, for each state. We then construct a scatter plot which gives the combination of aggregate state level gender association in skills demand with the female to male employment ratio for each Indian state in Figure 5. This scatter plot shows that states which have a gender association in skills demand favorable for females also tend to have a higher ratio of female to male employment, or there is a positive correlation between demand for skills associated with females vs males and relative female employment at the state level.

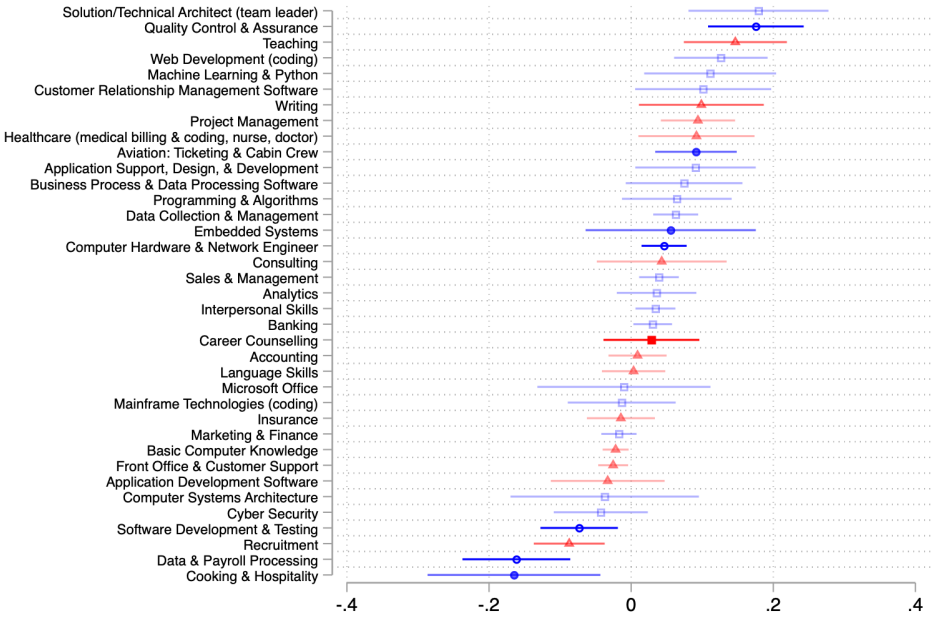
4.5 Posted wages

Figure 6 gives the estimated coefficients β_{1l} for the thirty seven skill groups (along with their confidence intervals) from estimation of equation (1).[31] It also shows how posted wages vary with gender association of skills. Male-associated skills are in blue while female associated skills are in red; darker shades of each color signify a stronger association with the respective gender they represent (or higher $|CS(i = l, female) - CS(i = l, male)|$). Panel (a) shows estimates after including state fixed effects while panel (b) estimates incorporate (state $\times$ occupation) fixed effects. Both panels control for education and experience requirements in a job ad, type and sector of the organization and type of job contract.

[31] In these results we are comparing the association of thirty seven different skill categories with log wages; in such cases, when the number of hypotheses being tested is large, multiple hypothesis testing is recommended. Hence, we also compute the sharpened False Discovery Rate (FDR) q-values to account for this and find that our results remain unchanged. These additional results are omitted for brevity but are available on request.



(a) Using State Fixed Effects



(b) Using (State × Occupation) Fixed Effects

Figure 6: Skills demand and posted wages. Coefficients are estimates of β_{1l} for each skill category from equation (1), after controlling for education, experience, type and sector of the organization and type of job contract. Darker red shades indicate a greater association of skills with women vs men while darker blue shades indicate a greater association of skills with men vs women. 95% confidence intervals are plotted for each coefficient.

Estimates in panel (a) indicate that eight of the top ten skill categories (in terms of log posted wages) are male-associated with $CS(i = l, female) - CS(i = l, male) < 0$, mostly belonging to the Information Technology domain; some of these include *Quality control & Assurance*, *Machine Learning Solution/Technical architect*, *Aviation*, *Application Support* and *Data collection and Management*. In fact, the skills associated with highest log wages have an extremely strong male association. Skills with a strong female association, or with $CS(i = l, female) - CS(i = l, male) > 0$, such as *Front office & customer support*, *Career Counselling*, *Recruitment* and *Language skills* have relatively low posted wages. Nevertheless some skills with a stronger male association (such as *Payroll processing*, *Cooking and Hospitality*, and *Software Development & Testing*) have low posted wages while others with a stronger female association (such as *healthcare*, *teaching* and *consulting*) have relatively high posted wages. Panel (b) shows that the differences in associations of skills with the log posted wage are attenuated after including occupation fixed effects which suggests that most of these differences can be explained by differences in skill requirements across occupations.

Table 2: Net female association in a job ad and the posted wage[a].

	(1)	(2)	(3)	(4)	(5)
CS^{diff}	–0.022***	–0.022***	–0.022***	–0.005**	–0.004*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	58068	58068	58068	58068	58068
Mean Y	11.565	11.565	11.565	11.565	11.565
Controls					
Job Ad Controls		✓	✓	✓	✓
State FE			✓	✓	
Occupation FE				✓	
State × Occupation FE					✓
Month-Year FE	✓	✓	✓	✓	✓

[a] The dependent variable is the logarithm of posted wage in a job ad. Job Ad Controls include the type and sector of the organization, type of job contract, required minimum qualification and experience specified in the job ad along with the square of required experience. Each column reports the effective number of observations after incorporating the included fixed effects. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

There might also be variation in the correlation between posted wages and gender associations of required skills *within* aggregate skill categories. Therefore, we check whether the overall correlation of posted wages with female associated skills is systematically different from male associated skills. Table 2 shows the estimation results from equation (2) where we estimate the relation between net female association in job ad j, CS_j^{diff} and log wages. Column (1) gives estimates with only month-year fixed effects while column (2) further

controls for education and experience requirements and type of job ad, sector of posting, and organization type of the posting firm. We further add state and occupation fixed effects in columns (3) and (4) respectively, and control for (state × occupation) fixed effects in column (5). The estimates show that there is a negative relationship between net female association of skills in a job ad and the posted wage. On average, posted wages are 2.2% lower if the net female association score increases by one unit when we control for the job requirements and variation in wages in a given location. Given that missing wage jobs are more likely to be relatively high-skill jobs, with high posted wages and a low net female association on average, our regression estimates using the sample of jobs with non-missing wages is likely to an under-estimate. Columns (4) and (5) show that after controlling for detailed occupation levels, the negative relationship between net female skill association and posted wages becomes attenuated though still remains marginally significant. On average, the posted wage decreases by 0.4% if the net female association score increases by one unit, after controlling for variation in wages within an occupation in a given location. These findings show that lower posted wages associated with female skills are mediated by distinct detailed occupational categories which contain these skill requirements.[32]

4.6 Firm size

Table 3 gives results from estimating equation (3) when the total number of skills demanded by a firm is the dependent variable. Columns (1)–(3) give results when the dependent variable is the total number of skills and an OLS model is used while columns (4)–(6) give results for the same dependent variable when a Poisson model is used for estimation. The estimates show that there is a positive association between the number of skills demanded and firm size. The OLS model estimates show that the number of skills demanded increase by 0.023, 0.012 and 0.0006 when postings, vacancies, and paid-up capital increase by one percent. This translates into an 0.81%, 0.435% and 0.022% increase in skills demand over the mean number of skills demanded (2.79) when postings, vacancies and paid-up capital increase by one percent, respectively. The Poisson model estimates shows similar results in columns (4)–(6) with slightly smaller elasticities—a one percent increase in firm postings, vacancies and paid-up capital increases skills demanded by 0.43%, 0.32% and 0.023%.

[32]These results continue to hold when we drop job ads that do not specify an education degree (Appendix Table A.6). While we continue to find that higher net female association of skills in job ads is associated with lower posted wages, the results within occupation and location become insignificant though the magnitudes are similar to our base specification. We also continue to find a negative relationship between net female association of skills in a job ad and the posted wage when using using 200, 250, 350, and 400 occupation categories obtained using k-means clustering. These additional results are available on request.

Table 3: Firm size and the number of skills demanded[a].

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS			Poisson		
ln(Total Post)	2.256***			0.430***		
	(0.048)			(0.009)		
ln(Total Vacancy)		1.215***			0.320***	
		(0.036)			(0.006)	
ln(Paid-up capital)			0.064***			0.023***
			(0.014)			(0.005)
N	11946	11946	11946	11946	11946	11946
Mean Y	2.786	2.786	2.786	2.786	2.786	2.786
Controls						
Firm controls	✓	✓	✓	✓	✓	✓
State FE	✓	✓	✓	✓	✓	✓
Business activity FE	✓	✓	✓	✓	✓	✓

[a] The dependent variable is the number of skills demanded by a firm. Columns (1)-(3) show the results from linear estimation. Columns (4)-(6) show the results from Poisson pseudo-maximum likelihood regression with the number of skills demanded as the dependent variable. Firm controls include organization type, industry classification and year of registration. We keep a set of firms that are matched to the Ministry of Corporate Affairs database and that make at least one job posting with required skills. Each column reports the effective number of observations after incorporating the included fixed effects. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4, panels A and B, give the relationship between firm size and demand for female and male associated skills. Columns (1)–(3) give estimates using an OLS model whereas columns (4)–(6) give estimates using a Poisson model. Clearly, demand for both female and male skills increases with firm size. In columns (1)–(3) while the absolute increase in skills for men with firm size is larger, the percentage increase (over the baseline mean) is slightly higher for female skills. Columns (4)–(6) confirm these findings and the percentage change for female skills with an increase in firm size is slightly larger in panel A when compared to that for men in panel B. Overall, an increase in firm postings, vacancies and paid-up capital by one percent increases demand for female skills by 0.45%, 0.34%, and 0.023%, respectively (Columns 4–6).

Additionally, Appendix Figure A.9 gives the estimated coefficients from thirty seven regressions for each skill category along with their confidence intervals from estimation of equation (3). Here the dependent variable equals one if a given category of skill is demanded by a firm and zero otherwise. It shows how firm size (in terms of total postings in the firm) is linked to skills demand by category and how this relation varies with the gender association of a skill. The color coding of the markers is similar to that of Figure 6. The estimates indicate that larger firms tend to have a higher demand for male-associated skills such as *Marketing & Finance*, *Sales & Management*, and *Data Collection & Management* and female associated

Table 4: Firm size and the number of female/male skills demanded[a].

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS			Poisson		
Panel A: Demand for female-associated skills						
ln(Total Post)	0.728***			0.450***		
	(0.018)			(0.010)		
ln(Total Vacancy)		0.396***			0.341***	
		(0.013)			(0.007)	
ln(Paid-up capital)			0.017***			0.023***
			(0.005)			(0.006)
N	11946	11946	11946	11942	11942	11942
Mean Y	.747	.747	.747	.747	.747	.747
Controls						
Firm controls	✓	✓	✓	✓	✓	✓
State FE	✓	✓	✓	✓	✓	✓
Business activity FE	✓	✓	✓	✓	✓	✓
Panel B: Demand for male-associated skills						
ln(Total Post)	1.529***			0.421***		
	(0.035)			(0.009)		
ln(Total Vacancy)		0.819***			0.311***	
		(0.025)			(0.006)	
ln(Paid-up capital)			0.047***			0.023***
			(0.010)			(0.005)
N	11946	11946	11946	11942	11942	11942
Mean Y	2.04	2.04	2.04	2.04	2.04	2.04
Controls						
Firm controls	✓	✓	✓	✓	✓	✓
State FE	✓	✓	✓	✓	✓	✓
Business activity FE	✓	✓	✓	✓	✓	✓

[a] The dependent variables are the number of female-associated skills demanded and the number of male-associated skills demanded in Panel A and Panel B, respectively. Columns (1)-(3) show results from a linear OLS estimation. Columns (4)-(6) show the results from Poisson pseudo-maximum likelihood regression with the number of female/male skills demanded as the dependent variables. Firm controls include organization type, industry classification and year of registration. We keep a set of firms that are matched to the Ministry of Corporate Affairs database and that have at least one job posting with required skills. Each column reports the effective number of observations after incorporating the included fixed effects. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

skills such as *Customer Support*, *Recruitment*, and *Accounting*. Whereas, the demand for male associated skills such as *Data and Payroll processing*, *Cooking and Hospitality*, *Aviation* and female-associated skills such as *Career Counselling*, *Teaching*, and *Writing* increases the least for larger firms. Appendix Figures A.10 and A.11 show the coefficient plots when the number of vacancies and paid-up capital of a firm are used as measures of firm size. We continue to observe broadly similar results.[33]

Table 5: Firm size and the demand for both male and female associated skills[a].

	(1)	(2)	(3)
ln(Total Post)	0.174***		
	(0.004)		
ln(Total Vacancy)		0.095***	
		(0.003)	
ln(Paid-up capital)			0.003**
			(0.002)
N	11946	11946	11946
Mean Y	0.325	0.325	0.325
Controls			
Firm controls	✓	✓	✓
State FE	✓	✓	✓
Business activity FE	✓	✓	✓

[a] The dependent variable takes the value one if the firm demands both male and female-associated skills and zero otherwise. The results are from a linear probability model. Firm controls include organization type, industry classification and year of registration. We keep the set of firms that are matched to the Ministry of Corporate Affairs database and that have at least one job posting with required skills. Each column reports the effective number of observations after incorporating the included fixed effects. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Lastly, Table 5 shows whether the joint demand for male and female skills changes with firm size. If bigger firms need both male and female skills, then the joint demand for both skills should increase in firm size. Clearly, we find that a firm is 0.174 percentage points (or ≈0.5% over the baseline mean) more likely to demand both male and female skills when the number of postings increase by 1%. Similarly, a one percent increase in vacancies and paid-up capital is associated with a 0.35% and 0.018% increase in joint demand for male and female skills. These results show that male and female skills are likely to become more complementary as firm size increases.

[33]As before, we also compute the sharpened False Discovery Rate (FDR) q-values to account for multiple hypothesis testing and find that our results remain unchanged. These additional results are omitted for brevity but are available on request.

5 Conclusion

Our work demonstrates how text analysis methods can be usefully employed to examine important questions that labor market researchers are interested in, for instance related to the demand for skills, and its role in earnings inequality as well as how this demand varies with firm size. In addition, we show how language models trained on online job ads can pick up subtle biases, or language associations in job ads. Thus, we highlight the need to exercise caution when using such models, for example, in content based recommendation systems.

We find that job ads which use female associated skills post a lower wage. This can contribute to a gender pay gap if women are more likely to apply for these jobs, indicating that promoting the use of gender neutral language can help in reducing gender segregation and earnings inequality. On the other hand, an increase in joint demand for male and female skills as firm size increases indicates that policies that promote larger firms may increase female employment. Although we focus on gender associations of skills in this study, domain specific word embeddings can also be used to examine ethnic or racial associations, or those related to older or disabled workers. As increasingly large amounts of labor market data become available in the form of text, these methods can be leveraged to study the nature of skills demand as well as its variation across regions and firm sizes.

References

Abraham, L., & Stein, A. 2020. *Words Matter: Experimental Evidence from Job Applications*. Unpublished manuscript.

Acemoglu, Daron, & Autor, David. 2011. Skills, tasks and technologies: Implications for employment and earnings. *Pages 1043–1171 of: Handbook of labor economics*, vol. 4. Elsevier.

Adenbaum, Jacob. 2022. Endogenous firm structure and worker specialization.

Afridi, Farzana, Bishnu, Monisankar, & Mahajan, Kanika. 2019. What Determines Women's Labor Supply? The Role of Home Productivity and Social Norms.

Ao, Ziqiao, Horváth, Gergely, Sheng, Chunyuan, Song, Yifan, & Sun, Yutong. 2023. Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions. *Information Processing & Management*, **60**(2), 103185.

Aquilina, Matteo, Klump, Rainer, & Pietrobelli, Carlo. 2006. Factor substitution, average firm size and economic growth. *Small Business Economics*, **26**, 203–214.

Ash, Elliott, & Hansen, Stephen. 2023. Text Algorithms in Economics. *Annual Review of Economics*, **15**.

Ash, Elliott, Chen, Daniel L., & Ornaghi, Arianna. forthcoming. Gender Attitudes in the Judiciary: Evidence from U.S. Circuit Courts. *American Economic Journal: Applied Economics*.

Balassa, Bela. 1965. Trade liberalisation and "revealed" comparative advantage 1. *The manchester school*, **33**(2), 99–123.

Banfi, Stefano, & Villena-Roldan, Benjamin. 2019. Do high-wage jobs attract more applicants? Directed search evidence from the online labor market. *Journal of Labor Economics*, **37**(3), 715–746.

Beaudry, P. andÂ D. Green andÂ B. Sand. 2016. The Great Reversal in the Demand for Skill and Cognitive Tasks. *Journal of Labor Economics*, **34(2)S1**.

Blau, F. D., & Kahn, L. M. 2017. The gender wage gap: Extent, trends, & explanations. *Journal of Economic Literature*, **55**(3), 789–865.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, & Mikolov, Tomas. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, **5**, 135–146.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y, Saligrama, Venkatesh, & Kalai, Adam T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, **29**.

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

Caliskan, Aylin, Bryson, Joanna J, & Narayanan, Arvind. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–186.

Campello, Ricardo JGB, Moulavi, Davoud, & Sander, Jörg. 2013. Density-based clustering based on hierarchical density estimates. *Pages 160–172 of: Pacific-Asia conference on knowledge discovery and data mining*. Springer.

Chakraborty, Pubali, & Mahajan, Kanika. 2023. *Firm Size and Female Employment*. Tech. rept. Ashoks Discussion Paper 103.

Chaturvedi, Sugat, Mahajan, Kanika, & Siddique, Zahra. 2021. Words matter: Gender, jobs and applicant behavior.

Chen, J., & Roth, J. 2023. *Logs with zeros? Some problems and solutions*. Unpublished manuscript.

Deming, David. 2017. The growing importance of social skills in the labor market. *Quarterly Journal of Economics*, **132(4)**.

Deming, David, & Kahn, Lisa B. 2018. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, **36**(S1), S337–S369.

Edwards, Aleksandra, Camacho-Collados, Jose, De Ribaupierre, Hélène, & Preece, Alun. 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. *Pages 5522–5529 of: Proceedings of the 28th international conference on computational linguistics*.

Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan, & Zou, James. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, **115**(16), E3635–E3644.

Gentzkow, M., Kelly, B., & Taddy, M. 2019. Text as Data. *Journal of Economic Literature*, **57(3)**.

Gonen, Hila, & Goldberg, Yoav. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *Pages 609–614 of: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Grave, Edouard, Bojanowski, Piotr, Gupta, Prakhar, Joulin, Armand, & Mikolov, Tomas. 2018. Learning Word Vectors for 157 Languages. *In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hansen, Stephen, Ramdas, Tejas, Sadun, Raffaella, & Fuller, Joe. 2021. The Demand for Executive Skills.

Hershbein, B., & Kahn, L. 2018. Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings. *American Economic Review*, **108**(7), 1737–1772.

Hidalgo, César A, & Hausmann, Ricardo. 2009. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, **106**(26), 10570–10575.

Kuhn, Peter, & Shen, Kailing. 2013. Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics*, **128**(1), 287–336.

Kuhn, Peter, Shen, Kailing, & Zhang, Shuo. 2020. Gender-targeted job ads in the recruitment process: Facts from a Chinese job board. *Journal of Development Economics*, 102531.

Marinescu, Ioana, & Wolthoff, Ronald. 2020. Opening the black box of the matching function: The power of words. *Journal of Labor Economics*, **38**(2), 535–568.

McInnes, Leland, Healy, John, & Astels, Steve. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, **2**(11), 205.

McInnes, Leland, Healy, John, & Melville, James. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, & Dean, Jeff. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, **26**.

Mitra, Aparna. 2003. Establishment size, employment, and the gender wage gap. *The Journal of Socio-Economics*, **32**(3), 317–330.

Ocampo, Sergio. 2022. *A Task-Based Theory of Occupations with Multidimensional Heterogeneity*. Tech. rept. University of Western Ontario, Centre for Human Capital and Productivity (CHCP).

Olivetti, C., & Petrongolo, B. 2016. The Evolution of Gender Gaps in Industrialized Countries. *Annual Review of Economics*, **8**(1), 405–434.

A Additional Figures and Tables

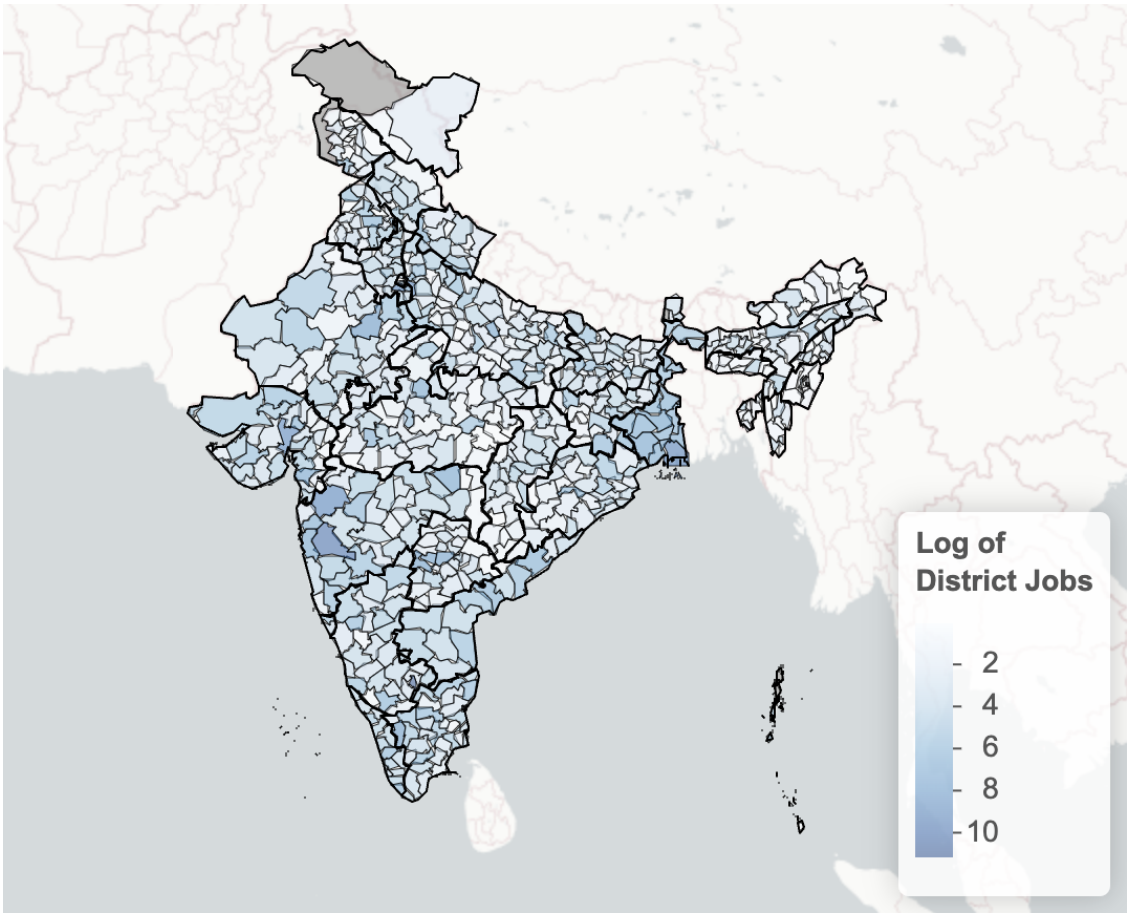
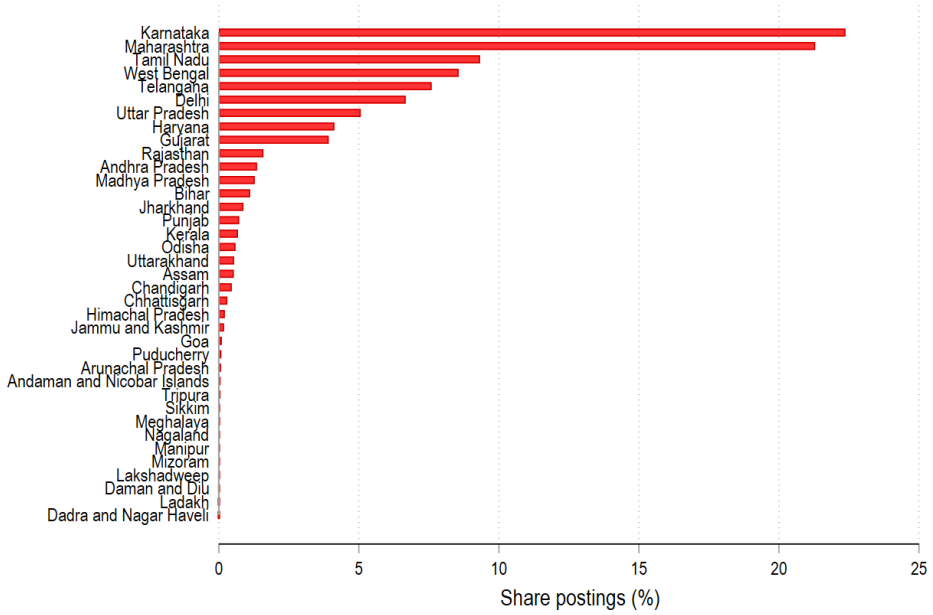
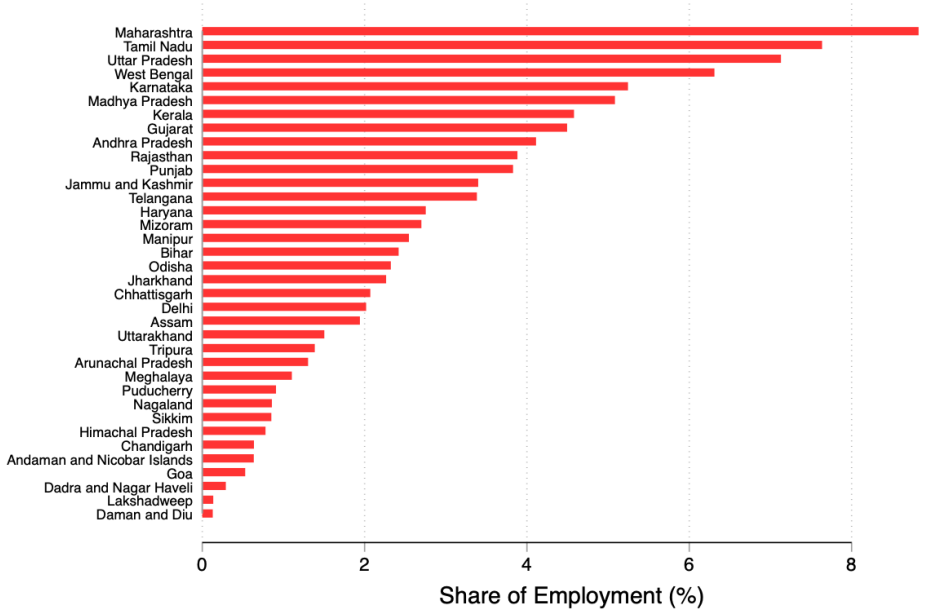


Figure A.1: Geographic distribution of online job ads across Indian districts posted on the NCS portal in India between July 2020 and November 2022.



(a) Job ads



(b) Employment

Figure A.2: The statewise distribution of online job ads posted on the NCS portal in India between July 2020 and November 2022 are given in the upper panel (a). The lower panel (b) reports statewise distribution of usual status employment in urban areas among working age population (15 to 59 years) from the Periodic Labor Force Survey (2020–2021).

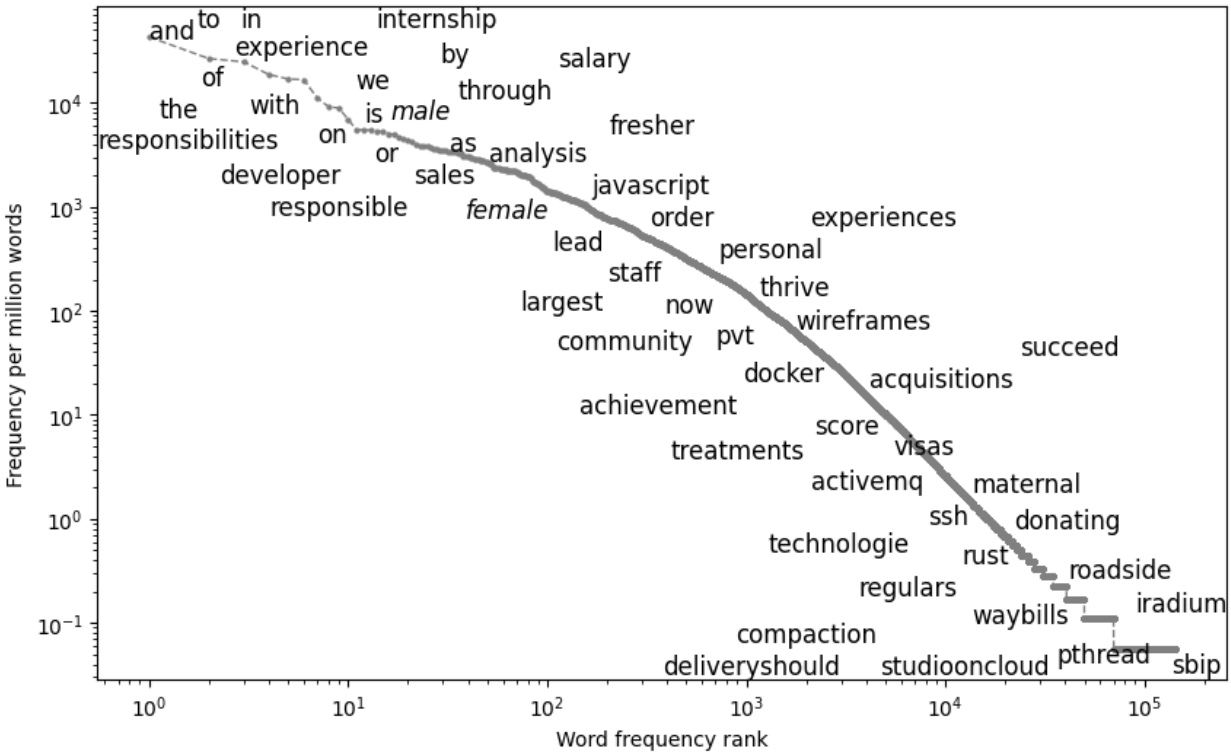


Figure A.3: Frequency per million words vs. word rank plotted on a logarithmic scale. This is based on word occurrences in the job descriptions corpus after removing duplicate job ads posted on the NCS portal in India between July 2020 and November 2022. Randomly chosen words in each rank bin are shown for illustration.

Job Id : 16Y61-1417009838206J | **Salary:** (₹) 100000 - 400000 (Yearly) | **Number of Openings:** 1 | **Posted on:** 06/02/2021 | **Last date to apply:** 28/02/2021

Company Name:	[redacted]	**Job Title**	Customer Relationship Executive
Organisation Type	Private	**Sector**	Manufacturing
Functional Area	Others	**Functional Role**	Others

Job Description

Requirement - Female tele caller in the age group of 22-30, graduate with good communication in English and Hindi, sufficient knowledge of computers. Soft spoken with pleasing mannerism with decent writing skills. Minimum 2-3 years? experience of working with a call centre. Accountability ? Talking to Consumer / Customer calls on Quality, Trade and other issues ? Email writing to the consumers ? Timely logging of calls in CRM and cascading them to stakeholders across the country for further action ? Follow up with stakeholders e.g Quality, Sales team, MU CREs on action taken ? Ensure closure as per turn-around time with appropriate comments and feedback from stakeholders ? Analyse nature of complaints and trends ? Follow up with external partners like Kankei and Autumn on escalations

Required Qualifications

Minimum Qualification Required: Graduate

Additional Information

Total Experience (in years)	2 - 7		
Job Location	Gurugram	**Key Skills**	Telecaller
Nature of job	Full Time		
Salary (₹)	100000 - 400000	**Salary/Wage Type**	Yearly
Gender Preferences	Any		
Ex-Servicemen preferred	No	**Number of Openings**	1

Figure A.4: Job ad with female preference (net female association or $CS_j^{diff} = 8.26$)

Job Id : 17P62-1211204503893J | **Salary:** (₹) 11500 - 12500 (Monthly) | **Number of Openings:** 50 | **Posted on:** 07/11/2022 | **Last date to apply:** 15/11/2022

Company Name:	[redacted]	**Job Title**	Hi Tech Lenses
Organisation Type	Company	**Sector**	Manufacturing
Functional Area	Operations and Maintenance	**Functional Role**	Operator

Job Description * Urgent Job Openings for (GKB) Hi - Tech Lenses* * Designation - Machine Operator (Associate)* * Shift Timings - 8 Hour Rotational Shift* * Double OT on Sundays* * Gender : Male Candidates can apply* * Salary: 11.5k to 12.5k In-hand + PF + ESIC with free accommodation* * Work Location: North Goa* * Experience: Fresher & Experience both can apply* * Qualification: SSC, HSC & ANY GRADUATES* * Number of Open Positions: 50*

Required Qualifications

Minimum Qualification Required: 10th Pass

Additional Information

Job Location	North Goa	**Key Skills**	Industrial Engineer Executive, Associate Operations Engineer
Nature of job	Full Time		
Salary (₹)	11500 - 12500	**Salary/Wage Type**	Monthly
Gender Preferences	Any		
Ex-Servicemen preferred	No	**Number of Openings**	50

Figure A.5: Job ad with male preference (net female association or $CS_j^{diff} = -5.02$)

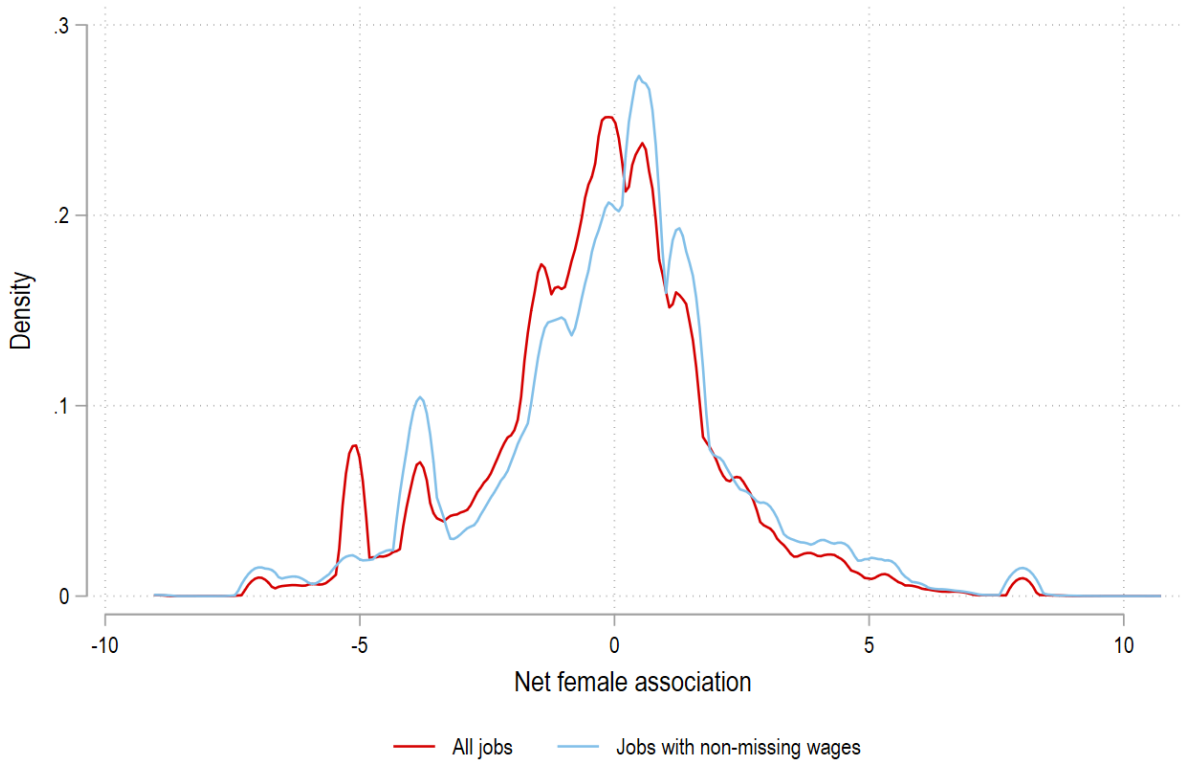
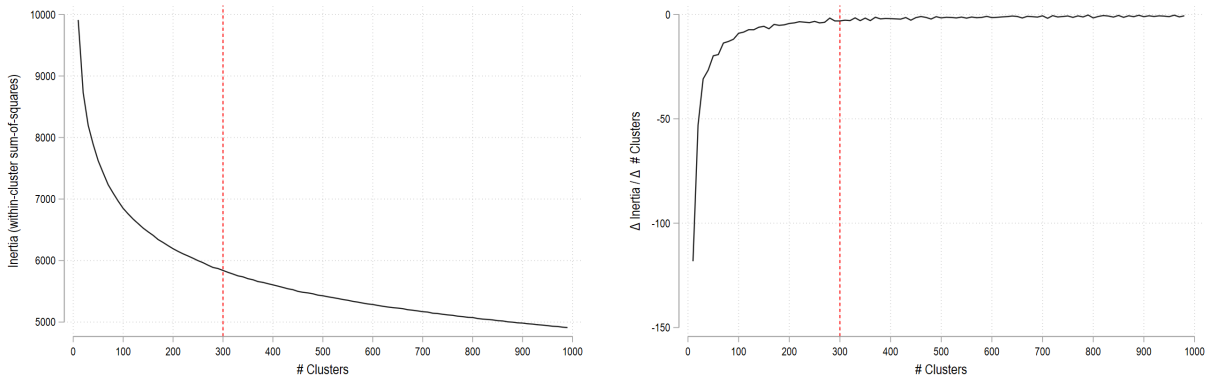


Figure A.6: Density function for net female association CS_j^{diff} for all job ads (in red) and after restricting the sample to ads with information on wages which are posted for a location within a single state (in blue).



(a) Inertia by number of clusters

(b) Change in inertia per cluster by number of clusters

Figure A.7: Panel (a) plots inertia (within-cluster sum-of-squares) against number of occupation clusters while Panel (b) plots the change in inertia per additional cluster obtained by applying k-means clustering on online job descriptions posted on the NCS portal in India between July 2020 and November 2022.

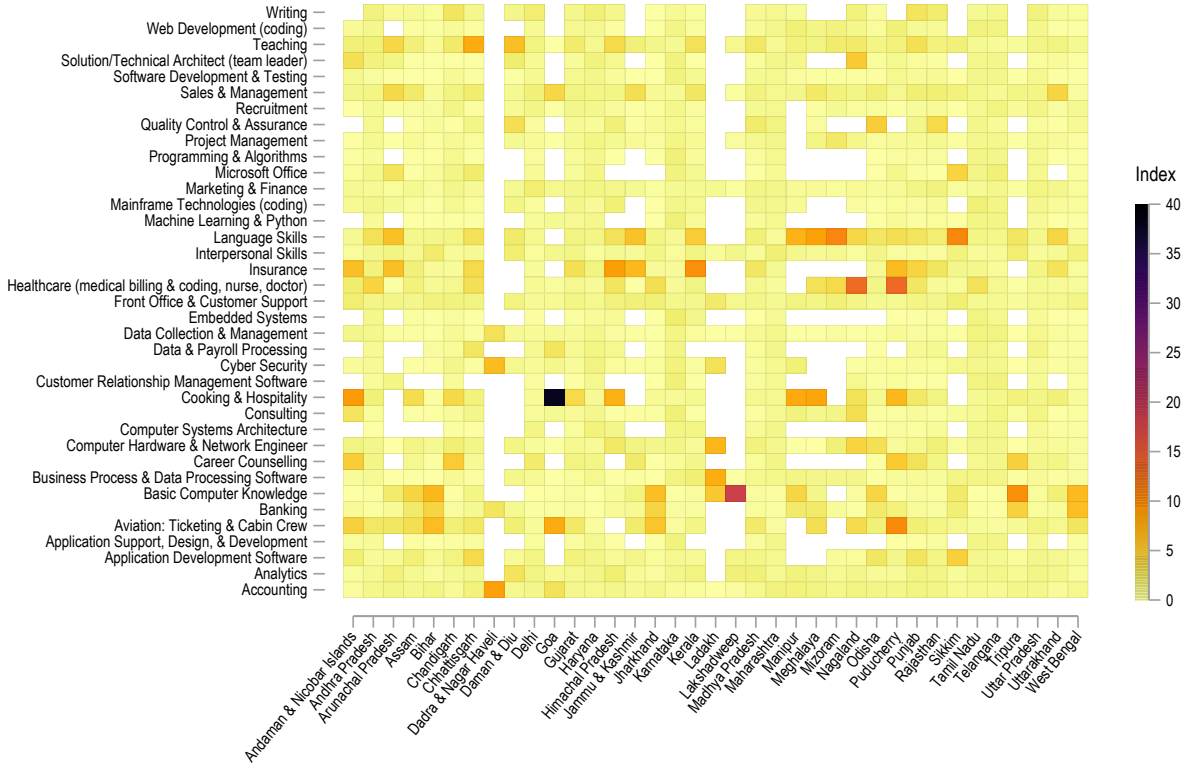


Figure A.8: Skills demand across Indian States. A heatplot of Balassa's Revealed Comparative Advantage (RCA) index, indicating relative skills demand, is given across Indian states/union territories. The heatplot is constructed using data on online job ads posted on the NCS portal in India between July 2020 and November 2022.

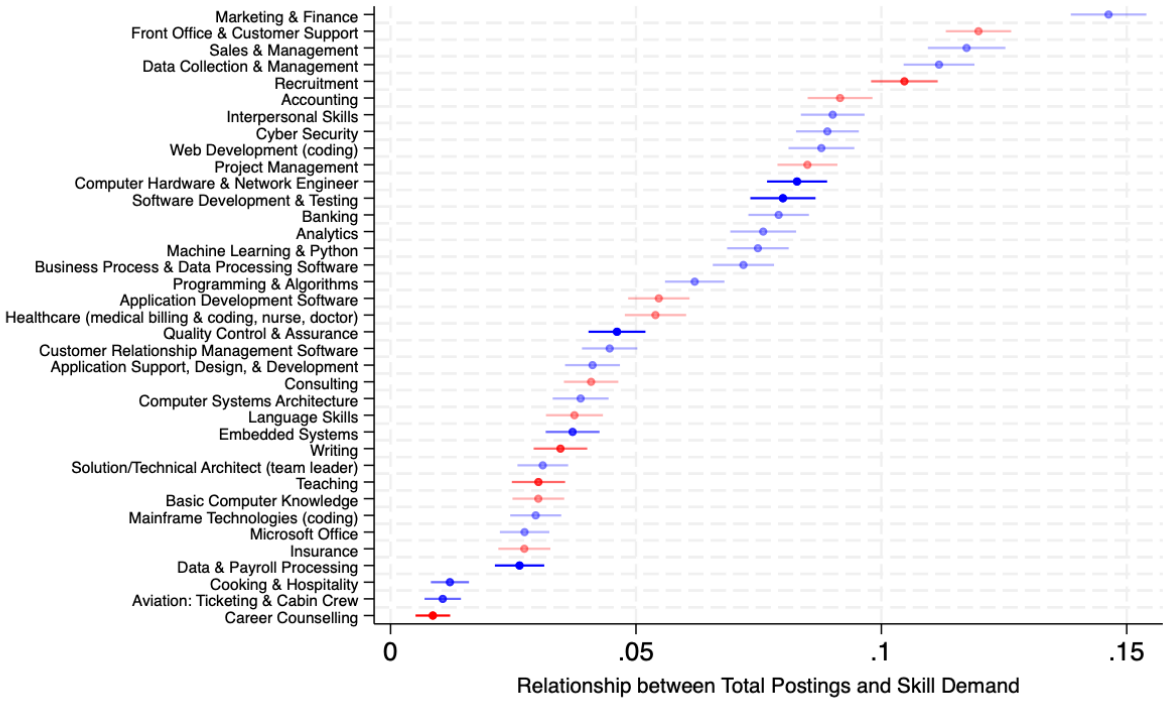


Figure A.9: Firm size (number of job ad postings) and skills demand, by gender association. The Figure reports coefficient estimates from estimation of equation (3) with a dependent variable that equals one if a given skill category (of 37 categories) is required by a firm. Estimated coefficients are for changes in the log of total job postings by a firm as the independent variable which measures firm size. We also control for the firm's industry, organization type, and year of registration. Darker red shades indicate a greater association of a skills category with females while darker blue shades indicate a greater association of a skills category with males. 95% confidence intervals are provided for each coefficient.

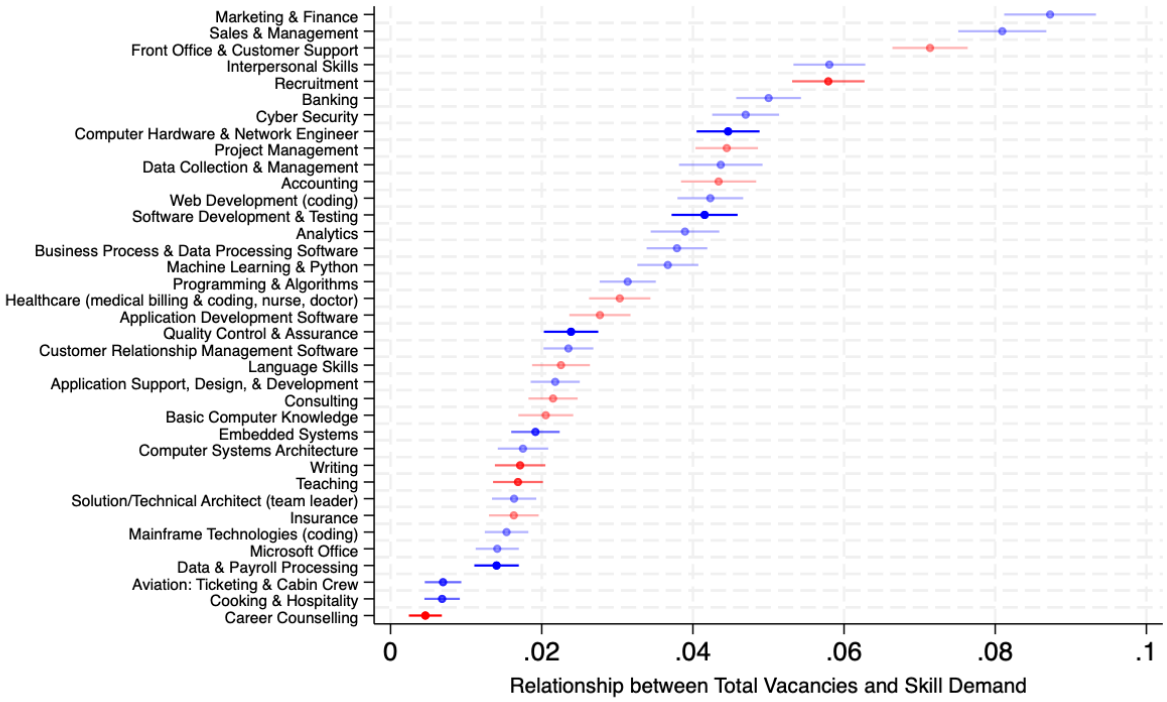


Figure A.10: Firm size (number of vacancies in job ads) and skills demand, by gender association. The Figure reports coefficient estimates from estimation of equation (3) with a dependent variable that equals one if a given skill category (of 37 categories) is required by a firm. Estimated coefficients are for changes in the log of total vacancies by a firm as the independent variable which measures firm size. We control for the firm's industry, organization type, and year of registration. Darker red shades indicate a greater association of a skills category with females while darker blue shades indicate a greater association of a skills category with males. 95% confidence intervals are provided for each coefficient.

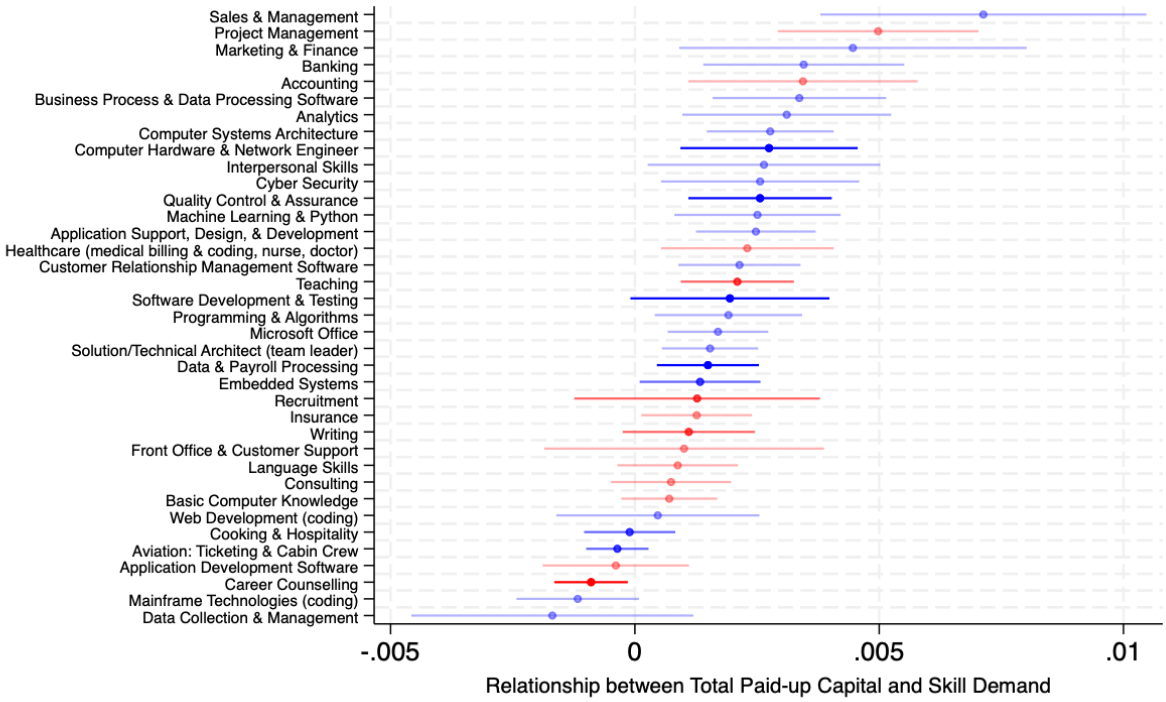


Figure A.11: Firm size (paid-up capital) and skills demand, by gender association. The Figure reports coefficient estimates from estimation of equation (3) with a dependent variable that equals one if a given skill category (of 37 categories) is required by a firm. Estimated coefficients are for changes in the log of paid up capital by a firm as the independent variable which measures firm size. We control for the firm's industry, organization type, and year of registration. Darker red shades indicate a greater association of a skills category with females while darker blue shades indicate a greater association of a skills category with males. 95% confidence intervals are provided for each coefficient.

Table A.1: Descriptive statistics (all jobs)[a,b].

	Mean	**SD**	**N**
No. of vacancies per ad	11.653	159.728	332044
Experience	4.490	6.640	313599
Yearly Wage	2,78,041	3,06,080	119740
Education:			
Secondary	0.023	0.150	332040
Senior Secondary	0.203	0.402	332040
Diploma	0.037	0.188	332040
Graduate	0.286	0.452	332040
Post Graduate and above	0.023	0.150	332040
Not Specified	0.428	0.495	332040
Organization Type:			
Government	0.024	0.152	331870
Private	0.432	0.495	331870
NGO	0.003	0.051	331870
Others	0.542	0.498	331870
Job Sector:			
Agriculture	0.002	0.048	330207
Construction	0.005	0.073	330207
Manufacturing	0.059	0.236	330207
Services	0.933	0.250	330207
Job Type:			
Full Time	0.810	0.393	332040
Internship	0.125	0.331	332040
Part Time	0.065	0.246	332040
Skills			
Skill requirement per ad	1.471	1.294	277700
Female Association	0.897	1.208	241799
Male Association	1.225	1.453	241799
Net Female Association	-0.328	2.265	241799

[a] Each cell gives the average value of a variable in the population of job ads for the observations that have a non-missing value of that variable. Wages are annual wages in Indian Rupees. Wages and experience are the mid-point of the range specified in the job ad. Number of positions advertised for in a posting shows the number of vacancies per job ad. Skill requirements show the average number of skills required by a job ad out of the 37 skills classified in the analyses.
[b] *Source:* Data from the population of all job ads.

Table A.2: Descriptive statistics (jobs with non-missing wages)[a,b].

	Mean	**SD**	**N**
No. of vacancies per ad	13.619	163.699	62958
Experience	1.489	2.595	61940
Yearly Wage	2,53,328	3,08,251	62958
Education:			
Secondary	0.024	0.152	62958
Senior Secondary	0.333	0.471	62958
Diploma	0.043	0.203	62958
Graduate	0.469	0.499	62958
Post Graduate and above	0.021	0.143	62958
Not Specified	0.110	0.313	62958
Organization Type:			
Government	0.003	0.058	62947
Private	0.688	0.463	62947
NGO	0.004	0.064	62947
Others	0.305	0.460	62947
Job Sector:			
Agriculture	0.004	0.067	62782
Construction	0.007	0.084	62782
Manufacturing	0.084	0.278	62782
Services	0.904	0.294	62782
Job Type:			
Full Time	0.584	0.493	62958
Internship	0.283	0.451	62958
Part Time	0.133	0.339	62958
Skills:			
Skill requirement per ad	1.419	1.090	62958
Female Association	1.081	1.404	59659
Male Association	1.137	1.506	59659
Net Female Association	-0.056	2.460	59659

[a] Each cell gives the average value of a variable in the population of job ads for the observations that have a non-missing value of that variable. Wages are annual wages in Indian Rupees. Wages and experience are the mid-point of the range specified in the job ad. Number of positions advertised for in a posting shows the number of vacancies per job ad. Skill requirements show the average number of skills required by a job ad out of the 37 skills classified in the analyses.

[b] *Source:* Job ads data after restricting the sample to those with information on salary, key skills and posted for locations within the same state.

Table A.3: Job characteristics and missing wages[a].

	(1)	(2)	(3)
Senior Secondary	0.228***	0.269***	0.269***
	(0.004)	(0.008)	(0.008)
Diploma	0.344***	0.448***	0.447***
	(0.006)	(0.009)	(0.009)
Graduate	0.275***	0.292***	0.292***
	(0.004)	(0.008)	(0.008)
Post Graduate and above	0.370***	0.265***	0.264***
	(0.007)	(0.012)	(0.012)
Education Not Specified	0.671***	0.682***	0.681***
	(0.004)	(0.008)	(0.008)
Experience	0.044***	0.039***	0.039***
	(0.000)	(0.001)	(0.001)
Experience^2	–0.001***	–0.001***	–0.001***
	(0.000)	(0.000)	(0.000)
Net Female Association		–0.007***	–0.008***
		(0.000)	(0.000)
Number of skills			0.002***
			(0.001)
Constant	0.062***	0.064***	0.060***
	(0.004)	(0.008)	(0.008)
N	313599	225584	225584
Mean Y	0.618	0.658	0.658

[a] The dependent variable takes the value one if the wage is missing in the job ad and zero otherwise. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.4: Descriptive statistics (jobs with additional firm-level information)[a,b].

Panel A			
	Median	**SD**	**N**
Firm Size:			
Total No. of Posts	2	154.649	12,000
Total No. of Vacancies	4	1532.555	12,000
Total Paidup Capital (million INR)	0.1	24,776.85	12,000

Panel B			
	Mean	**SD**	**N**
Firm Size:			
Total No. of Posts	10.164	154.649	12,000
Total No. of Vacancies	96.977	1532.555	12,000
Total Paidup Capital (million INR)	674.54	24,776.85	12,000
Organization Type:			
Government	0.005	0.072	11,975
Private	0.733	0.443	11,975
NGO	0.001	0.022	11,975
Others	0.262	0.440	11,975
Skill Demand:			
Female-associated skills	0.748	1.324	12,000
Male-associated skills	2.038	2.489	12,000
Female and Male skills (joint)	0.326	0.469	12,000

[a] Each cell gives the average value of a variable in the firm level data for the observations that have a non-missing value of that variable. The table shows the average number of postings, vacancies and paidup capital of the firm. The organization type shows the ownership type for the firm. Female associated skills show the average number of skills associated with a female that are posted in the skill requirements across all job ads by that firm. Male associated skills show the average number of skills associated with a male that are posted in the skill requirements across all job ads by that firm. At most these variables can take a value of 37. The total skills demanded by a firm are the sum of male and female associated skills. Female and Male skills (joint) shows whether a firm demands both male and female associated skills.

[b] *Source:* All job ads posted on the data portal which had a firm name that could be matched with the MCA database. The set of firms are those that report a non-missing paidup capital in the MCA data.

Table A.5: Gender attribute word frequency and similarity[a].

Word	# Job Ads	Share of job ads (%)	Cosine Similarity (%)
Panel A. Reference word: **female**			
female	23,012	6.93	100.00
females	449	0.14	62.46
woman	51	0.015	20.73
women	689	0.21	23.51
girl	46	0.014	31.23
girls	129	0.039	34.78
lady	49	0.015	26.14
ladies	30	0.009	23.49
feminine	13	0.0039	21.96
Panel B. Reference word: **male**			
male	19,736	5.94	100.00
males	185	0.056	57.99
man	446	0.13	13.26
men	145	0.044	28.04
boy	551	0.17	35.43
boys	158	0.048	39.27
gent	0	0	27.50
gents	2	≈ 0	19.44
guy	83	0.025	24.62
guys	286	0.086	20.88
masculine	0	0	27.16

[a] The frequency of gender attribute words in our job ads corpus and their cosine similarity with the words "female" and "male" is given in Panels A and B respectively.

Table A.6: Net female association and posted wages (robustness check after dropping job ads without education requirements)[a].

	(1)	(2)	(3)	(4)	(5)
CS^{diff}	–0.026***	–0.019***	–0.019***	–0.003	–0.003
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
N	51478	51478	51478	51478	51478
Mean Y	11.389	11.389	11.389	11.389	11.389
Controls					
Job Ad Controls		✓	✓	✓	✓
State FE			✓	✓	✓
Occupation FE				✓	✓
State × Occupation FE					✓
Month-Year FE	✓	✓	✓	✓	✓

[a] The dependent variable is the logarithm of the posted wage in a job ad. Job Ad Controls include the type and sector of the organization, type of job contract, required minimum qualification and experience specified in the job ad along with the square of required experience. Each column reports the effective number of observations after incorporating the included fixed effects. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.