Posted Wage Inequality*

Xuanli Zhu†

February 3, 2023
(Latest Version Here)

Abstract

This paper introduces a novel method to study the determinants of wage inequality in the labor market, utilizing online job posting data and machine learning algorithms. This method offers new insights by opening the black box of the multidimensional worker and job heterogeneity, which also enables to identify the most important skills and tasks that account for sorting between worker/job and firm components. Applying our method to the posting data of a Chinese online job board, we estimate different wage dispersion components and find clear evidence of firm wage premiums and positive firm-job sorting, with the shares consistent with those observed in many developed countries. During the estimation procedures, our machine learning approach reveals a data-driven skill and task structure featured by different levels of specificity. Variations in occupation-specific skills and tasks are the primary determinants of wage variances, contributing through both channels of job differences and sorting with firm pay policies. This is especially prominent in high-skill occupations, which are characterized by large wage dispersions. Meanwhile, in low-skill occupations, education- and experience-related skills and tasks, which are relatively less specific, play an equal or sometimes more important role. In contrast, variations in the most general skills, whether cognitive or interpersonal, have minimal impact on posted wage inequality.

Keywords: wage inequality, skills and tasks, firm wage premium, firm-worker sorting

*I am very grateful to Tetsuji Okazaki for his guidance throughout the development of this project. I thank Daiji Kawaguchi, Naoki Wakamori, Ryu Kambayashi, and participants in the seminar of University of Tokyo, the 4th Monash-Warwick-Zurich Text-as-Data Workshop, the "Labor, Firms, and Macro" Workshop, the Tokyo Labor Economics Workshop, The 17th Applied Econometrics Conference, and Asian and Australasian Society of Labour Economics 2022 Conference for valuable feedback and comments. I thank Adam Oppenheimer for his great help in using the Pytwoway package. This paper is split from the previous version of the working paper "Posted Wage and Compensation Inequality", which is further previously circulated with the title "Job Task Variation, Firm Wage Premium, and Firm-Provided Compensation".

†University of Tokyo, Graduate School of Economics. Email: zhuxuanli46@gmail.com.

1 Introduction

Workers are paid differently in the labor market. The determinants of wage inequality have long been a key research agenda of labor economists. One major econometric problem that economists often face when studying the labor market inequality is unobserved worker and job characteristics. A common approach to resolve this problem is to use fixed effects and panel data to control for those time-invariant confounding factors that are unobserved in the data. Abowd et al. (1999) (hereafter AKM) pioneered to use two-way fixed effects, i.e. both worker fixed effects and firm fixed effects, to separate wage inequality into different components. The following literature in general find that while workers' observed and unobserved characteristics ("worker effect") account for a majority share of the wage differentials, the different levels of firm wage premium ("firm effect") and the sorting between worker quality and firm wage premium ("sorting") are also important determinants of wage inequality (see the review in Card et al. (2018) and more recent papers discussed in Section 2). Despite these substantial progresses, relatively less is known about the more granular factors and determinants within these board components. For example, we know that the worker fixed effects capture those potentially high-dimensional heterogeneous skills of workers beyond formal education levels, but we know little what are those skills and what structures or features do they have. Without knowing the details about what is behind the worker effect, we will also have a limited understanding on how and why workers and firms are assortative matching with each other.

In this paper, we develop an alternative way to study the determinants of labor market inequality by looking at the online job market and show that this new method can bring important new insights on wage inequality and determination by directly exploiting those most granular wage drivers. In short, our approach applies machine learning algorithms to online job vacancy/advertisement data to distill all the wage-predictive information from the job description texts, which turns out to be mainly a large set of various skills and tasks, and to generate direct controls for the captured job characteristics so that we can replace the worker effect in the AKM framework with a "job effect".[1] The key idea behind this new approach is that while we cannot observe many important characteristics of workers in the census or survey data, in vacancy data firms actually document all the information about the job and the worker they are looking for—the skills that the worker should hold and the tasks that the worker will conduct—so that they can attract and match with their ideal worker. Moreover, firms' posted wages (often wage ranges) in their vacancies will reflect their valuations on these job characteristics, and also will work as the justification for their posted contents. This perspective is natural under the view of directed job search models where firms post wages and other job properties and workers direct their search on different submarkets segmented by different post

[1] To be specific, except skills and tasks, our algorithms also find a set of non-wage compensations that hold strong predictive power for the job wages. However, we show in the appendix that, after accounting for all the extracted skills and tasks, these non-wage amenities hold little explanatory power on the reported pay differences, and their predictive powers mainly come from their correlations with the worker effects and firm effects. Therefore, we will exclude these features from our analysis and focus on the features that directly link to workers' ability and productivity. In a companion paper Zhu (2022), we use the same data and methods to document some additional stylized facts about the non-wage compensation provision across different types of firms and jobs, and build a theoretical model to explain how different firms decide their compensation provision and what are the implications for the labor market inequalities.

contents.[2] As a result, we can replace the real worker with the posted job or the ideal worker, of which we have the full information that matters for the firms' wage determination observed, and replace the real wage with the posted wage as well.[3] The difficulty, however, is to correctly capture those useful information from the high dimensional text data and to bring them back to the otherwise typical econometric estimation. We tackle this task by taking advantage of a series of machine learning algorithms to automatically find various skills and tasks embedded in the job vacancy text and then to generate a set of low dimensional proxy variables for them. These proxy variables allow us to estimate the job effect, firm effect, and firm-job sorting of the posted wage inequality, which correspond to the worker effect, firm effect, and firm-worker sorting in the AKM framework.[4]

There are several advantages of our new method which make it complementary to the popular AKM-style two-way fixed effect approach used in many recent studies. First, while employer-employee panel data has been widely used in the recent literature with a main focus on rich countries, such data is often either unavailable or unaccessible in many developing countries. For example, in this paper we will study the labor market in China, where as far as we know there is no linked employer-employee data available and hence no studies on systematically investigation of the impact of both worker heterogeneity and firm heterogeneity on labor market inequality. In comparison, online job vacancy data is more easily to access and also more up-to-date, and although inevitably subject to some sample bias, it represents the major channel of matching between firm and worker in the recent labor market of many countries. Second, because in our approach there is no worker fixed effects but only firm fixed

[2]Although random search and wage bargaining have been a typical setting in the job search models, recently there are increasing evidences showing that directed search and wage posting is more realistic way to thinking about job search and wage setting in recent labor market (Banfi and Villena-Roldan, 2019; Marinescu and Wolthoff, 2020). In fact the fast development and the prevalence of online job boards in recent decades is itself the best evidence that firms and workers recognize such matching process as the efficient way to match the ideal counterparts. One potential concern may be if there could be misinformation or strategic information posting by firms. However, note that in an online job portal, a job post often attracts dozens or hundreds of applicants and a typical jobseeker also applies to dozens or over hundred different jobs. Given the resulted large screening costs (and opportunity costs), it is quite unlikely that firm will post wrong or misleading job information which would generate mismatches. Our assumption on wage setting is also consistent with recent empirical results that previous jobs have very limited impact on the starting wages in new jobs and that bargaining has an only moderate role on wage setting (see e.g. Di Addario et al., 2022; Lachowska et al., 2022).

[3]In other words, our implicit presumption is that the information documented in the job vacancies reflects the firms' true demand and pricing on the various worker and job characteristics, and, at least in expectation, represent the skills owned and the task conducted by the workers that the firm will eventually hire. Even if there exists mismatches between firm's idea workers and actually hired workers, our approach can still represent the true labor market demand and pricing in expectation as long as the level of such mismatches are not systematically different across different types of firms and workers.

[4]Because we replace the controls for worker characteristic in the AKM framework with the controls for job characteristics, the firm effect estimated in our method also does not hold exactly the same interpretation as the firm effect obtained in the AKM framework. To be specific, while the firm effect in the AKM framework is the firms' systematic pay differences after controlling for all worker characteristics, the firm effect in our framework is the firms' pay differences after controlling for everything documented in the job vacancies (except the firm name). The difference could occur when some job characteristics are rather firm-specific. For example if a firm pays higher wage because it assigns the otherwise similar workers with some specific and productive tasks that other firms cannot imitate and such specific tasks are documented in the vacancy text, then such higher wage level will be counted as firm effect in the AKM framework but as job effect in our framework. However, in practice we do not find such firm-specific job characteristics in our distilled job characteristics.

effects, the restriction of connected firm set and the contentious assumption of exogenous mobility that are required in the AKM approach are no longer necessary here. Also, the finite sample bias, which is stemmed from high dimensional fixed effects and known as the "limited mobility bias" in the AKM framework, will be moderate as long as firms in the data do no post too few vacancies. The last and perhaps the most important advantage is that through this new approach we can now open the black box of the worker effect in the previous studies and examine how important are different types of skills and tasks in accounting for the labor market wage inequalities. In a similar vein, it can also help to improve our understanding on the firm-worker sorting by examining what are the most important part of the job or worker characteristics that contribute to the firm-worker sorting.

In this paper, we apply our new approach to the job vacancy data of a national IT-centered online job board in China, which is called Lagou.com. In total, we collected over 6 million job vacancies posted on the website between 2013 and 2020, and use a bunch of cleaning procedures to obtain our final sample of 4 million vacancies for the main analysis. Due to the nature of the job board, one third of the job vacancies in our sample belong to Computer occupations like IT engineers or programmers, but the typical firms in our data also post a large amount of vacancies in other occupations including Design & Media, Business Operations, Financial, Legal, Sales, Administrative, etc. This allows us to study the job characteristics and the wage inequality both at firm level and across occupations with different skill levels. With fully acknowledging that our data can only represent a submarket but not the entire labor market in China, we posit that our approach can be easily applied to any other job vacancy data in any other countries, and that our analysis in this specific labor market uncover many important new facts about wage inequality that we believe hold general implications for other labor markets. The key information in the job vacancy for our analysis is the raw texts of job description in which employers document their skill requirements and task descriptions in order to match with their ideal workers. We also use the systematic information including the post wages and the requirements on education and experience that firms must enter or select from the website system when posting vacancies. Therefore, we are in fact using almost all the information that the potential jobseekers observe when they direct their job searches to study firms' wage posting behavior. To better illustrating both the intermediate results generated during our analysis and the main results on posted wage inequality, we conduct all the analysis on and show the results for both the pooled sample and three subsamples of different (major) occupations. These three subsamples/occupations are Computer, Design & Media, and Administrative, which are the typical high-, medium-, and low-skilled occupations in our data.

In order to distill the useful information embedded in the job texts and to generate the proxy variables, we apply a series of machine learning algorithms to the vacancy text data. The first step is feature selection. The aim is to limit the entire vocabulary of the job vacancy texts into a subset of words or terms (i.e. tokens or features in the machine learning or textual analysis terminology) that matter for wage determination. We achieve this by running a least absolute shrinkage and selection operator (Lasso) regression of posted wage on the token indicator vector of each vacancy and then selecting those features with nonzero estimated coefficients. To avoid overfitting and to reduce the randomness in the selection, we use the Bayesian information criterion (BIC) to tune the Lasso model. This procedure shrinks the entire vacancy vocabulary set of over 0.1 million tokens to a subset of only a few thousand, which mainly contain different types of skills and tasks. Although the estimation results in a

high-dimensional penalized model like Lasso are in general uninterpretable and not necessarily casual due to multicollinearity and high model-flexibility in the high-dimensional context, we verify our selected features through subsampling and sanity check and find that these features are rather robust and intuitive. Our second step then is feature clustering. We need this step because the still high-dimensional selected features are difficult for human understanding and a necessary procedure is thus mapping them to some low-dimensional concepts that we can easily understand. Our key deviation here from the previous literature is that we want to achieve this through a way that does not rely on any prior domain knowledge and let the text data speak for itself. In order to do so, we first train a natural language processing (NLP) model—the word embedding model—on our entire vacancy documents. The word embedding model learns the relationships between terms through the context of each term (i.e. the adjacent terms within a sentence) and represents each term in a latent embedding space based on such relationships. We then apply an unsupervised K-Mean clustering algorithm on this latent embedding space to classify our Lasso-selected features into eight clusters. In essence, the terms are gathered in the clusters based on if the employers talk them in a similar context in the job vacancy text. After inspecting the tokens within these auto-generated clusters, we use our human knowledge to label these eight clusters as the following: a cluster of general human capital terms on cognitive, noncognitive, and interpersonal skills; a cluster of terms about education and other relevant terms; a cluster of terms featured by experience- or position-related skills and tasks including managing, subordinating, or coordinating specific tasks; four clusters of occupation-specific skills and tasks which perhaps require occupational domain knowledge to assign a suitable name; and a cluster of compensations and amenities that we exclude from the our following analysis but are studied in the companion paper Zhu (2022). We interpret this structure as that our data-driven approach discovers a skill and task structure featured by different levels of specificity and confirm this claim by inspecting the occurrence frequencies of the features in different clusters across different occupations. The last step of our machine learning procedures is dimensional reduction, which is purely for computational reason. In particular, we split our feature indicator matrix that are used in the Lasso regression into submatrices based on our clustering results and then use the partial least squares regression (PLS) algorithm to transform each sub-matrix into a low dimensional representation with only three proxy variables, so that we can easily add them into the standard wage differential estimation.

We recognize the proxy variables obtained through above procedures as a full set of controls for the job characteristics—job tasks and (ideal) worker skills—that affect firms' wage posting. We then embed those proxy variables of skills and tasks into a posted wage regression along with education and experience requirement dummies and firm fixed effects, and conduct the variance decomposition to distinguish the job effect, firm effect, and firm-job sorting as well as further granular components within the job effect and firm-job sorting. Our main findings on the components of the post-wage inequalities are the following. First, our estimation on the pooled sample show that the total share of the wage variance can be accounted 45.0 percent by the job effect, 13.6 percent by the firm effect, and 14.2 percent by the firm-job sorting. The levels of the firm effect and firm-job sorting is consistent with the findings in the recent literature that use the employer-employee data in the U.S. and European countries and bias-corrected AKM approach, suggesting that at least in this high-end labor market in China, the composition of wage inequality is similar to other developed countries. Second, despite the fact that we extract way more job characteristics for the high-skilled sample of the Computer

occupation than the low-skilled sample of the Admin occupation, the estimation results show substantially smaller share from job effect and larger share from firm effect and firm-job sorting in Computer occupation comparing to Admin occupation. This result suggests that the firm wage premium and firm-worker sorting observed in the labor market are potentially linked with how different firms adopt different specific skills and tasks. Third, we find that while most of the explanatory power of the education dummies are absorbed by the proxy variables that directly extract education information from the job text, the experience dummies still account for nearly half of the job effect and the sorting between job effect and firm effect and are highly correlated with our proxy variables. We suggest that this is because our machine learning approach mainly extracts the extensive margins of different skills and tasks while the experience requirement can represent the intensive margins of those occupation-specific skills and tasks and thus complements to our proxy variables. Fourth, our further decomposition on the extensive margin show that those occupation-specific skills and tasks account for an important share of the job effect and firm-job sorting in the pooled sample and the high-skilled sample of Computer occupation, but their importance declines significantly in the low-skilled sample of Admin occupation. In comparison, experience- and position-related skills and tasks, which arguably have medium levels of specificity, account for a major share of the effects of the extensive margin in low-skilled Admin sample, and also have non-negligible effects in the pooled and high- or medium-skilled sample. However, those most general skills turn out to play little roles in explaining posted wage differentials, in spite of the fact that firms do mention these cognitive and noncognitive terms in their job ads. The third and the fourth findings in combination suggest that occupational specific skills and tasks are not only a key part of the potential job or worker differences that directly account for the posted wage inequalities but also a key factor that generates the assortative matching between firms and workers or jobs. Fifth, we find that our estimated firm effects can be partially explained by firm size and location dummies, which is again consistent with the estimated firm effects in the AKM framework. Finally, we conduct several robustness checks and show that our results are unaffected by the finite sample bias or the compositional differences across different samples.

To validating, generalizing, and extending of our main results, we also conduct four extensive analyses, including tests on more flexible posted wage regression specifications, a shortcut way of estimation, the heterogeneity of posted wage components across occupations, and the trend of the posted wage inequality in our data. We find that there are a large increase in the posted wage variances accounted by firm effect and firm-job sorting when we allow for occupation-specific firm wage policies, suggesting that firms are not necessarily to pay the same wage premiums to all types of jobs within the firm. In contrast, allowing for occupation-specific skill prices have limited effects on the shares of posted wage components, largely because of the large amount of occupational specific skills that are not overlapped across different occupations. Borrowing the idea of using unsupervised clustering algorithms to reduce the dimension of the estimation supposed in Bonhomme et al. (2019), we also develop a shortcut method of estimation by first clustering both firms and job posts into low-dimensional classes and then replacing the high-dimensional job characteristics and firm fixed effects with these classes. The job classification algorithm uses the embedding space generated from the work-embedding model introduced earlier, and can be seen as a way to classify the job vacancies into arbitrary numbers of "occupations". We find that while even fairly low number (say 10 or 20) of firm classes can give a good approximation, a substantially larger number of job classes are

required to approach the baseline results, further supporting that jobs are well distinguished by specific skills and tasks. Similar to the results in Bonhomme et al. (2019), we also find that there are limited evidence of quantitatively important complementarities in our posted wage data. Relying on our job classification algorithm developed for the shortcut estimation, we can now generate enough numbers of different occupations for statistical analysis and examine how the posted wage components change across different occupations. We find that the higher the mean wage of the occupations, the larger the variance and covariance values of both three main components of the posted wage dispersions. And this positive relationship is more significant for the job effect and firm-job sorting, further suggesting the potentially important role played by those specific skills and tasks in high skilled occupations. Finally, consistent with many studies using administrative data in developed countries, we also find increased wage inequality in our job vacancy data. The major contributor of the increased posted wage variance in our data is the increased sorting between job qualities and firm pay policies, and through our decomposition, we find that, again, those specific skills and tasks contribute the most for this increased sorting.

The outline of the rest of our paper is following. In next section we discuss the related literature and our contributions. Section 3 introduces our data. In Section 4 we set up our econometric model and discuss relevant issues on the estimation. In Section 5, we apply a series of machine learning approaches to the vacancy text data to exploit wage-predictive information and generate proxy variables. Section 6 shows our main results on the posted wage inequalities as well as a bunch of results from robustness checks, and Section 7 conducts the extensive analyses. Finally, we provide some concluding remarks in Section 8.

2 Related Literature

This paper mainly links to and contributes to two board literature that focus on two major determinants of the wage inequalities in the labor market, namely the worker heterogeneity and the firm heterogeneity (in fact three but the third element worker-firm sorting is a natural derivation of study of firm heterogeneity). While these determinants are often studied separately, our integrated examination here shows that they are in fact closely linked with each other, and thus it is important to investigate their interactions for a better understanding on the mechanisms behind wage inequality and determination.

The first literature strand is the voluminous literature that use heterogeneous workers to explain the wage inequalities in the labor market. Since Mincer (1958) and Becker (1964), human capital, whether general or specific, has long been recognized by economists as the main factor behind wage differences. However, observed worker characteristics, including education, experience, occupation, and other demographic factors, often explain only a fraction (around 30%, Mortensen (2005)) of the total wage variation in a typical wage regression. While various types of specific human capital like industry- or occupation-specific human capital have been examined in the early literature, more recently, there is a converging consensus in the labor literature that occupation and industry categories are just serving as measurable proxies for the underlying tasks performed and skills required across different jobs and firms, and that multidimensional task-specific skills are the most natural way in thinking about hu-

man capital (see Sanders and Taber (2012) for an early survey on this literature). Following this idea, the recent empirical studies have begun to stress on the importance of multidimensional skills and tasks for wage determination and discrepancies (see Spitz-Oener, 2006; Autor and Handel, 2013; Deming and Kahn, 2018; Yamaguchi, 2012; Lise and Postel-Vinay, 2020, among others). Despite this surging popularity, in practice the entire space of the multidimensional tasks and skills are often classified and compressed into a very limited number of pre-determined board and abstract dimensions of cognitive, social, abstract, manual, routine, etc. And the potential specificity of skills and tasks (i.e. the necessary width of the space), though once discussed intensively in the literature, are now often completely circumvented. As a result, the examination of multi-dimensional skills and tasks in many studies are often eventually constrained in a pre-defined and fairly low dimension, which puts even distinctive occupations into very similar positions.[5] Also, many studies on multidimensional skills and tasks have limited their attention on between-occupation skill and task variations, potentially due to data limitation, even though the recent empirical studies find clear evidences of within-occupation task or skill variations and their significance in wage prediction (see e.g. Autor and Handel, 2013; Deming and Kahn, 2018). Our paper thus contributes to this literature by developing a method to investigate the indeed high dimensional skill and task space spanned both between occupations and within occupations with no priors holding on what are the most important dimensions. Indeed, we let the online vacancy text data to tell us what are the structure of the skill and task space based on how employers document the skills and tasks about their jobs. Our approach thus generates a data-driven skill and task structure, and it turns out that this structure is distinguished by different levels of specificity. Our following estimation results show that it is those most specific skills and tasks that play the most important role in accounting for the posted wage inequalities. On the other hand, general skills, whether cognitive, interpersonal, or noncognitive, matter little for the posted wage differential in our data.[6] Therefore, our results suggest that the specificity is still an important dimension when considering high dimensional skill and task variations, and those highly specific skills and tasks

[5]Although such simplification has been proven very useful in studying some labor market issues including what types of workers are or will be substituted by machines or robots or AI, it can be potentially misleading when studying worker heterogeneity and wage differences. It is because that the wage determination in the labor market are potentially based on very detailed skills and tasks which, even if similar in a low dimension, can be completely different and thus largely nontransferable at a high dimension. Such distinction is particular important when thinking about issues like job assignment, job mobility, and human capital investment, all of which could be potentially important for wage determination. For example, skill indexes calculated in those low and board dimensions often recognize that economists and biologists or electronic engineers have very similar skill compositions and skill levels. Consequently, using an analogy similar to the one in Sattinger (1993), those low dimensional skill index would indicate that this paper can be equally written by a biologist or an electronic engineer. More recently, Frank et al. (2019) suggests that studying skills and tasks with further increased specificity could provide better insights even for understanding the technological impact on labor market.

[6]Note that here we are not arguing that those cognitive, interpersonal, or noncognitive skills and their relevant dimensions, which have been extensively used in the literature, are not important or uninformative in wage determination. Actually occupation-specific skills and tasks (or skills and tasks with any levels of specificity) can be classified as cognitive or interpersonal or etc., and thus the specificity that we stress here is just another dimension that is orthogonal to those previously studied board dimensions. These different dimensions could have different importance when facing different economics questions about the labor market. Moreover, although those most general skills turn out to be not important in our results, these general skills can be important for workers' developing their occupational specific skills and workers' wage changes within firms (since these general skills are likely to be cheap talks and firms need time to confirm them).

are especially important when thinking about within-occupation skill and task variations.

The second closely related literature is a recently booming literature on estimating the firms' role in wage inequalities at both cross-sectional level and at chronological level (see Abowd et al., 1999; Card et al., 2013; Barth et al., 2016; Song et al., 2019; Bonhomme et al., 2020, among others).[7] In order to overcome the unobserved worker abilities and characteristics, these papers use linked employer-employee panel data and both worker and firm fixed effects to estimate and decompose the entire wage differential into worker effect, firm effect and sorting between firms and workers. Although the initial results of AKM show no evidences for firm-worker sorting, more recent studies equipped with better data and bias correction methods generally find that both the firm wage premiums and the assortative matching between firms and workers are important to account for wage inequality.[8] Our paper contributes to this literature by providing an alternative method to deal with the problem of unobserved worker characteristics and to estimate the firm pay differences. Instead of estimating the worker effect, we apply machine learning methods to online job vacancy text data to generate a full set of controls on the firm-documented job skills and tasks and then estimate a job effect as a replacement. The estimated wage components using our Chinese IT-centered job vacancy data are consistent with those found in the previous literature that use employer-employee panel data and AKM approach (see Bonhomme et al., 2020). Moreover, our method allows us to open black box of the worker fixed effect in the AKM framework and to examine what are the important skills and tasks that contribute to the sorting between workers and firms. Our estimation results find that those occupation-specific skills and tasks contribute for a major amount of firm-job sorting in our pooled sample and in high-skilled computer occupations, while experience- and position-related skills and tasks are the most important drivers of firm-worker sorting in low-skilled administrative occupations.[9]

Given that the model structure in the AKM approach are quite restrictive, some recent studies have tried more flexible specifications to examine its validity. For example, Bonhomme et al. (2019) designs an alternative way of estimation by first clustering all firms into small numbers of classes using the information of within-firm wage distribution, and then estimating the worker classes and the conditional wage distributions through a flexible statistical model. In our extensive analyses, we borrow this idea and develop a shortcut way of estimating by directly classifying the job clusters through the representations of the job vacancies in the embedding space of a word-embedding model. Our results show that different from the firm

[7]Before the pioneered work of AKM, labor economists had discovered strong evidence of significant and consistent wage differentials at industrial level even after controlling for all observed worker characteristics. This stylized fact called for many theories to explain, and one major explanation at that time was efficiency wage theory which generally argues that high wage can elicit workers' effort or avoid turnover costs. For detail see for example Krueger and Summers (1988) and Katz (1986). Since AKM, the main focus of the literature has turned into the differentials in firm level wage premiums.

[8]See Card et al. (2018) for the review on the findings in this literature. For recent improvements in econometric methods, see Kline et al. (2020); Bonhomme et al. (2019) and also the comparison of different methods in Bonhomme et al. (2020).

[9]Note that another feature we observe in our results is that high-skilled professional occupations have significantly more shares of wage differentials accounted by firm effects and firm-worker sorting, and thus contributing more to the aggregate results of firm effects and sorting in the pooled sample. Therefore, our results in combination suggest that those specific skills and tasks in those high-skilled professional occupations are perhaps key for understanding the firm-worker sorting in the labor market, either in terms of cross-sectional levels or trends over time.

classes that can work as a good approximation with with a very small number, the job classes requires a larger magnitude in the number to well distinguished all the job types. We also find that consistent with the real wage components in the administrative data, our posted wage data also show little evidence for an important quantitative role played by the complementarity between firm wage posting policies and job qualities. Another more straightforward way of relaxing the AKM approach is to allow for occupation-specific firm wage policies, as shown in Torres et al. (2018); Hou and Milsom (2021). We find that the evidences of firm pay policies varying across occupations also exist in our posted wage data. Moreover, we find that the all three posted wage components, namely job effect, firm effect, and firm-job sorting, are increasing in the mean wage of occupations, which is further consistent with the results found in Hou and Milsom (2021), and we suggest that this positive correlations are likely to be stemmed from increased use of specific skills and tasks in those high skilled and high wage occupations.

Finally, beyond these two board literature, our paper also contributes the recent literature that use vacancy data to understand demand of skills and tasks and/or wage inequalities in the labor market (Hershbein and Kahn, 2018; Deming and Kahn, 2018; Deming and Noray, 2020; Marinescu and Wolthoff, 2020; Atalay et al., 2020; Braxton and Taska, 2020; Bloesch et al., 2021, among others). Different from many of these previous papers that study pre-defined skill and task categories, our new method here illustrates a new way of utilizing online job vacancy data to study the multi-dimensional skill and task variations in workers and jobs.[10] Another important contribution of our paper to this literature is that we show the possibility and usefulness of using online vacancy data to study not only the skill and task demands but also the firm wage policies and firm-worker or firm-job sorting in the labor market. Moreover, we show that there are clear linkages between these different components of wage differentials in the labor market, and thus an integrated framework is likely to be required to fully understand the determinants and drivers of wage inequality. In these senses, the closest papers to us in the literature of online job vacancy data are perhaps Marinescu and Wolthoff (2020) who show that skill and task information in the vacancy text combined with the firm identifier can account for a majority of wage differentials, and Bloesch et al. (2021) who show that different occupations have different levels of skill specificity affecting firms' wage policies.

3 Data

In this paper, we use the vacancy data from a Chinese online job board called Lagou.com, which is the first and the largest information technology (IT)-centered online job board in China. The Lagou website starts its service at 2013 and grows fast by specializing on the labor market towards the relatively less-experienced workers in the Chinese Internet industry and acquiring

[10] In particular, in this paper we do not hold any such prior and let job text data to tell us what are the structure of skills and tasks that potentially matter for wage determination. Part of the reasons that why earlier papers use pre-defined skill and task indexes, in additional to the huge impact of previous multi-dimensional skill and task studies that we have talked above, may be that several pioneer works in this literature use the online vacancy data from the Burning Glass Technology firm which has already constructed occupation information and skill indexes but has the procedure of the construction keep in secret.

a large customer base of both IT-producing and IT-using firms.[11] Until the end of 2020, about 8 million vacancies have been posted on the website, and we successfully collected the information of over 6 million vacancies between 2013 and 2020.[12] For each vacancy we observe the information of the job name, the wage range, the job location and address, the education level and experience years required, if full-time or part-time or intern, the job descriptions on the tasks and skills required, the job benefits or firm amenities, the firm name, the firm industry category, the firm size category, and the posted time of this vacancy.[13] Different from many other papers in the literature that use pre-processed skill indexes or generate specific skill indexes based on a pre-specified dictionary of terms capturing certain pre-defined skill categories, we will fully utilize the raw text data of each vacancy's job descriptions by adopting a completely data-driven method to distill and classify the important skills and tasks embedded in the text.

One inevitable drawback of any vacancy data is that it does not constitute the whole labor market in an economy. It is well known and fully discussed in the literature that in most if not all cases the job vacancy data are biased to high skilled and high education jobs, to internet-related jobs, to jobs from large firms and in large cities, and to jobs targeting young or less-experienced workers.[14] Given that our data here is a highly professional part of the online job market, the labor market we study is thus even more biased in this way than other vacancy data. To be specific, our vacancy data is mainly composed a variety of jobs required from 0 to 10 years experience posted by Chinese IT-producing firms and IT-using firms, a majority of which locate in large cities in China.[15] One third of the vacancies in our data belongs

[11]The slogan of the website (https://www.lagou.com/) is "Find an Internet Job—Go to Lagou Recruitment". In 2017, 51job, a leading provider of integrated human resource services in China listed in the NASDAQ stock market and also the owner of one of the largest general online job board in China, announced that it will acquire a 60% equity interest in the parent company of Lagou for $119 million because they think the IT labor markets that Lagou specializes on will be a large complement to their general job board.

[12]The amount of posted vacancies per year grows over time along with the growing popularity of the website. As a result, vacancies between 2013 and 2016 account for less one third of the data, and vacancies between 2017 and 2020 accounts for over two thirds of the data. Our scrapper successfully collected around 60 percent of the vacancies for the 2013-2016 period and over 80 percent of all vacancies posted in the 2017-2020 period. In Appendix A.1 we explain the details of the data collection and show the patterns of both collected vacancies and the missing vacancies over time.

[13]A sample of the job vacancy posted in Lagou can be found in Figure A2. The information of the wage range, the requirements on education and experience, whether full-time or not, and the job location is either selected by firms within given choices provided by the website or be filled in with certain formats when posting the vacancy. This setting ensures that almost all the vacancies in our data have the unambiguous information on the level of post wage and the required education and experience, making it straightforward to generate consistent job variables. In contrast, the format of the job name and the descriptions on job skills, tasks, and amenities is arbitrary and as a result these text contents vary in the length and structure and are often entangled together and hard to distinguish. For example, while there is a certain space for entering job or firm amenities, firms sometimes also mention amenities along with job skill or task descriptions or, in the inverse cases, mistakenly write skill or task terms in the space of amenities. This problem, which is often seen in the real-world text data, partially incentivizes our machine learning methods introduced in the later sections, i.e. we will simply combine all the descriptions on job skills, tasks, and amenities as one integrated text for each vacancy and conduct textual analysis to distinguish the different information types of different words or phrases.

[14]See the relevant discussion in Kuhn and Shen (2013) for the Chinese vacancy data and the discussion in Hershbein and Kahn (2018) for the U.S. vacancy data.

[15]IT-using firms mainly incorporate firms in a variety of industries in the tertiary sector like finance, real restate, retail, etc.

to Computer occupations, and the other two-thirds of the jobs come from both other professional occupations like Design & Media occupations, Business Operation occupations, Financial Occupations, and Legal Occupations, and low-skilled occupations like Sales occupations and Administrative occupations.[16]

Given the popularity and the low charge of the website, these IT-producing and IT-using firms are likely to post all types of jobs that they demand, allowing us to study the firm-level wage premium. In our main analysis and results, we will show both the result for pooled sample including all vacancies along with the results for three typical major occupations in our data: Computer occupation, Design & Media occupation, and Administrative occupation. We pick these specific occupations because they are the representative high-, middle-, and low-skill occupations in our data and thus allow us to study how the skill composition and wage determination vary across occupations with different level of skills that are normally defined in the literature. Unlike Computer occupations, the jobs in Design & Media and Administrative occupations are largely confined to those jobs in the IT-producing or IT-using companies and may not be representative for those occupations in the entire labor markets. With full awareness of the limited coverage of our vacancy data, we argue that our aim in this paper is to illustrate how our new method can be applied to vacancy data and provide new insights on wage determination, and we anticipate our results being examined or validated under other vacancy data in other labor markets or other countries.

We explain our method of occupation classification, describe our sample cleaning procedures, and show the summary statistics of our final sample of 4 million job vacancies used for analysis in Appendix A.2.

4 Econometric Setting

In this section, we set up our baseline econometric model by following the literature of wage differential (Abowd et al., 1999; Card et al., 2013; Barth et al., 2016; Song et al., 2019) and discuss the relevant empirical issues when conducting the estimation. One major problem of the estimation is the unobserved worker or job characteristics that motivates the use of panel data and fixed effects in the literature. We suggest that we can resolve this problem by exploring the information embedded in the job texts (Section 5) and assuming that all the information about the job and the potential worker used by firms to determine their posted wage is documented on the job vacancies. We will also discuss how the issues of exogenous mobility, limited mobility bias, and mis-specification, which are the common problems in the AKM literature, should be considered in the case of vacancy data. This discussion also illustrates a hidden correlation between our approach and the AKM approach: both approaches estimate the wage dispersion through the information of job mobility or new jobs.

[16]Here we define these occupations by following the U.S. SOC classification 2018 and using most board two-digit or three-digit categories. For example, our "Computer occupations" refers to the 2-digit "15-0000 Computer and Mathematical Occupations" and it will incorporate all the occupations in the further 3-digit "15-1200 Computer Occupations" but only some occupations like data scientists in the 3-digit "15-2000 Mathematical Science Occupations". We explain the details of the occupation classifications below and in Appendix A.3.

Our baseline specification of the log wage regression is

$$\ln w_{i,j,t} = X_i\beta + \psi_j + \iota_t + \epsilon_i \tag{1}$$

, where $j \equiv j(i)$ is the firm that posts vacancy i, and $t \equiv t(i)$ is the year that the vacancy was posted. X_i is a vector of job vacancy characteristics, which can incorporate the traditional observed worker characteristics like education, experience, and occupation, as well as other usually unobserved worker and job characteristics that are collected in alternative ways.[17] β is the coefficients for job characteristics, ψ_j is the firm fixed effect, ι_t is the year effect, and ϵ_i is the residual wage.[18] After correctly estimating the model, we can then conduct variance decomposition such that we can divide the total dispersion of the posted wage in our data into components of the job characteristics, the firm pay policies, and the sorting between them. From Equation (1), and by ignoring the year effects ι and denoting $X_i\beta \equiv \theta_i$, we obtain

$$\text{var}(\ln w_i) = \underbrace{\text{var}(\theta_i)}_{\text{Job Effect}} + \underbrace{\text{var}\left(\psi_j\right)}_{\text{Firm Effect}} + \underbrace{2\,\text{cov}\left(\theta_i, \psi_j\right)}_{\text{Sorting}} + \text{var}(\epsilon_i) \tag{2}$$

. We denote the variance component due to job characteristics, $\text{var}(\theta_i)$, as the job effect, corresponding to the worker effect in the literature. Consequently, the variance component due to the firm fixed effects, $\text{var}\left(\psi_j\right)$, is the firm effect due to firm wage premium, and the covariance of these two variances, $2\,\text{cov}\left(\bar{\theta}_j, \psi_j\right)$, is the sorting between job quality and firm wage premium. The Equation (2) also allows for further decomposition when θ_i can be modeled as a linear combination of different components. For example, if we assume $\theta_i = \theta_i^A + \theta_i^B = X_i^A\beta^A + X_i^B\beta^B$, we can then rewrite Equation (2) as $\text{var}(\ln w_i) = \text{var}\left(\theta_i^A\right) + \text{var}\left(\theta_i^B\right) + \text{var}\left(\psi_j\right) + 2\,\text{cov}\left(\theta_i^A, \theta_i^B\right) + 2\,\text{cov}\left(\theta_i^A, \psi_j\right) + 2\,\text{cov}\left(\theta_i^B, \psi_j\right) + \text{var}(\epsilon_i)$. We will use this type of decomposition to present our main results in Section 6.

There are several econometric problems that one could face when estimating the Equation (1) and the variance and covariance terms in Equation (2). The first, and perhaps the most important one, is the notorious unobserved worker characteristics that would potentially bias the estimation through its correlation with the observed variables.[19] While the common

[17] Note that in the AKM framework with worker fixed effect, X_i, or in fact X_{it}, often only contains age and/or potential experience because any time-invariant worker characteristics like education (or occupation if there is no occupational switching during the observation period) will be wiped out under the fixed effect estimator. Also note that in our vacancy data the experience variable is the years of experience required by the employer, and thus is presumably occupational experience and more accurate in representing the proficiency of the job skills than the often-used potential experience in the literature.

[18] The employer identifier used here is likely to be coarser than the one used in other studies using administrative or census data. In our data, while in some cases different establishment establishments or subsidies of one firm are identified differently, in other cases they are labeled as the same firm. As a result, some part of the between-establishment wage differential might be identified here as within-firm wage difference if firm have different pay policies across different branches.

[19] For example, we can assume the error term in our main specification has the structure $\epsilon_i = \alpha_i + \varepsilon_i$, where α is unobserved skills and tasks, ε is the real random error. Then, if either $\text{cov}(\alpha_i, X_i) \neq 0$ or $\text{cov}(\alpha_i, \psi_j) \neq 0$, we would not have $E\left(\epsilon_i \mid X_i, \psi_j\right) = 0$ and the estimated β and ψ will be biased. In fact, it is likely that both observed worker characteristics and job characteristics are positively correlated with the unobserved job characteristics, and thus both $\hat{\beta}$ and $\hat{\psi}$ would be overestimated.

approach in the literature is to use panel data and worker fixed effect under the assumption that the unobserved worker characteristics are time-invariant and thus identified, here we suggest that our data provide us an alternative way to resolve the problem by directly observing all the "worker" heterogeneity. The key idea here is to assume that firms post all the information about the job and the ideal worker who they will eventually hire or expect to hire to do the job, and that firms also post the wages based on their valuation of their jobs and the corresponding ideal workers. The information can potentially contain both the usually observed one like education and experience, and the often unobserved one like various skills and tasks as well as job amenities. There are two theoretical perspectives popular in the labor economics that support such an presumption. The first is that an occupation or a job can be considered as a bundle of different tasks and/or skills. And thus the unobserved worker abilities are the unobserved tasks conducted or unobserved skills held by the worker, which are likely to be documented in the job descriptions and requirements. The second theory related is the perspective of directed search. There are large search frictions in the labor market and non-trivial search cost for both firms and workers to match with each other. Given that workers can direct their search—an reasonable argument especially in the case of the online job market, firms have the incentive to document the correct information so that they are more likely to attract and match with their ideal workers.[20] Therefore, as long as we can extract all the information that employers document in their job vacancy texts and use to determine their posted wages, we can control for all the job and (ideal) worker heterogeneity and circumvent the potential omitted-variable bias. Even better, by doing this we are able to open the black box of the unobserved worker characteristics masked by the worker fixed effect and examine what are the most important worker abilities attributable to wage differentials. In fact, we will show in the Section 5 that, those job characteristics extracted by our machine learning algorithms, as we expected, are exactly workers skills, job tasks, and job amenities. Hence, the estimated $\hat{\beta}$ can be seen as the average price of various skills and tasks and other job characteristics in the job market, and if a firm overprices or underprices some skills and tasks evenly for jobs within the firm, such overpricing or underpricing will be accounted as part of the firm wage premiums.

Next, we discuss several other issues that researchers often encounter when dealing with the AKM framework. Because in the AKM framework the firm fixed effects are identified from the job movers, the key assumption required for the identification is thus exogenous mobility, i.e. the job mobilities are uncorrelated with time-varying wage components in the residuals. This rather restrictive assumption has been under debate for a long time given that in many job search model, job switch is an optimal choice of agents and likely to be correlated with the match quality drawn upon matching between worker and firm. However, Card et al. (2013) and the following studies show supportive evidences for this assumption by using an event study to find that there are symmetric wage changes for the movers of inverse directions between any two groups of firms with different pay levels. In our framework with job vacancy data, the firm effect is identified through all the new jobs posted by a firm in the observation period, and if such new job posts accomplish as successful hiring, then similar to the AKM

[20]However, one perspective of the directed search theory that we do not consider seriously here is that in many directed search models, the wage can be also used as a tool to control the length of the queue and the hiring efficiency. To what extent the segregation of the labor market is by the job contents and the wage levels is an open questions. In our framework here, if a firm use such wage premium to attract more job seekers evenly across all its jobs, it will be accounted in the firm fixed effect as the firm wage premium.

framework, we are essentially also using the information of job mobility to identify the firm pay policies. However we suggest that, whether one find the evidence of the event studies in the literature convincing or not, the exogenous mobility is less a concern in our data and our framework because both the job contents or the ideal worker and posted wage are determined before the real firm-worker match. In other words, the job mobilities in our case are designed to be exogenous as long as firms do not foresee certain types of workers that they will match and deliberately hide such information in their job advertisements. Of course there could exist the cases of mismatch where a firm matches and hires a worker who deviates from its ex-ante expectation and may or may not change the wage correspondingly, but as long as such mismatches are rather random and exogenous, our estimations from the job vacancy data can still largely represent the real wage dispersions in the labor market. In fact, such ideal situation of our data and framework helps to relieve us from the consideration of many other mechanisms of wage determination such as bargaining or discrimination, and thus generate a rather clean environment to study the impact of worker heterogeneity and firm heterogeneity in wage dispersion.

Another related issue is the so-called limited mobility bias, which has been thoroughly studied in the literature (Andrews et al., 2008; Kline et al., 2020; Bonhomme et al., 2020). Since the nature of the limited mobility bias is the finite sample bias due to limited job moves to identify the firm fixed effects, our framework would have the similar problem if there are too few job vacancies posted by a firm. However, in our framework this problem is easily resolvable by limiting firms in the sample to have enough number of job posts, which is itself a common data cleaning procedure for job vacancy data to remove unreliable job postings. To show this empirically and practically, we will apply all three correction methods that have been proposed in the literature, namely the homoscedasticity correction in Andrews et al. (2008), the heteroscedasticity correction in Kline et al. (2020), and the clustering method in Bonhomme et al. (2020), and compare their results with the plug-in estimates. Also, because our framework does not trace the origin of the job movers, there is no need to construct connected sample set as the AKM approach, which will cause the drop of a certain amount of samples.

The final issue co-existed in the AKM framework and our framework is the assumption of additive separability in Equation (1), which presumes that the common firm wage effects will apply to all types of workers or jobs. This linearity restriction can be relaxed in two steps. The first step is to allow a firm vary its levels of wage premium among different occupations, as suggested by Torres et al. (2018); Hou and Milsom (2021). We will firstly test this possibility by estimating on both the pooled sample and the sample of individual major-occupation, partially because our job characteristic extracting algorithm will also be tested on these two different sample levels. Then, we will formally test it by using an extensive specification of Equation (1) to estimate our pooled sample, which allows both occupational-specific firm pay policies and occupational-specific skill prices, and comparing the results with our baseline model. Given that our first step of relaxation still largely preserves the additivity between job effect and firm effect, in our second step we will try fully allow nonlinearity between these two terms, i.e. there can be arbitrary complementarity between job effect and firm effect. As such an fully flexible specification is prohibited under a high-dimensional setting, we will utilize a clustering method, which applies the idea behind the approach proposed in Bonhomme et al. (2019) to our case, to easy the computation. We will also show that this clustering approach can also be used to understand the differences of the wage dispersion components across different types

of job vacancies.

5 Use Machine Learning To Understand Vacancy Text

In this section, we apply a series of machine learning algorithms to the job vacancy text data to extract information on detailed job skill requirements and task descriptions. Our first aim here is to select wage-predictive terms from the high-dimensional vacancy text data where both informative and meaningless information about wage determination coexist. In other words, we want the data to tell us what are the useful job characteristics embedded the vacancy text, whether skills, tasks, amenities, or any other terms, that can explain the posted wage variations. We achieve this in Section 5.1 by using regularized linear regression which reduce the effective feature dimension from the whole vocabulary in the data to a few thousand terms. Our second aim is to understand what are the job characteristics that we selected in the last step and, if possible, to classify them into board genres, again through a data-driven perspective. We achieve this in Section 5.2 by using both natural language processing (NLP) algorithm and unsupervised clustering algorithm to conduct feature clustering based on how firms talk about different things in the job vacancy. In this process, we show that our algorithms automatically separate different job skills and tasks, and generate a data-driven skill-task hierarchical structure. Our final aim in this section is to construct low dimensional proxy variables for the useful job characteristics that we have identified and classified in above steps so that we can bring these information back to our wage differential estimation and show how these previously unobserved skills and tasks could improve our understanding on the wage determination and total earning inequality. This further dimensional reduction is achieved by using supervised dimensional reduction algorithm in Section 5.3. Throughout this section, our selections on a variety of machine learning algorithms largely follow the suggestions in Gentzkow et al. (2019), in which the authors review the applications of a wide range of machine learning techniques on text data and economics topics.

5.1 Features Selection

Our first step is to select important features from the raw text of the job descriptions on the job vacancies that incorporate a variety of skills, tasks, non-wage benefits, and perhaps other contents, and by important here we mean holding some predictive power for the posted wage, whether causal or not. Because a feature is often called a "token" in the textual analysis and means either a word or a phrase (or more generally a term), we will use these words interchangeably throughout the paper. To this end, we need first transform the raw job vacancy texts, denoted by $\mathbf{D}$, into a numerical token matrix $\mathbf{C}$ which has dimension $N \times K$. Here N is the number of vacancies in the sample data, and K is the number of tokens in the whole vocabulary set V. V is tokenized from all vacancies' texts of the data after standardization and removing words that do not convey meaningful or interpretable information.[21] Each entry of

[21] For example, all numerical numbers either in Arabic or in Chinese are removed because we have no idea what they interpret without the context. We also remove all the firm name from the vocabulary because it will catch the firm effect and distort the clustering of selected features.

C, indexed by c_{ik}, is an indicator of the presence of token k in vacancy i (1 if present otherwise 0). The details of this transformation are described in Appendix B.1.

Next, we regress the log posted wage on the token matrix **C** to estimate the explanatory power of each tokens in V. Unlike a normal regression problem, the high dimensionality of **C**, in which many dimensions could be totally irrelevant, makes standard techniques like Ordinary Least Square (OLS) infeasible and unsuitable. We thus apply the penalized (also called regularized) linear models to this high-dimensional regression problem for feature selection. The penalization here add additional costs for deviations of any estimators from zero, which helps to shrink the effective dimension of explanatory variables, and the linearity retains the model to be rather intuitive and interpretable. In particular, here we choose the least absolute shrinkage and selection operator (Lasso) regression which extends the Gaussian linear regression and uses a L_1 penalization. The L_1 penalization here means that the penalization cost function is linear (zero curvature and constant shrinkage), and thus the non-differentiable spike of the additional cost at zero leads to sparse estimators, with some coefficients to be exactly zero. This strong form of penalization are particularly suitable for feature selection in text analysis because it limits nonzero estimators for prediction to a rather reasonable size and thus helps to throw out a large amount of potentially uninformative tokens in the raw text. Our lasso estimator is written as

$$\hat{\zeta} = \arg\min_{\zeta} \sum_{i=1}^{N} \left(\ln w_i - \sum_{k=1}^{K} c_{ik}\zeta_k \right)^2 + \lambda \sum_{k=1}^{K} |\zeta_k| \tag{3}$$

, where $\lambda > 0$ is a parameter of the model that indicates the level of the "penalty". Note that the first part within the minimization is the residual sum of squares (RSS) in the normal OLS estimator, which is an unregularized objective proportional to the negative log likelihood, $-\log \mathrm{P}(\ln w_i \mid \mathbf{c}_i)$, and the second part is the penalization term.

A key difference in the estimation of Lasso comparing to the estimation of traditional econometric models like OLS is that there is now a pre-determined hyper-parameter λ, i.e. the parameter is to be set before the estimation through other procedures. This parameter controls the extent to which the model penalize non-zero estimators. The larger the λ the more sparse will be the selected non-zero estimators, and as $\lambda \to 0$ it approaches to the usual maximum likelihood estimation. The standard practice to determine this prior parameter (or "model turning" in the jargon of machine learning field) is to define a criterion to measure the performance of the estimates from different values of λ and then choose the best one from them. Although the commonly used criterions in the machine learning literature is some metrics of the model's out-of-sample prediction power, such approach has been argued to be more suitable for achieving the best predictive performance rather than for selecting features used for further analysis because it often leads to λ too small and overfitting.[22] Instead, we follow the suggestion in

[22]To be specific, this most popular tuning approach that follow the out-of-sample accuracy idea in the machine learning literature is called cross-validation. The basic step is to randomly partition a part of the data as the test sample separated from the training sample that are used to train the model, and then to apply the trained model back to test sample to calculate the results of the pre-defined measure, such as mean squared error. The one repeats this procedure for a large set of parameter values and for different random partitions to get the optimal parameter values. Although the idea of out-of-sample efficiency is exactly designed to target the problem of overfitting, empirical findings in the machine learning literature often suggest that such prediction-based approach is still

Gentzkow et al. (2019) and use the Bayesian information criterion (BIC) as the criterion to choose the optimal λ for our feature selection. Similar to the well-known Akaike's information criterion (AIC), BIC is an approximation to the Bayesian posterior marginal likelihood subject to an adjustment on degrees of freedom. In our Lasso case, the BIC is defined as

$$\text{BIC}(\lambda) = \frac{\|\ln \mathbf{w} - \mathbf{C}\hat{\zeta}_\lambda\|^2}{\sigma^2} + \widehat{df}_\lambda \log N \tag{4}$$

, where σ is the common variance of Gaussian noises, and $\widehat{df}_\lambda$ is the degrees of freedom of the estimation with λ.[23] In practice, we pass a grid of different λ values to the Lasso regression and find the λ^* that yields the lowest BIC score.

Table 1: BIC Tuned Lasso Models

	Pooled	Computer	Design_ Media	Admin
λ^*	332.0	190.3	238.5	155.0
MSE	.162	.149	.142	.100
R^2	.566	.494	.461	.418
BIC/N	.446	.527	.561	.613
df	3,144	1,922	929	691
K	109,123	51,602	39,306	24,896
N	3,999,005	1,330,001	561,236	277,932

Notes. For each major occupation, the hyperparameter λ^* of the Lasso model is tuned by minimizing BIC(λ) as defined in the text. The smaller the λ, the less the penalization for nonzero coefficients and the more features are picked by the Lasso model.

The tuned Lasso models and their estimation results for both Pooled sample and three selected occupations are shown in Table 1. The tuned λ^* ranges from 155 to 332 in different samples due to the different levels of tradeoff between decreased normalized RSS and increased penalty from the increased degree of freedom with a higher $lambda$. As a result, the number of tokens with nonzero estimated coefficients also varies across samples, ranging from over 3100 tokens in the Pooled sample to less than 700 tokens in Administrative occupation, a substantial reduction from the norm of the original vocabulary K. The R-squared values

very likely to overweight the predictive power to model rigidity and interpretability in many real world settings. Also note that although we also partition samples in our occupation classification approach, we do not need to conduct such model tuning because the Naive-Bayes classifier is too simple and requires no hyper-parameters.

[23] More generally the BIC is defined as $\text{BIC} = -2\log(\hat{L}) + \log(N)\widehat{df}$, where $\hat{L}$ is the maximum likelihood under estimation. For a linear Gaussian model, the maximum log likelihood can be derived as: $\log(\hat{L}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\ln\left(\sigma^2\right) - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{2\sigma^2}$, where y and $\hat{y}$ are any true and predicted targets, and σ is the "true" error variance. By bringing this term back to the definition and removing the constant terms one obtains the formula in the text. Note that there is no general way to estimate σ^2, which works as a baseline unit for the RSS so that there is a "fair" comparison between the reduction in estimation error and the increase in number of parameters. In practice, we simply estimate the σ to be Var($\ln \mathbf{w}$).

for the estimated models are between 57% (Pooled) to 42% (Admin), indicating that the job characteristics extracted by our Lasso model can account for around half of the entire wage variance.

One caution that has been repeatedly raised from the literature is that the selected features and their coefficients of any high-dimensional penalized models are generally uninterpretable (see for example Belloni et al., 2014; Mullainathan and Spiess, 2017). In general the results in these models and thus in our Lasso model suffer two problems for further interpretation: multicollinearity and flexibility.[24] The first is that given the high-dimensionality and the penalization, and especially in our case with a linear regularization, the penalized regressions will likely to pick one feature at random from a highly correlated group. In other words, multicollinearity among features in such high-dimensional penalized model could cause the set of nonzero variables selected to be highly unstable. As a result, the tokens selected by our Lasso model in general do not necessarily indicate any casual relationships. The second problem is on the interpretation of the coefficient levels. Even after regularization, our models still left hundreds or thousands nonzero control variables, which makes the both the levels and the signs of the coefficients hard to be taken for any serious interpretation.[25]

Because we do want to go beyond wage prediction and to learn something about the contents of various job characteristics and their impact on wage determination from the selected features, we now check how severe are these problems of statistical uncertainty in our Lasso estimation results. To this end, we use subsampling, which is a nonparametric approach of inference and has been argued can retain robust in the cases where the estimator has non-differentiable loss function and potential model selection.[26] In practice, we randomly partition our sample into ten pieces and re-estimate the Lasso model equipped with previously tuned λ^* separately on each subsample. We repeat this procedure for ten times and gather all the results to calculate the standard deviation of our parameters of interest—the coefficients of the nonzero features selected in our Lasso estimation with the full samples.[27] The result of the

[24] On the top of these two, there could still be unobserved bias in that our captured job characteristics predict the wage not because they have direct casualty links but because they link to some other unobserved casual factors or to the firm effect. For the first possibility, we suggest that this will affect our main results because our models have extracted a fairly large amount of skills, tasks, amenities and other potential job characteristics, and eventually we will cluster these job characteristics based on their textual relationship in the text. For the second possibility, as we explained before, in our vectorization process we have already removed the terms of firm names from the vocabulary and perhaps more directly, in our selected tokens by the Lasso model we don't find specific terms are that can be used to identify any particular firms. However, if different types of firms with different levels of firm wage premiums also post certain job characteristics, being either skills, tasks, or non-wage compensations, then some features selected here will hold strong prediction power on the posted wage simply because of this indirect relationship. We show later that this hypothesis is actually true in our data.

[25] One simple but intuitive example of this problem is that if one put both age and experience variables into a high dimensional wage regression along with one hundred of other individual variables, there are chances that the coefficient of one of the age and experience variables might turn out to be negative. Although this result can still tell something informative but one need to fully acknowledge what has been conditioned on to give a reasonable interpretation.

[26] Within the inference computation algorithms that approximate the sampling distribution, the commonly-used nonparametric bootstrap uses with-replacement resampling and thus fails for the statistics models that involve non-differentiable loss functions like Lasso here. In comparison, in subsampling each subsample is a draw from the true data generating process, and thus it works for estimation algorithms even with non-differentiable losses. For more details about subsampling, we refer to Gentzkow et al. (2019) and Gentzkow et al. (2019).

[27] To calculate the statistics of interest, one needs to translate the uncertainty in the subsamples to the one in the

subsampling is shown in Figure E1, from which we can see that the coefficients of the tokens selected in our full-sample Lasso estimation are generally robust—they don't easily flip the sign in different subsamples and their standard deviations are actually quite small in most cases. Although the robustness of our Lasso results shown in this uncertainty check does not necessarily dispel all the potential problems due to the multicollinearity and flexibility in our high-dimensional regularized feature selection, and hence any causal inference for the estimates is still largely forbidden, we think that at least it gives us some confidence that our selected features are likely to represent some stable and consistent patterns pertain to the posted wage determination.

Acknowledging the potential interpretability problems of the Lasso results, we then inspect the most important features estimated by the model to see if they make some intuitive sense and if they expose some stylized features that desire more validation and reasoning. In Table 2 and Table 3 we show the top positive and negative tokens that have the largest absolute coefficient and occurs in more than one percent of the vacancies.[28] We next document several patterns found from these top tokens and discuss their potential implications. First, there is a few compensation terms appear in the top tokens. The potential reason that these terms are able to predict posted wage can be compensating differential or some correlation with the job or firm qualities. Given the main focus of this paper is about worker heterogeneity in skills and tasks and firm heterogeneity in wage policies, we leave the careful analysis of these non-wage compensations in our companion paper Zhu (2022). Second, perhaps not surprisingly, education terms appear to be the highly predictive features in our Lasso estimation. And the sign of these coefficients make intuitive sense: bachelor degree and master degree in the top positive tokens and vocational college degree in the top negative tokens. Also, several terms related to fresh graduates appear to be negatively correlated with the posted wage, which may represent the effect of working experience.[29] These intuitive features thus provide some additional sanity checks verifying that our Lasso models do find the key features that are important for posted wage determination in our data.

The third stylized fact is that there are many occupation-specific and professional terms in the top tokens. In top positive tokens we observe for example "deep learning", "golang", and "c++" in Computer occupation, "engine", "3d", and "journalist" in Design & Media occupation, and "translation" and "business negotiation" in Admin occupation. And in top negative tokens we can observe for example "installation" and "computer" in Computer occupation, "photoshop"

full sample because each subsample will be smaller than the entire sample of interest. We follow the convention to assume the estimator's rate of convergence to be $\sqrt{n}$ so that the corrected standard deviation of the coefficients can be calculated as $\text{sd}(\hat{\zeta})\sqrt{\frac{B}{N}}$, where B is the sample size of each subsample. The robustness for our estimation results would not change significantly even if we choose a higher rate of converge.

[28] There are other ways to define the importance of coefficients of Lasso estimation, for example the absolute coefficient values scaled by the associated standard deviation or the order in which the coefficient of a covariate first turns to nonzero in a series of Lasso estimations with decreasing penalties. Here by simply showing the top positive and negative tokens with larger than 1 percent occurrence we aim to display the tokens with the highest prediction power that are not rare. We show the full list of all nonzero tokens selected by the Lasso estimations on the author's personal website.

[29] While we have the education terms in our features, we don't have any direct work experience terms in our vocabulary V because the working experience is often documented in the vacancy as "required n year experience" where n is pure number and thus dropped from the vocabulary because they are also used in many other cases and hard to interpret.

Table 2: Top Positive Tokens (Frequency > 1%) in Lasso Regression

	Pooled			Computer			Design_Media			Admin		
	token	coef	freq	token	coef	freq	token	coef	freq	token	coef	freq
1	14薪(14th month pay)	.152	.014	15薪(15th month pay)	.181	.010	14薪(14th month pay)	.193	.011	大学本科(undergraduate)	.161	.014
2	三餐(three meals)	.143	.014	三餐(three meals)	.148	.014	带领(lead)	.155	.025	本科(undergraduate)	.157	.156
3	大平台(large platform)	.131	.019	14薪(14th month pay)	.140	.017	三餐(three meals)	.129	.015	总裁(president)	.120	.014
4	硕士(master degree)	.126	.015	硕士(master degree)	.109	.027	c++(c++)	.121	.017	ceo(ceo)	.117	.010
5	带领(lead)	.107	.041	带领(lead)	.089	.038	危机(crisis)	.113	.011	搭建(build)	.117	.016
6	c++(c++)	.092	.051	golang(golang)	.080	.017	游戏(games)	.098	.180	带领(lead)	.105	.017
7	算法(algorithm)	.082	.061	大牛(guru)	.079	.047	欧美(europe & america)	.090	.011	政府(government)	.103	.030
8	大牛(guru)	.082	.028	深度学习(deep learning)	.078	.022	引擎(engine)	.090	.046	高薪(high salary)	.089	.018
9	知名(famous)	.079	.019	知名(famous)	.070	.014	4a(4a)	.090	.014	翻译(translation)	.083	.012
10	机器学习(machine learning)	.077	.016	高薪(high salary)	.070	.018	六险一金(six insurance & one fund)	.086	.046	本科学历(bachelor degree)	.082	.018
11	组建(formation)	.076	.013	牛人(maestro)	.068	.012	财经(finance)	.084	.016	战略(strategy)	.077	.015
12	本科(undergraduate)	.074	.319	海外(overseas)	.067	.010	本科(undergraduate)	.078	.238	大型(large scale)	.076	.030
13	海外(overseas)	.072	.026	go(go)	.065	.027	上市公司(listed company)	.076	.021	落地(landing)	.070	.018
14	react(react)	.072	.020	c++(c++)	.064	.144	金融(finance)	.076	.031	项目管理(project management)	.067	.011
15	开发(development)	.071	.374	算法(algorithm)	.064	.164	外包(outsourcing)	.074	.012	海外(overseas)	.066	.021
16	大学本科(undergraduate)	.066	.029	react(react)	.064	.061	大牛(guru)	.070	.022	背景(background)	.064	.032
17	高薪(high salary)	.063	.028	机器学习(machine learning)	.061	.045	海外(overseas)	.068	.024	制定(develop)	.063	.097
18	落地(landing)	.060	.067	落地(landing)	.061	.037	记者(journalists)	.068	.011	13薪(13th month pay)	.063	.019
19	战略(strategy)	.057	.047	开发(development)	.059	.776	13薪(13th month pay)	.068	.023	统招(unified recruitment)	.058	.031
20	直播(live streaming)	.056	.014	音视频(audio & video)	.058	.012	c4d(c4d)	.066	.021	预算(budget)	.057	.021
21	上市公司(listed company)	.055	.027	统招(unified recruitment)	.054	.044	知名(famous)	.065	.023	重大(major)	.055	.019
22	大型(large scale)	.055	.072	北京(beijing)	.053	.012	unity(unity)	.065	.043	装修(decoration)	.055	.016
23	职责(responsibilities)	.055	.048	直播(live streaming)	.052	.011	高薪(high salary)	.064	.016	资源(resources)	.053	.043
24	班车(shuttle)	.054	.018	推荐(recommend)	.052	.023	管理工作(management)	.063	.010	推动(promote)	.051	.029
25	金融(finance)	.054	.070	管理工作(management)	.051	.016	3d(3d)	.063	.106	金融(finance)	.051	.036
26	六险一金(six insurance & one fund)	.053	.055	ai(ai)	.051	.015	大型(large scale)	.063	.043	英语(english)	.050	.054
27	python(python)	.052	.066	股票(stock)	.049	.025	性能(performance)	.063	.016	商务谈判(business negotiations)	.048	.010
28	总监(director)	.052	.022	本科(undergraduate)	.048	.365	统招(unified recruitment)	.059	.019	优化(optimization)	.046	.079
29	统招(unified recruitment)	.051	.042	薪资(salary)	.048	.049	大学本科(undergraduate)	.059	.023	职责(responsibilities)	.046	.035
30	hive(hive)	.051	.013	补充(supplementary)	.045	.019	ip(ip)	.057	.017	统筹(integrated planning)	.046	.028
31	技术(technology)	.049	.285	金融(finance)	.045	.057	指导(guidance)	.054	.047	上市公司(listed company)	.045	.020
32	引擎(engine)	.049	.017	建设(construction)	.045	.078	设计(design)	.054	.546	出差(business trip)	.045	.038
33	团队(team)	.048	.552	高级(advanced)	.045	.022	职责(responsibilities)	.054	.043	集团(group)	.044	.018
34	期权(options)	.047	.052	大型(large scale)	.043	.113	主导(leading)	.052	.025	指标(indicators)	.043	.033
35	收入(revenue)	.047	.019	六险一金(six insurance & one fund)	.041	.057	动效(dynamic effects)	.050	.016	整体(overall)	.042	.023
36	集团(group)	.046	.022	职责(responsibilities)	.041	.049	数值(numerical value)	.050	.012	规划(planning)	.042	.036
37	生态(ecology)	.045	.012	期权(options)	.041	.062	作品集(portfolio)	.049	.021	转化(transformation)	.042	.011
38	主导(leading)	.045	.025	指导(guidance)	.040	.076	角色(roles)	.049	.053	梳理(combing)	.041	.016
39	增长(growth)	.044	.021	架构设计(architecture design)	.040	.133	落地(landing)	.049	.041	公关(public relations)	.040	.021
40	股票(stock)	.044	.022	广告(advertisement)	.040	.015	产出(outputs)	.048	.033	管理工作(management)	.039	.110

Notes. These are the tokens selected our Lasso models that have highest (or lowest) coefficients and occurs in more than 1 percent of the sample vacancies. Although the Lasso coefficients of our model means the percentage rise of the expected wage for the occurrence of the certain word in the vacancy text, these coefficients generally do not indicate any casual relationship due to the strong multicollinearity among features and the flexible structure of the Lasso model (see our discussion in the main text). The more recommended way of interpretation is that these features hold strong prediction power for the posted wage and potentially are or are correlated with some important factors of wage determination. On the author's personal website we list all the nonzero features selected by Lasso for reference.

Table 3: Top Negative Tokens (Frequency > 1%) in Lasso Regression

	Pooled			Computer			Design_Media			Admin		
	token	coeff	feq	token	coeff	feq	token	coeff	feq	token	coeff	feq
1	应届生(freshmen)	-.155	.018	毕业生(graduates)	-.205	.013	应届生(freshmen)	-.188	.017	五险(five insurance)	-.070	.052
2	五险(five insurance)	-.136	.030	五险(five insurance)	-.197	.016	实习(internship)	-.133	.011	毕业生(graduates)	-.061	.082
3	毕业生(graduates)	-.128	.033	大专(vocational college)	-.134	.072	五险(five insurance)	-.132	.033	中专(vocational school)	-.059	.038
4	专科(vocational major)	-.100	.036	社保(social insurance)	-.121	.012	毕业生(graduates)	-.132	.030	应届生(freshmen)	-.057	.048
5	双休(two-day weekend)	-.098	.166	专科(vocational major)	-.119	.030	双休(two-day weekend)	-.090	.176	实习(internship)	-.056	.012
6	大专(vocational college)	-.094	.148	双休(two-day weekend)	-.115	.147	应届(recent graduate)	-.072	.026	实习生(interns)	-.053	.017
7	助理(assistant)	-.079	.011	应届(recent graduate)	-.106	.011	大专(vocational college)	-.070	.144	双休(two-day weekend)	-.051	.214
8	客服(customer service)	-.075	.030	测试用例(test cases)	-.067	.068	社保(social insurance)	-.068	.023	玩家(player)	-.046	.024
9	社保(social insurance)	-.073	.028	安装(installation)	-.067	.048	专科(vocational major)	-.066	.041	普通话(mandarin)	-.046	.172
10	会计(accounting)	-.071	.019	th(th)	-.066	.014	有限公司(ltd.)	-.059	.012	女性(women)	-.038	.015
11	住宿(accommodation)	-.067	.016	电脑(computer)	-.065	.011	专业不限(any major)	-.055	.011	社保(social insurance)	-.037	.060
12	行政(administration)	-.067	.027	售后(after sales)	-.061	.011	人性化(humanization)	-.055	.019	qq(qq)	-.037	.036
13	专员(commissioner)	-.063	.011	年轻(young)	-.060	.013	漫画(comics)	-.053	.014	轻松(easy)	-.035	.043
14	淘宝(taobao)	-.059	.015	五险一金(five insurance & one fund)	-.059	.273	cad(cad)	-.052	.010	网站(website)	-.033	.032
15	协助(assistance)	-.058	.164	出差(business trip)	-.051	.030	photoshop(photoshop)	-.049	.235	清洁(cleaning)	-.030	.015
16	ps(ps)	-.056	.029	记录(records)	-.048	.015	cdr(cdr)	-.047	.012	卫生(health)	-.029	.024
17	有限公司(ltd.)	-.056	.012	吃苦耐劳(hardworking)	-.048	.015	网站(website)	-.047	.180	文员(clerks)	-.029	.014
18	安装(installation)	-.055	.020	节日(holidays)	-.046	.059	协助(assistance)	-.046	.131	考勤(attendance)	-.029	.104
19	photoshop(photoshop)	-.052	.039	客户(clients)	-.046	.078	ps(ps)	-.045	.142	电子商务(e-commerce)	-.029	.031
20	细心(careful)	-.050	.032	轻松(easy)	-.043	.017	吃苦耐劳(hardworking)	-.044	.023	录入(input)	-.028	.044
21	吃苦耐劳(hardworking)	-.050	.032	软件测试(software testing)	-.043	.047	动漫(anime)	-.044	.019	轮班(shift)	-.028	.013
22	核对(verification)	-.048	.011	微信(wechat)	-.041	.042	轻松(easy)	-.044	.033	接听(answer the phone)	-.027	.101
23	人力资源(human resources)	-.047	.032	.net(.net)	-.041	.034	接触(contact)	-.042	.011	行政(administration)	-.027	.256
24	网站(website)	-.047	.090	耐心(patience)	-.040	.023	编辑(editor)	-.039	.204	全勤奖(perfect attendance award)	-.026	.032
25	专业不限(any major)	-.047	.020	网站(website)	-.039	.101	美工(artwork)	-.038	.032	应聘(apply for the job)	-.025	.018
26	人性化(humanization)	-.046	.012	专注(focused)	-.038	.011	论坛(forum)	-.038	.034	移动(mobile)	-.025	.013
27	excel(excel)	-.046	.047	网络设备(network equipment)	-.037	.016	淘宝(taobao)	-.038	.024	吃苦耐劳(hardworking)	-.025	.055
28	普通话(mandarin)	-.045	.027	bug(bug)	-.036	.053	年轻(young)	-.038	.034	加入(join)	-.024	.041
29	交代(explanation)	-.044	.013	作品(works)	-.035	.023	提成(commission)	-.037	.017	游戏(games)	-.024	.039
30	年轻(young)	-.044	.025	节假日(holiday)	-.034	.037	客户(clients)	-.037	.096	前台(front desk)	-.023	.088
31	接触(contact)	-.044	.010	分红(dividend)	-.034	.012	微信(wechat)	-.037	.172	部门经理(department manager)	-.023	.014
32	轻松(easy)	-.043	.027	故障(failure)	-.033	.055	玩家(player)	-.037	.017	资料(information)	-.023	.122
33	致力于(commitment)	-.043	.014	自主(autonomy)	-.033	.014	coreldraw(coreldraw)	-.037	.041	倒班(shift)	-.023	.015
34	应届(recent graduate)	-.043	.029	双薪(double pay)	-.033	.035	上级(higher)	-.036	.034	淘宝(taobao)	-.022	.047
35	五险一金(five insurance & one fund)	-.043	.294	培训(training)	-.033	.076	上传(upload)	-.036	.014	广阔(wide)	-.022	.024
36	编辑(editor)	-.042	.042	ssh(ssh)	-.033	.010	细心(careful)	-.033	.028	服从(obedience)	-.022	.029
37	招聘(recruitment)	-.041	.057	xcode(xcode)	-.033	.016	加入(join)	-.033	.048	客户档案(customer profile)	-.022	.016
38	seo(seo)	-.041	.010	细心(careful)	-.032	.015	耐心(patience)	-.031	.036	社会保险(social insurance)	-.022	.015
39	成立(established)	-.041	.011	专业优先(professional priority)	-.032	.024	节日(holidays)	-.031	.084	档案(archives)	-.022	.046
40	电脑(computer)	-.039	.014	测试报告(test report)	-.032	.037	文字(text)	-.031	.229	地点(location)	-.022	.045

Notes. See the note in Table 2

and "editor" Design & Media occupation, and "mandarin" and "answer the phone" in Admin occupation. Again the signs of these features take intuitive sense in that within the occupations, those of positive tokens are high level skills and those of negative tokens are relatively low level skills. This thus confirms our prior that firms describe the detailed skills and tasks that they demand in their job posts and these terms are important for the posted wage determination. This result is also consistent with the perspective of multi-dimensional skills and tasks, as we have argued earlier, in which occupations are different compositions of various skills and tasks and there could also have important within-occupation skill and task variations.

Finally, in additional to the occupation-specific skill or task terms, we also observe two groups of terms in the top tokens that consistently appears across different major occupations. The first group is a set of terms related to management and within-firm hierarchy. In the top positive tokens we observe terms like "lead", "management", and "team" across all samples. Whereas in the top negative tokens we observe terms like "assistance" and "supervisor" that represent the position of the job within the firm.[30] The second group is a set of non-cognitive human capital terms like "hardworking", "careful", "patience", or "focused" in top negative tokens. These are general human capital that should be valued in any jobs, and one possible explanation for their negative relationship with posted wage can be that for some reasons they are more likely to be required in low-skill jobs or by firms with low wage premiums but less likely to be mentioned in high-skilled jobs or by firms with high wage premiums.

To sum up, our Lasso models select the most predictive features for the posted wage from the vacancy text data. In general, these features are constituted by various skills and tasks, along with some terms about compensations or amenities that we will exclude from our analysis. Although each coefficient is not interpretable due to multicollinearity and flexible structure embedded in our high-dimensional and penalized model, we do observe intuitive and interesting patterns within these selected features. Our next step is to classify these features into different types so that we can not only understand the underlying structure of these job characteristics but also together them into different bundles to study different questions about wage inequality.

5.2 Features Clustering

The penalized linear model in the last subsection reduces features of interest to less than 3 percent of the entire token vocabulary. However, the number of remaining tokens is still large, and it is thus hard to get a general picture about what are these features and what relationships do they hold. Although one can simply look at those selected nonzero tokens and decide for each what type it is based on some prior knowledge, here we will show how we can achieve this in a less arbitrary way by using natural language processing (NLP) model to learn the

[30] Here we check the raw data and find the term "supervisor" does not necessarily mean a supervisor job in many cases but mainly occurs in sentences like "follow the order of supervisor". This case is a good example showing the tricky part of textual analysis: all tokens should be interpreted by its meaning in the context rather than taking its superficial meanings. Another example is the positive feature "subcontracting" in Design & Media occupation. We find that it often occurs in sentence like "assigning, supervising and checking the subcontracted works" and thus means the tasks and skills of managing subcontracted workers rather than doing subcontract works.

associations between terms in our vacancy text and then using unsupervised machine learning algorithm to search for potential clusters and patterns within our selected tokens.

In the last subsection we have represented our job text documents through the presence indicator matrix **C**. Though simple and useful in many cases, this matrix does not tell us anything about the relationships between the tokens.[31] This motivates the development of the word embedding models in NLP, which go beyond simple counts of individual words or phrases and learn from the rich syntactical structures embedded within the human-written text to understand the "meanings" of the words. In particular what these models do is to map each word to a latent vector space in $\mathbb{R}^H$ where the dimensions of this latent vector space H correspond to some hidden aspects of meaning of which different words or phrases will hold as the endowment to fulfill their content, and where the relationships between words can be represented through some internally consistent arithmetic calculations. Among many methods to generate this mapping, we will use the most basic neural network method, the Word2Vec model, for our vacancy text.[32] The key idea of the Word2Vec model is that words in similar contexts, represented by the words with close sets of adjacent words, share the similar semantic meanings in the vocabulary, and vice versa. Consequently, we can obtain such relationships by training a neural network with a single hidden layer to perform either a task of given an input word, predicting the probability distribution of the nearby words, or the mirror task of given inputs of context words, predicting the center word. The projection weights that turn the input word or context words to the hidden layer are then interpreted as the word embeddings.[33] In practice, we use the version of the Word2Vec model which predict the center word given the surrounding context words, which is also called continuous bag-of-words (CBOW) because the order of context words does not influence prediction (bag-of-words assumption).[34] The details

[31]For example, consider two terms have proximate meanings. They can simultaneously occur in the same vacancy if this meaning is rephrased several times in the vacancy text. But they might also never simultaneously occur in the same vacancy if only either word would be used even though they mean very similar things. Therefore, simply through the vectors of presence or counts we cannot have enough information to tell any two words are proximate or distant.

[32]Our choice is based on a suitability and performance combined consideration. Although models with deeper neural networks like Bidirectional Encoder Representations from Transformers (BERT) are more powerful, the training of such models is significantly computation-demanding and time-consuming, making many researchers directly use already trained models based on internet text contents like Wikipedia or web news. One major strength of such more sophisticated models is that they can learn the different meanings of one token in different contexts (a trouble feature of the human nature language), while the Word2Vec model can assign only one context meaning for one token. However, given that vacancy text data is a very specific environment for language usage, such compound issue would less likely to happen in our case. More importantly, many words about job characteristics might have specific meanings deviated from the one for normal usage, and many specific terms could not exist in the vocabulary of any pre-trained models at all. Therefore, it is important to directly train any work embedding models on our specific job vacancy text data to get the best results.

[33]Note that the task here is often called synthetic or auxiliary task because we are not actually going to use that neural network for the task we trained it on—the problem of predicting surround words or center word. Rather our aim is just to learn and obtain the weights of the hidden layer. Therefore, although the Word2Vec model itself is an unsupervised machine learning task—unsupervised extraction of semantics for words from the corpus, the way it is phrased is using an auxiliary supervised machine learning task to learn the embeddings as useful representations of the words.

[34]The another version of the Word2Vec model that predict the adjacent words given a single word is called skip-gram. This architecture weighs nearby context words more heavily than more distant context words and performs well in the cases of infrequent words. We choose the CBOW architecture mainly because its generic

of the CBOW word embedding algorithm is described in Appendix B.2.

The result of our word embedding model is a $K \times H$ embedding weight matrix $\mathbf{U}$, where each row of the matrix, $\mathbf{u}_k$, is the representation vector of the word or phrase k in the latent embedding space. Note that although we will only use the embedding vectors of those nonzero tokens that are selected by our Lasso estimation in Section 5.1, i.e. $\mathbf{U}' \equiv \{\mathbf{u}_k\}$ where $k \in V'$ and $V' \subset V$ is the set of selected features, each of these embedding vectors is jointed estimated with and thus defined by all the words in the entire vocabulary V. With these embedding vectors in hand, we now can apply unsupervised clustering algorithm to classify our nonzero tokens into different clusters based on their meanings in the text. Here we also use a simple and popular method, K-Means, which find the centroids for the clusters in the target space (here the embedding space) to minimize the sum of within-cluster Euclidean distances. To conduct K-Means clustering, we first need to decide a primary parameter, the number of the clusters, denoted as P. Then we look for the P-partition of the selected vocabulary V', $\{V'_1, V'_2, \ldots, , V'_P\}$, to minimize the distance from each token to the centroid of the cluster it belongs to:

$$\underset{\{V'_1,V'_2,\ldots,V'_P\}}{\arg\min} \sum_{p=1}^{P} \sum_{k \in V'_p} \left\| \mathbf{u}_k - \frac{1}{\left|V'_p\right|} \sum_{j \in V'_p} \mathbf{u}_j \right\|^2 \tag{5}$$

. The pre-determined parameter P is the only hyper parameter of the algorithm and is arbitrary unless we know the number of the "true" clusters of the data, which often does not even exist.[35] In practice, we select $P = 8$, i.e. eight clusters for each samples of analysis, in order to avoid some obvious entanglements, but our main findings hold for selecting other close numbers. To visualize the clustering result, we use t-distributed stochastic neighbor embedding (t-SNE) algorithm to first reduce the embedding matrix $\mathbf{U}$ to a two-dimensional representation and then plot all tokens in V' on this reduced two dimensions with their assigned clusters labeled in different colors. We show this for the Pooled sample and for Computer occupation in Figure 2, and same plots for other occupations can be seen in Figure E2.

We then document several consistent patterns that we find in the results of the K-Means clustering across different samples. Firstly, in each sample we can find a cluster that contains all the compensation words and phrases given that they have a rather special context in the job vacancy text. This gathered cluster, labeled as V'_1, thus helps us to remove all the job characteristics that could be potentially affect posted wages through the channels that we are not interested here. Secondly, we can observe a similar cluster cross all major occupations that contains a combination of words about cognitive skills, noncognitive skills, and interpersonal skills. These words include "hardworking", "patient", "responsible", "challenging", "logic", "critical thinking", "self-learning", "problem-solving", "open mind", "communication", etc. Some of these words have been used in the prior studies (e.g. Deming and Kahn (2018)) to measure the level of cognitive and social skills and found important in determining job wage. It might be a little surprising that although the terms in this cluster have slightly different stress between

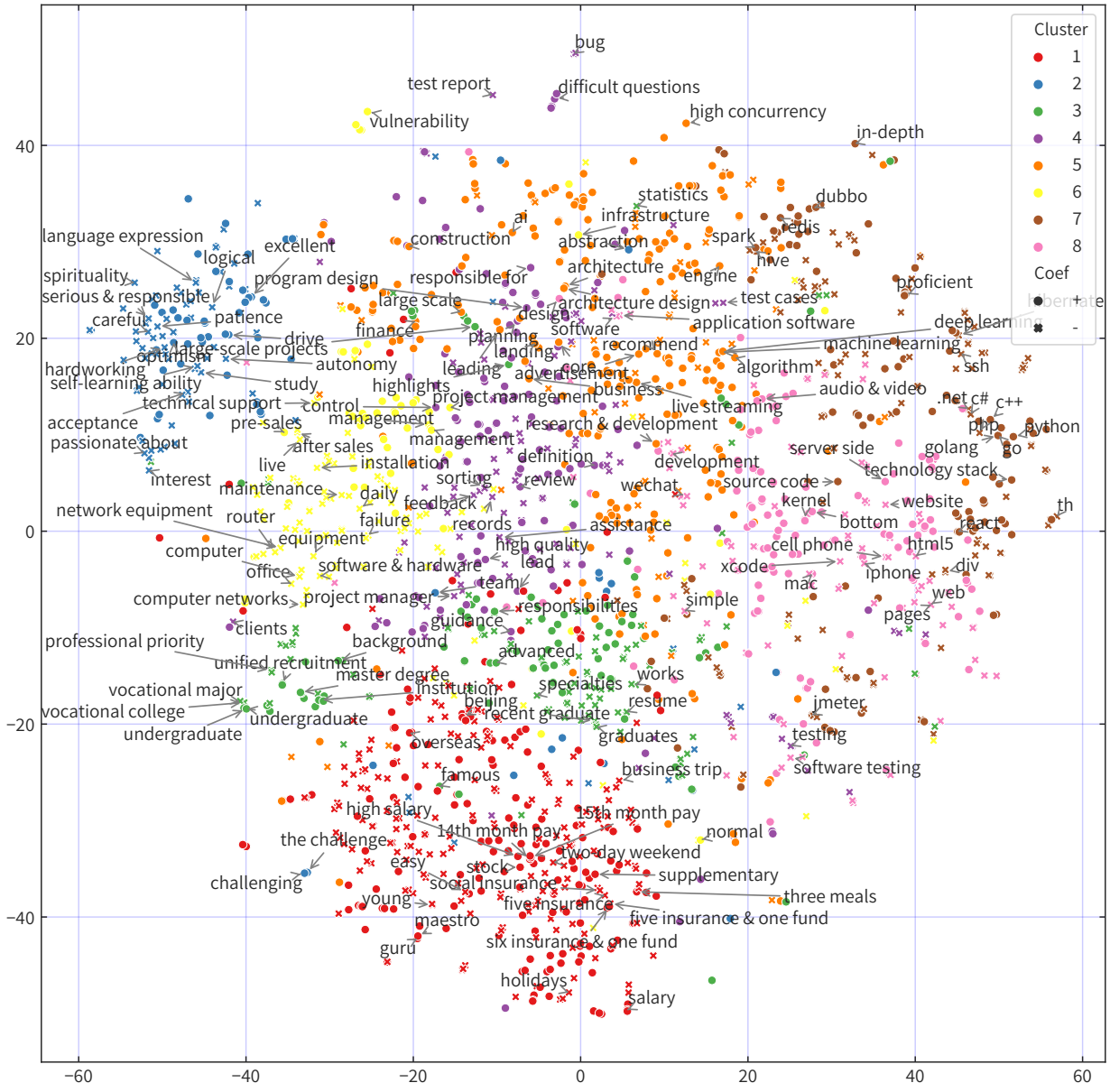
model is more simple and nature, and its algorithm is faster.

[35] This procedure is analogous to decide the hierarchy of the occupation categories by human knowledge. Both board categories and granular categories make some sense for understanding the structure of the occupational space and there is no one particular ideal number of the categories of occupations.

Figure 1: Feature Clustering on the Embedding Space

(a) Pooled Sample

(b) Computation Occupation



Notes. The clustering of tokens is created by applying K-means algorithm to the latent dimensions of token context obtained from the word embedding model. Here, we visualize all the nonzero features on the 2d dimension by using TSNE algorithm to reduce the latent dimensions of word vectors from one hundred to two. This further dimensional reduction is only for illustrating the results of the clustering and is not used in any results of the main text. The features with text labels are the top 20 words in each cluster with the highest absolute value of coefficients and with frequency larger than 1%. The same plots for other occupations are shown in Figure E2.

different occupations, in general the composition of the words are similar across occupations, indicating that these skills are fairly general and firms of all occupations require a similar set of these general skills whether cognitive, noncognitive or interpersonal. We index this cluster as V_2'.

Thirdly, we can find a cluster that contains the education related tokens in all occupations. It incorporates tokens about the general education levels, like high school, vocational college, college, new graduates, etc., and tokens on more specific education requirements like college majors and professional certificates. It also includes requirements on experience in certain fields and the most fundamental skills or tasks in the board occupations, probably because firms often write these terms in together with the education requirements. Therefore, this cluster can be seen as an extensive education control, which indicates relatively more specific skills and tasks than the ones in V_2', and we index it as V_3'.

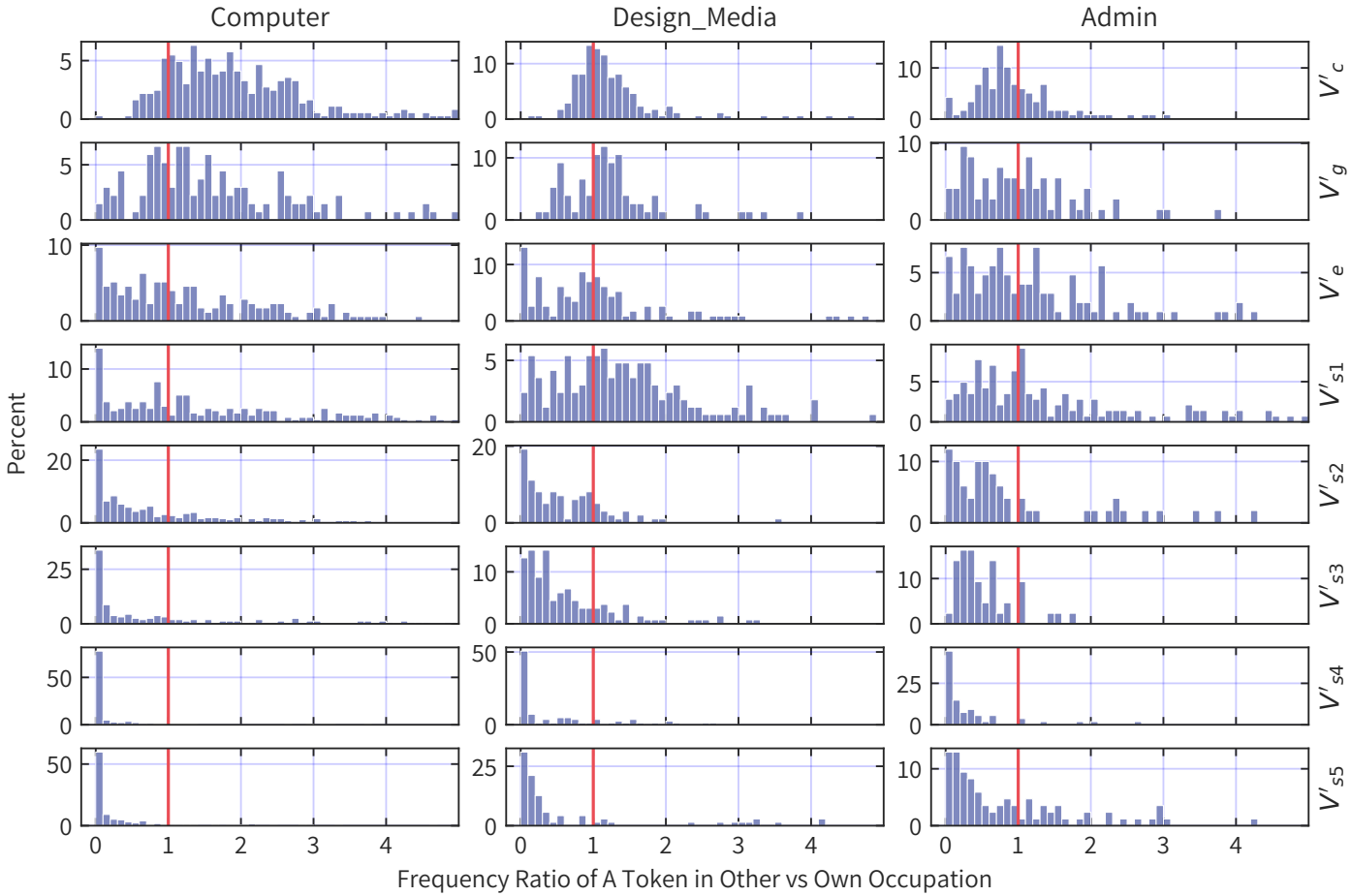
The forth cluster that we are able to identify from our clustering result is less consistent and more ambiguous comparing to above three clusters. To be specific, in each major occupations we can find a cluster that incorporates words or phrases related to within-firm hierarchy and coordination like management, planning, allocation, collecting, subordination, and assistance. However, for each major occupation this cluster also incorporates occupational-specific tasks that are linked with these hierarchical or organizational terms, and for the Pooled sample it includes a variety of administrative tasks. This cluster is gathered together by the algorithm likely due to the fact that whatever the occupation is, there are always similarly stated tasks about (manager) assigning tasks, (subordinate) following manager's order to accomplish tasks, and coordinating different tasks for departments within the firm or between firm and outsider clients or suppliers, although these tasks can be specific to different occupations. We consider this cluster as an extensive or complementary control for (potentially occupational) experience, which largely indicates the job position in the firm hierarchy or the job ladder, and index it as V_4'.

For the rest of the clusters, it becomes difficult to find the similar counterparts across different samples and occupations. In particular, for the Pooled sample, the rest four clusters (and also the V_4' to some extent) seem to be the clusters of skills and tasks stemmed from different major occupations. And for single occupation, the rest four clusters seem to be further partition of the skills and tasks in that major occupation into distinguished groups.[36] In other words it seems that our clustering algorithm conducted on the word embedding space of vacancy text mimics what the official occupation categories do: classifying jobs into different hierarchies based on skills and tasks. As a result, we recognize these rest clusters as occupation-specific skills and tasks and label them $V_5', \ldots, V_8'$ arbitrarily.

Our definition above on different clusters derived from the algorithm is based on our human learning on the terms in those clusters and one may doubt that to what extent do they make sense. To confirm, we measure the specificity of a token $k \in V'^o$ selected in occupation o by compare its occurrence rate in V'^o with the weighted mean of its occurrence rate in $V'^{o'}, \forall o' \neq o$. We plot the distributions of this other-vs-own occupation frequency ratio for all

[36] This is easiest to see in Computer occupation, where these clusters contain many terms about programming languages and other IT-specific technical words. Whereas in other occupations, the skills and tasks might not be completely specific. For example, analysis and planning could be important for many occupations although for different occupations the content for analysis or planning might be very different.

Figure 2: The Distribution of The Ratio of Feature Frequency in Other Occupations to in Own Occupation



Notes. We calculate this ratio by dividing each token's occurrence percentage in the vacancies out of the major occupation of this token by its occurrence percentage in this own major occupation. The cluster index is the same as the one in . In particular, the cluster 1 is the compensation cluster, the cluster 2 is the general human capital cluster, the cluster 3 is the education related cluster and the cluster 4 is the management and subordination cluster. Both cluster 3 and cluster 4 also contains some occupation specific skills and tasks. Cluster 5 to 8 are undefined occupation specific skills and tasks.

tokens in each cluster separately and for all three major occupations in Figure 2.[37] It shows that for the compensation cluster (V_1') and the general skills cluster (V_2') the token's relative frequency ratios are concentrated in value 1 with a shape close to normal distribution, indicating that these tokens are close to equally mentioned in different major occupations. For the education-related cluster (V_3') and the experience or position-related cluster (V_4'), the distributions become more dispersed and have more concentration close to 0 in high-skill Computer occupations, suggesting that they likely contain both general and occupation specific skills and tasks. For the rest of the clusters $V_5', \ldots, V_8'$, their tokens mainly concentrated close to 0, indicating that a majority of these words are likely to be very occupational-specific as they are way more likely or sometimes only to be mentioned in their own occupations. This left-skewed distribution is again more significantly in Computer occupations than in Administrative occupation, which makes intuitive sense because while specific skills in Computer are more likely to be some specific programming languages and thus very unlikely to be mentioned in other occupations, the specifics skills in Admin occupation involve more general terms like analysis, arrangement, or report which would likely to be used in many other occupations.

To sum up, in this section we classify the features selected in Lasso models to different types without any prior (except for the number of categories) and completely based on their associations in the job vacancy context, i.e. how firms write their vacancy text. We find that the results indicate a data-driven skill and task structure that is featured by skills and tasks with different levels of specificity. This skill and task structure or space distinguish with the official occupation categories in that it add very general skills that are irrelevant to occupation, and that it fulfills the within-occupation variations with detailed skills and tasks. It also distinguishes with the skill structures used in some recent labor literature that summarize the entire skill and task space using several board abstract categories like cognitive, noncognitive and interpersonal skills by showing that such classification will lose the dimension of skill and task specificity, which could potentially be important for thinking about issues like how the workers obtain different skills and to what extent do different skills transferable across different jobs. Moreover, in this clustering process we separate a cluster of compensation along with other skill and task clusters, which allows us to focus on the impacts of skills and tasks in the posted wage determination and discrepancy. In next subsection, we further reduce the dimension of the indicator matrices of these clusters so that we could bring these clusters of different job characteristics back to our wage differential estimation.

5.3 Dimension Reduction

In order to bring the selected hundreds or thousands of features in V' back to the wage regression in Section 4, we now further reduce the dimension of the indicator matrix of the selected tokens, $\mathbf{C}' \equiv \{\mathbf{c}_k\}, k \in V'$, to a reasonable size to ease the estimation. Relying on the clustering results that we have derived in Section 5.2, we will do this dimensional reducing separately for each cluster in each sample, i.e. reducing the dimension of $\mathbf{C}'_p \equiv \{\mathbf{c}_k\}, k \in V'_p$ for all p, so that we can distinguish the effects from the different types of job characteristics. For the

[37] It's not possible to plot the same figure for the Pooled sample but given that the structure of the clustering in the Pooled sample is similar to those in the occupational samples, our evidence here is also suggestive for the Pooled sample.

task of dimension reduction, unsupervised methods like principal component analysis (PCA) are often used. In fact, PCA projects the target data onto a lower dimensional space so that the variance of the projected data is maximized along each axis In other words, PCA finds a low-rank representation of $\mathbf{C}'_p$ that best preserves its covariance structure, but use no information about the structure of its predictive power and thus could generate unsatisfied results for our purpose here.[38] Instead, here we follow another suggestion in Gentzkow et al. (2019) to use a supervised method, partial least squares regression (PLS), to achieve a better performed dimension reduction.

In contrast to PCA, PLS performs dimension reduction by taking account of the information in the relation between the predictive and target variables. In particular, PLS projects both predictive and target variables into a lower-dimensional subspace such that the covariance between these two projections is maximized. Here because our target variable log wage has dimension one, this boils down to simply project each $\mathbf{C}'_p$ to 1D dimension and maximizing the covariance between this projection and the log wage. This procedure is iterated with orthogonalization to reach the desired number of PLS components Q. The details of the computation procedure are described in Appendix B.3. In essence, PLS forms the components by taking all features into a small set of linear combinations where the weights are decided by the predictive power of the features. We denote the resulted matrix of each cluster as $\Xi_1, \Xi_2, \ldots, \Xi_8$, whose indices correspond to the vocabulary clusters $V'_1, V'_2, \ldots, , V'_8$. In practice, we choose Q to be three, which means each Ξ will contain three vectors that represents three most useful dimension of the cluster in wage prediction.[39] Therefore, for each sample, we can now replace the indicator matrix of hundreds or thousands features with only twenty-four synthetic continuous variables. Running an OLS regression of all twenty-four variables on the posted wage, we find that, for all major occupations, the obtained R-squared is over 95 percent of the R-squared obtained in our Lasso regressions in Section 5.1 which use the full set of tokens (see Figure E3), indicating that our dimension reduction successfully preserves the majority of the predictive power of the tokens selected by the Lasso estimator.[40]

6 Main Results On Posted Wage Inequality

In last section we exploit machine learning methods to distill all wage-predictive job characteristics from vacancy text, to classify them into different clusters of skills and tasks, and to

[38] In particular, a predictive regression using principal components may perform poorly in data where the prediction target is strongly correlated with directions that have low variance because these directions, despite that their high predictive power, will be dropped in PCA. This problem could happen in our case due to the fact that the ill-understood features of the indicator matrix $\mathbf{C}$ that we have talked about in our motivation of using the word embedding model in the last subsection can carry over to each $\mathbf{C}'_p$.

[39] This choice of Q is again somehow arbitrary. We choose $Q = 3$ because three reduced variables under PLS are already able to account for most of the prediction power of the original token matrix of each cluster. Increasing Q further has little marginal improvement in the R-squared of the linear regression that use these reduced variables. Changing Q to two or four would not affect any of our results qualitatively.

[40] In comparison, the result from PCA or LSA (Latent Semantic Analysis, a direct singular value decomposition on the data without normalization and thus often used for sparse data in textual analysis) with three principal components ($Q = 3$) is significantly worse, achieving only around 50 percent of the R-squared in the Lasso regression.

generate the low dimensional proxy variables that preserve most of the information of these features. In this section we bring these job skill and task variables back to the econometric model in Section 4 such that we can accomplish our wage differential estimation and examine how do these newly-obtained and often-observed information improve our understanding of the wage determination and inequality in the labor market.[41] In order to better show our main results, in Section 6.1 we first bundle our skill and task clusters into different groups and specify the final composition of X. We then show our main results of posted wage differential in Section 6.2, where we not only illustrate the major components of posted wage variance but also further decompose the job effect to examine how different types of skills and tasks contribute to the job effect and firm-job sorting. In Section 6.3, we use the firm effects estimated in Section 6.2 to test some more features of the firm wage premium. Finally, in Section 6.4, we conduct several robustness tests on our estimation results.

6.1 Grouping Job Characteristics

The set of skill and task proxy variables obtained in Section 5, denoted $\tilde{\Xi} \equiv \{\Xi_2, \ldots, \Xi_8\}$, can be recognized as a representation of the full set of skills and tasks documented on the job vacancies by which firms use to match with their ideal workers and justify their posted wage. These variables thus incorporate not only the between-occupation skill and task variations as usually captured by the occupation dummies but also the within-occupation skill and task variations that are barely observed in the administrative or census data. For example, the features in $V_3^{'}$ used to generate Ξ_3 incorporate not only formal education information but also other related information such college major, certificates, and past projects, and features that takes $\{\Xi_4, \ldots, \Xi_8\}$ further contains dozens of detailed skills and tasks that will be required and conducted on a certain job. Despite this informational richness, listing one thousand of different skills and tasks and suggesting that they are all important for the wage determination does no help in improving our understanding of the source of wage dispersion. As a result, we need to further bundle these thousands skills and tasks, which have already been dimensional reduced into $\{\Xi_2, \ldots, \Xi_8\}$, along with other information we directly collect from the job vacancies, into some groups that we can give informative and meaningful interpretations. We will do this bundling procedure in two steps.

First, we set our specification of X in Equation (1) as $X = \left\{X_e, \tilde{\Xi}\right\}$, where $X_e \equiv \{\text{EXP}\}$ is the dummy variable of the experience level required in the job posts, and $\tilde{\Xi} \equiv \{\text{EDU}, \Xi_2, \ldots, \Xi_8\}$ is a bundle of the dummy variable of the education level required and all the skill and task clusters we extracted from the job texts.[42] The reason why we have this division is that $\tilde{\Xi}$ mainly

[41] Although during our machine learning procedures we also discover a cluster of non-wage compensations and amenities, which indicates that non-wage compensation provision might also be an important potential driver of wage determination, here we focus on job skills and tasks and leave the investigation of that specific aspect to Zhu (2022).

[42] We do not include any occupation information because, as we discussed in Appendix A.3, that the our occupation assignment algorithms are principally using the information of the skills and tasks documented in the job texts, and thus the extracted more granular variables $\{\Xi_2, \ldots, \Xi_8\}$ should include all the information that occupation groups embody. In fact we did a preliminary check by comparing the R-squared values of posted wage regressions with different specifications (Figure E4) and found that after controlling for our skill and task proxy variables $\Xi_2, \ldots, \Xi_8$, adding our constructed occupation dummies now almost generate no further increase

contains the indicators of the existence of one skill and task, while the experience required, X_e, can be seen as an indicator of the proficiency and competency of the typical skills and tasks required on a certain job.[43] We thus refer $\tilde{\Xi}$ as the extensive margin of the job skills and tasks, and X_e as the intensive margin. We will study the importance of both margins in the posted wage dispersions as well as the correlations between them two.

Second, to further distinguished the extensive margin of thousands of skills and tasks, we split $\tilde{\Xi}$ into three groups based on their levels of specificity that we have shown in Section 5.2. In particular, we set the group of the most general skills as $\Xi_g \equiv \{\Xi_2\}$, the group of the medium specific skills and tasks (or of board education information) as $\Xi_m \equiv \{\text{EDU}, \Xi_3\}$, and the group of the most specific skills and tasks as $\Xi_s \equiv \{\Xi_4, \ldots, \Xi_8\}$. The classification here may be considered a little bit arbitrary, especially for the cluster Ξ_4, which in some cases contains some organizational and hierarchical skills and tasks seemingly common across occupations along with other occupational specific skills and tasks. Given this difficulty, in Section 6.4 we will show that allowing for Ξ_4 to be included in the Ξ_m have no qualitative impact on our main results. A final note on the $\tilde{\Xi}$s is that the clusters and groups within $\tilde{\Xi}$ are not orthogonal variables but can be correlated with each other. In fact, the sum of the R-squared value for individual cluster (Figure E4) is way larger than the Lasso results, indicating strong complementarity or sorting between different clusters of skills and tasks. We will thus also check such correlations between Ξ_g, Ξ_m, and Ξ_s in our results.

6.2 Decomposition of Posted Wage Dispersion

The main results of our posted wage variance decomposition with full controls on job characteristics are shown in Table 4. Panel A, displaying the four most fundamental components of wage dispersion, shows that in our Pooled sample, the total share of wage variance is accounted by 45 percent job effect, 14 percent firm effect, 14 percent sorting, and 27 percent residual wage. Speaking differently, our estimation claims that the most important source of posted wage differences in our job vacancy data is the differences in the tasks and skills described in the job vacancy texts, and that either firm pay policies or the positive sorting between job differences and firm pay policies also play an important role in posted wage determination. Moreover, the estimated shares of these components are largely consistent with the results in the recent literature that use employer-employee panel data in rich countries and (bias-corrected) AKM approach.[44] For example, in Bonhomme et al. (2020) the authors use a

in the R-squared value. In addition, adding the education dummies has only limited improvement (about 1 to 2 percent points) to the R-squared value, because the information is largely overlapped with and thus absorbed by the cluster Ξ_3. However we still keep education in $\tilde{\Xi}$ because in some cases requirement of education is not documented in the vacancy text and thus cannot be captured by our machine learning procedure into Ξ_3. In fact, we will combine EDU and Ξ_3 into one bundle.

[43] For example, $\Xi_2, \ldots, \Xi_8$ may contain the information that if a certain type of programming language such as python is required or not, but does not contain the information of how many years of experience of using python. Although there are some words or phrases in the job post tests can indicate or represent the experience required, in a simple OLS wage regression test we find that our skill and task variables can absorb only a small part of the explanatory power of the experience dummies.

[44] One small deviation from the results in the literature is that under the AKM approach with worker fixed effect, the estimated variance of worker effect for more recent periods usually accounts for slightly over 50 percent and the estimated variance of residual wage accounts for 15 to 20 percent (see, e.g. Song et al., 2019). The under-

Table 4: Posted Wage Variance Decomposition

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.281	-	.253	-	.164	-
Panel A: $X = \{\textbf{EDU}, \textbf{EXP}, \Xi_2, \ldots, \Xi_8\}$								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.331	.067	.407
$\text{Var}(\epsilon_i)$	.098	.272	.074	.264	.071	.279	.058	.352
$\text{Var}(\psi_j)$	.049	.136	.071	.251	.056	.219	.028	.168
$2\,\text{Cov}(\theta_i, \psi_j)$	.052	.142	.054	.194	.044	.173	.012	.073
Panel B: Decompose θ **Terms**								
$\text{Var}(X_{int})$	.042	.115	.028	.099	.030	.119	.016	.096
$\text{Var}(X_{ext})$	.072	.199	.035	.126	.030	.118	.030	.183
$2\,\text{Cov}(X_{int}, X_{ext})$	.049	.136	.019	.067	.024	.094	.021	.129
$2\,\text{Cov}(X_{int}, \psi_j)$	.017	.048	.017	.060	.018	.072	.004	.025
$2\,\text{Cov}(X_{ext}, \psi_j)$	.034	.094	.038	.134	.026	.101	.008	.047
Panel C: Further Decompose X_{ext} **Terms**								
$\text{Var}(\Xi_g)$	.001	.002	.000	.001	.000	.002	.000	.002
$\text{Var}(\Xi_m)$	.006	.017	.004	.015	.002	.009	.005	.033
$\text{Var}(\Xi_s)$	.039	.108	.021	.074	.020	.078	.013	.082
$2\,\text{Cov}(\Xi_g, \Xi_m)$	.002	.005	.000	.002	.000	.002	.001	.004
$2\,\text{Cov}(\Xi_g, \Xi_s)$	.006	.017	.002	.007	.002	.007	.002	.010
$2\,\text{Cov}(\Xi_m, \Xi_s)$	.018	.049	.008	.028	.005	.020	.008	.051
$2\,\text{Cov}(\Xi_g, X_{int})$	.004	.011	.001	.004	.001	.005	.001	.006
$2\,\text{Cov}(\Xi_m, X_{int})$	.011	.031	.004	.014	.005	.019	.007	.041
$2\,\text{Cov}(\Xi_s, X_{int})$	.034	.094	.014	.049	.018	.070	.013	.081
$2\,\text{Cov}(\Xi_g, \psi_j)$	.002	.007	.002	.007	.001	.005	.000	.001
$2\,\text{Cov}(\Xi_m, \psi_j)$	.009	.026	.011	.038	.007	.028	.004	.025
$2\,\text{Cov}(\Xi_s, \psi_j)$	.022	.062	.025	.089	.017	.068	.004	.022
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Notes. The covariates X for the job effect θ_i now include education and experience dummies, X_e, and the proxy variables for different skill and task clusters, $\tilde{\Xi} \equiv \{\Xi_2, \ldots, \Xi_8\}$, that are derived from the machine learning algorithms in Section 5. In Panel B we decompose the variance and covariance terms that involve θ_i for X_e and $\tilde{\Xi}$ separately. And in Panel C we further decompose the terms that involve $\tilde{\Xi}$ by splitting it into three groups, $\{\Xi_g, \Xi_m, \Xi_s\}$,based on the level of skills and tasks specificity. All results here are corrected for finite sample bias by using KSS (leave-out) correction method (see Section 6.4 for the comparison between the plug-in estimates and the KSS estimates).

variety of matched employer-employee datasets from the US and several European countries, and find that despite substantial differences in labor market institutions and regulations, the estimated variance share of firm effects lay in the range of 5 percent to 15 percent and the estimate variance share due to firm-worker sorting concentrate within the range of 10 percent to 20 percent. Our results here suggest that at least in this high-end labor submarket in China, the wage inequality composition is similar to the broad labor markets in other developed countries. In Appendix C, we also compare the results here with the results from a specification where only information of education, experience, and detailed occupation is included, and find that the within-occupation job skill and task variations can explain about 5 percent of the total wage variances, which is over one-third of the variances accounted by between-occupation job variations.[45]

The estimation results in Panel A for three different occupation-level samples show that there are substantial differences in estimated components of posted-wage inequality both between the pooled sample and occupation-level samples and across different occupation samples. Comparing to the results in the Pooled sample, the results in all three occupation samples display lower shares (and values) of wage variance accounted by the job effect and higher shares (and values except the Admin occuptaion) by the firm fixed effect. A simple explanation for the reduced job variances is that for a certain board occupation the differences in job skills and tasks may be rather limited, but this cannot explain the increased accountability of firm fixed effects in the individual occupation samples.[46] One possible explanation for the higher share accounted by the firm effects estimated in individual occupational level is that if firm have different wage policies across different occupations, then our estimates on the Pooled sample will discard any variations of firm fixed effects across different occupations within the firm and produce underestimated share for firm effect variances.[47] We will explore this possibility and show the posted wage decomposition under a flexible wage regression specification in Section 7.1. Next we compare the differences in the posted-wage composition across three occupation samples with varying wage and skill levels. We find that while high skilled occupations like the Computer occupation have larger absolute value of variances due to job effect

estimated job components and overestimated residual components in our data can be due to (i) the measurement error of the mean of the posted wage range in our data comparing the real wage in the administrative data; (ii) there are other job differences that are not documented on the job texts; and (iii) there are additional variance due to worker effect which is not captured by the job differences due to mismatch between the job description and the real hired worker. Whatever the reason here, we think this relatively small deviation are not likely to qualitatively change any of our main results.

[45] In Appendix C we explain why this explanatory power of within-occupation job skill and task variations can be a lower bound and document that for individual major-occupation samples the within-occupation can account for substantially more wage differentials.

[46] One might wonder if the differences in the variance of firm effect is simply due to the stronger finite sample bias in the individual occupation samples, where we have less observations per firm. This is not true because the results in Table 4 are already under the KSS leave-one-out bias correction and in Section 6.4 we show the impact of the finite sample bias is limited in our framework.

[47] However, one may argue that this is also a problem of the definition of firm premium. Strictly speaking, a firm wage premium might be defined as a fixed premium equally given to all its employees. But if a firm decides only pay half of its member a wage premium that is higher than market level, how do we decide the level for this firm's wage premium? If such cases do exit, it also raises the following question that why and how a firm decides its wage premium across different employees, and more fundamental question eventually goes to where do firm wage premiums come from.

than the one in relatively low skilled occupations like the Admin occupation, the variance share of job effect is lower in the Computer occupation than in the Amin occupation, which may be a little bit surprised because we explore significantly more job features in high-skilled occupations than low skilled occupations. In fact, this is because the levels of firm effect and firm-job sorting are significantly higher in the high-skilled occupations than the low-skilled occupations. If this relationship holds generally, it implicates that the skill premiums in many high-skilled occupations or jobs are not only due to the fact that these jobs require high level skills and tasks which are priced highly in the labor market, but also due to the fact that these jobs are more likely to be attached with high wage policies and be sorted with firms that pay such high wage premiums. However, the limited number of occupations in Table 4 prevents us from ensuring this relationship, and thus we will leave this question to Section 7.3 where we use one method to generate enough numbers of occupations from our data.

The merit of our machine learning and vacancy data approach is that, different from the methods that using a worker fixed effect, we can unmask our job effect and investigate the effects of different types of skills and tasks on both job effect and firm-job sorting. In panel B we first distinguish the experience controls X_e and our constructed skill and task controls $\tilde{\Xi}$. The results show that in the Pooled sample, the wage variance share due to the intensive margin, X_e, is 11.5 percent, the share due to extensive margin, $\tilde{\Xi}$, is 20 percent, and the covariance terms between these two margins also account for 13.6 percent. For occupation-level samples, the variance share due to experience is similar to the level of the Pooled sample, though the variance share attributed to the extensive margin of job skills and tasks is lower, which is intuitive as the extensive margin will be more limited conditional on jobs from a certain occupation. Therefore, both margins of the job differences significantly contribute to the wage differentials, and there are positive correlation (0.44) between these two sources of job variations, indicating that firms which require high-wage skills and tasks are also more likely to require high-experience workers.[48] Moreover, the covariance terms with the firm fixed effect show that these two margins contribute to the posted wage variance due to job-firm sorting with a similar relative importance as their contribution to the job effect. Therefore, high wage premium firms are sorted with high-skilled workers in terms of both high-valued skills and high experiences on the skills.

Next, we further decompose all terms related to $\tilde{\Xi}$ in Panel C to examine the roles played by different types of skills and tasks. The decomposition results make it clear that at both pooled level and occupational level, general skills, as represented by Ξ_g, almost does not explain for any job effect and account for a very small fraction of sorting with firm effect or with X_e. In contrast, the most specific group of skills and tasks, Ξ_s, account for a majority of both job effect and sorting with firm effect and with external margin X_e, and the board education cluster with medium-level of specificity, Ξ_m, account for a relatively small part of all the effects and sortings. In particular, for the Pooled sample and the high-skilled Computer occupation sample and the medium-skilled Design & Media occupation sample, the variance and covariance terms of Ξ_s account for over 80 percent of the total variance of $\tilde{\Xi}$ and contribute to over 60 percent of the covariance of $\tilde{\Xi}$ with firm effect and X_e, and the rest percents basically goes to Ξ_m. For

[48] Although in our Pooled sample it seems that the experience variations explain only over half of the extensive job skill and task variations, we think our results might underestimate the importance of experience because our data is relatively preoccupied by low-experience jobs comparing to the real labor market. Hence we think the safer takeaway here is that both margins are important for posted wage dispersions.

the low-skilled Admin occupation sample, the medium-specific education cluster accounts for significantly more shares posted wage variances, and even contribute slightly more than Ξ_s for the firm-job covariance terms. These results indicate that the most important source of various job skills and tasks that generate large posted wage variances in the labor market are those most specific skills and tasks, potentially linking with firm technologies and productivities. For those low-skilled occupations with relatively lower wage dispersions, there are limited specific skills and tasks, which is perhaps why they are low-skilled and low-wage at first place, and thus the worker characteristics like education, certificates, or other basic human capital indicators play an important role in wage dispersions.

We believe that our findings here provide some new and intuitive evidences on the micro-foundation of both the firm pay premiums and firm-work sorting in the wage dispersion literature and the popular argument of skilled biased technological change (SBTC) in the wage inequality and labor demand literature. For high skilled occupation like Computer occupation, firms' different usages on new skills and tasks like machine learning or AI largely derive the differences in wage, generating the observed firm pay premiums and/or firm-worker sorting. And because such different requirements are positively correlated with firms requirements on education-related skills, they could potentially help to generate results like college premium. Whereas for the low skilled occupation like Admin occupation, technology advance has rather less impact on the skills and tasks, if not directly replacing some, and thus firms are likely to ask for a bundle of skills and tasks which are relatively less specific and have limited extent of firm wage premium and firm-worker sorting. Our results that under the hood of worker effect which contains all types of skills and tasks, it is those most specific ones that contribute the most to the wage inequality seems to contradict with the results in the earlier study (Deming and Kahn, 2018, e.g.) that find general skills like cognitive skills and social skills hold statistically significant predictive power on the wage differentials. In order to resolve this inconsistency, in Appendix D we replicate the main analysis in Deming and Kahn (2018) and show that we can replicate the significantly positive correlation between the posted wage and keyword-based cognitive and social skill indicator variables in our data. However, we also show that (i) many keywords behind the indicator variables can actually indicate specific skills, (ii) the values of the coefficients will be reduced significantly if we further controls for other job characteristics documented in the jobs ($\{\Xi_2, \ldots, \Xi_8\}$), and the most importantly, (iii) that such statistically significant correlation matters little for the entire posted wage variance simply because the variations of those general skills across the vacancies are too small to account for any nontrivial wage dispersion component.[49]

To sum up our main findings, equipped with the full controls for job skills and tasks derived from the job text, the estimation on our entire data sample generates wage inequality components, namely the job effect, the firm effect, and the firm-job sorting, consistent with the results in the previous studies that use (bias-corrected) AKM approach and employer-employee panel data in rich countries. However, we also find that there are significant heterogeneity of wage differential components across board occupations, such that higher skilled occupations

[49] As we have discussed in the introduction, our results does not necessarily mean that those general skills are not important at all. In fact, the cultivation of specific skills is likely to require general skills and the within-firm wage changes can be potentially affected by the exposure of signals of general skills over tenure. Our results here are more about saying that for the posted wage dispersions, those terms relevant to general skills that employers document on their job posts seem to not matter.

with more job features detected have also larger variance of firm effect and firm-job sorting, and we will further examine this possibility in Section 7. By decomposing the job effect, we find that both the extensive job skill margin captured by our constructed skill and task variables and the intensive job skill margin captured by the experience dummies contribute significantly to the wage dispersions and firm-job sorting, though in our data the extensive margin contributes more. Further decomposition on the extensive margin of various skills and tasks makes it clear that it is not the variations in those general skills but the variations in those specific skills and tasks that explain the posted wage differentials. In particular, the bundle of most occupational-specific skills and tasks account for a majority of all types of effects, and the bundle of education-related medium specific skills explain for a smaller share except in the low-skilled occupations, whereas the contribution from the most general bundle is rather trivial. We also find strong and positive correlations between extensive and intensive skill margins and between all types of skills and tasks bundles, which indicates that there could have important complementarities across all different dimensions of the human capital space, and raises the caution for the importance of taking account of all types of skills and tasks when identifying the wage effect of any individaual type of skills or tasks. Overall, our results offer a detailed picture on how different dimensions and types of skills and tasks contribute to the posted wage inequality and provide new insights and hints for understanding the deep nature of other components of wage inequality such as firm wage premium and firm-worker sorting.

6.3 Firm-Specific Posted Wage Policies

The results in last subsection show that even after controlling for almost all the information in the job vacancy text we can still observe a substantial part of posted wage variation attributable to firm fixed effects, i.e. firms have different pay policies even if they document exactly the same job tasks and worker requirements in the job vacancies. This indicates that the firm premiums does not only exist in real wages but also in the posted wages that firms post in the labor market . There are several explanations that have been suggested in the literature, including compensating differentials, efficiency wages, search frictions, and rent sharing, but few consensus has been achieved. The literature also documents positive linkages between firm pay policy and firm productivity or firm size (see e.g. Barth et al., 2016; Kline et al., 2020, among others) and firm location (see e.g. Dauth et al., 2022; Hou and Milsom, 2021). In this subsection, we examine if these correlations are also hold for our posted firm wage policies.

In order to do this, we regress the estimated firm fixed effects on firm size and firm location dummies, which are the available firm characteristics in our job vacancy data. Because from our wage variance decomposition we already know that firm fixed effect is positively correlated with job quality, we also try to include our estimated firm-average level of job characteristics, $\bar{\theta}_j$, into the regression in case that the variables of interests are correlated with both firm wage premium and job quality. The results of our regression are shown in Table 5. Firm employee size is significantly and positively correlated with firm fixed effects in all types of specifications, and coefficients for firm size dummies only decrease by a limited part after adding job effect $\bar{\theta}_j$. This positive correlation is again consistent to the results in the literature that use employer-employee data and AKM framework (see e.g. Kline et al., 2020). Despite the statistical significance, the R-squared in the specification with only firm size categories is less

than 2 percent, indicating that the part of firm wage premium that can be explained by firm size categories is very limited. In contrast, we find that the dummies of work location can still explain a large part of the firm wage premium even after controlling for average job quality and firm size categories, increasing the R-squared for slightly less than 30 percent points in most samples except for Admin occupation. This suggests that the firm wage premium may be partly due to different bargaining power under outside option differences across different regions, or due to different levels of productivities and rents under different levels of geographical agglomeration, along with other geographical reasons.

In short, similar to other studies that use administrative data, our estimated firm-specific wage policies are also correlated with firm size and can be partly explained by the firm location. More data on the firm-side as well as better econometric settings or economics models are necessary to further identify the exact sources of the differences in firm wage premiums, and we leave it for future study. However, in Section 7 we will provide some more empirical features about the firm posted-wage effects in the occupation level, which hopefully can shed some light on the nature of the firm-specific wage policies.

6.4 Robustness

In this section we provide several robustness tests on our results in Section 6.2 and Section 6.3.

Finite Sample Bias. As we have mentioned in Section 4, given the high-dimensional firm fixed effects in our regression model, the variance and covariance terms of firm fixed effect could be biased especially when the observed vacancies of a given firm are limited. In fact this finite sample bias can be also called "limited mobility bias" following the AKM literature due to the fact that the deep nature of limited mobility bias in the AKM approach is the finite sample bias for identifying firm-level wage differences and that limited vacancies observed can be also regarded as one type of limited job mobility for a certain firm. As a result, we can use the methods that are developed in the AKM literature to resolve the finite sample bias to correct the finite sample bias here. In particular, we use both the homoscedasticity correction approach suggested by Andrews et al. (2008) and the heteroscedasticity leave-out correction approach suggested by Kline et al. (2020), and the main results shown in Section 6.2 is under the heteroscedasticity leave-out correction. The comparison of two different types of corrections with the plug-in results are shown in Table E2. The results show that both corrections have very similar results in which the firm effects are reduced and the part accounted by the error terms are correspondingly increased. The corrections are not substantial in the pooled sample, where the changes are about 0.5 percent point, but more significant in the Admin occupation, where the changes are around 5 percent points. This difference is simply due to the fact that in the pooled sample we have less firms with a very small number of vacancies posted, and thus the finite sample bias is rather limited. Our estimated job effects are not subject to any corrections because our controls on job characteristics are either sparse categories or continuous variables. The correction also has very insignificant impact on our estimated firm-job covariance because empirically the finite sample bias in our case will only have second-order effect and thus only appear when the finite sample bias is very large.

In addition to our results in Section 6.2, the finite sample bias also matters for the results in

Table 5: Firm Fixed Effect and Firm Characteristics

	Pooled			Computer			Design_Media			Admin		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
fsize.15-50	.019**	.018**	.023**	.011+	.013*	.019**	.022**	.013**	.020**	.006	.005	.005
	(.004)	(.003)	(.003)	(.006)	(.005)	(.004)	(.005)	(.005)	(.004)	(.006)	(.006)	(.006)
fsize.50-150	.042**	.037**	.050**	.037**	.032**	.038**	.050**	.033**	.045**	.020**	.018**	.021**
	(.004)	(.003)	(.003)	(.006)	(.005)	(.004)	(.005)	(.005)	(.004)	(.006)	(.006)	(.005)
fsize.150-500	.067**	.057**	.067**	.072**	.054**	.051**	.086**	.058**	.063**	.035**	.031**	.030**
	(.004)	(.004)	(.003)	(.006)	(.005)	(.005)	(.005)	(.005)	(.004)	(.006)	(.006)	(.006)
fsize.500-2000	.095**	.078**	.085**	.108**	.074**	.066**	.127**	.087**	.086**	.050**	.043**	.040**
	(.005)	(.004)	(.004)	(.007)	(.006)	(.005)	(.006)	(.006)	(.005)	(.007)	(.007)	(.006)
fsize.2000+	.121**	.102**	.120**	.140**	.084**	.082**	.161**	.107**	.108**	.064**	.055**	.058**
	(.005)	(.005)	(.004)	(.008)	(.007)	(.006)	(.007)	(.007)	(.006)	(.008)	(.008)	(.007)
Job Effect ($\tilde{\theta}$)		.287**	.201**		.643**	.498**		.391**	.292**		.118**	.063**
		(.004)	(.003)		(.007)	(.006)		(.006)	(.005)		(.008)	(.008)
const	.146**	-1.115**	-.633**	.222**	-2.684**	-1.905**	-.030**	-1.759**	-1.208**	.024**	-.478**	-.166**
	(.003)	(.016)	(.015)	(.005)	(.030)	(.027)	(.004)	(.028)	(.024)	(.006)	(.036)	(.033)
Location FE			✓			✓			✓			✓
Adj. R^2	.016	.096	.377	.016	.168	.436	.022	.100	.390	.006	.014	.229
No. Obs	86165	86165	86165	62628	62628	62628	55664	55664	55664	41448	41448	41448

Notes. The baseline group for firm size fixed effect is the group of firms with less than 15 employees. In the case that a firm has multiple entries of size or location information we use the categories that are most recorded in the vacancies of the firm.

the firm wage premium regression in Section 6.3. Kline et al. (2020) shows that under strong finite sample bias, inference of the regression will be biased and lead to potentially wrong conclusions. However given that the finite sample bias is rather limited in our case, it is likely that this bias would also be small. To confirm, we remove firms with less than ten vacancies in each sample and then redo our tests in Section 6.3, as what we have done for our pooled sample in the sample cleaning. The results are in Figure E5 and Table E5, which show that our results in Section 6.3 largely maintains under these limited samples.

Compositional Differences. One potential concern on our results in Section 6.2 is that the different importance of different types of skills and tasks across different occupations might be driven by the compositional differences across different occupations. In particular, because in our data high-skilled occupations contain more vacancies with requirements of higher education and experience levels than low-skilled occupations, our results could be misleading if those specific skills and tasks are only important for wage differential in high education and high experience vacancies jobs and if our data does not correctly represent the true composition in the labor market. To resolve this concern, we slice our sample given certain education or experience level and redo our estimations on these sliced samples, and we find our main findings remain. For example, Table E3 shows the estimation results when the experience is conditional to be 0 (no requirement), where we can still observe that for the pooled sample and Computer occupation, those most specific skills and tasks account for the most share of total wage variance, while for the Admin occupation, it's those medium specific skills and tasks account for the major shares.

Specification of Skill and Task Groups. As we have stated in Section 6.1, it is difficult to decide the divide line between the bundle of medium-specific skills and tasks and the bundle of the most occupation-specific skills and tasks. In our baseline specification, we include the Ξ_4 into the most specific group Ξ_s, because although it contains some tasks of management, supervision, coordination, and subordination, which may share some generality across different occupations, we also find that organizational or positional tasks are often gathered together with occupational specific tasks. But still one may wonder how the results change if we assign Ξ_4 into the medium-specific group Ξ_m. To test how important are the classification of this cluster for our results, in Table E4 we show the variance decomposition results when Ξ_4 is moved from Ξ_s to Ξ_m. It turns out that such change will change our results mainly quantitatively but not qualitatively. In specific, for the Pooled sample and the high-skilled or medium skilled occupation sample, while the most specific group Ξ_s still is the most important attributor of different components of wage dispersion, the medium specific group Ξ_m now increase a lot in its contribution. And for the low-skilled Admin occupation, now it is the medium specific skills and tasks Ξ_m that account for the most share of the wage dispersions, leaving the most specific skills and tasks in the occupation Ξ_s a rather small role. These results reconfirm our earlier finding that there is linkage between the level of the skill and task specificity and the wage dispersion for an occupation. For high skilled occupations, both the high wage level and wage variance is likely achieved though a large variation of specific skills and tasks, while for low skilled occupations, firms reply more on medium-specific skills like education or general experience to determine wage and sort with workers, hence generating a lower dispersion.

Linear additive specification. One concern about the AKM framework that have been mentioned in the literature is that the assumption of additive separability between the worker term and the firm term could be too restrictive and prohibit any flexible interactions. As discussed in Section 4, we also have this similar linearity between the job term and the firm term in our specification. As a result, we follow Card et al. (2013, 2016) to test the validity of this assumption by checking the mean residuals of different job-firm cells. The results in Figure E7 show that while in individual occupation samples the mean residual for most firm-worker cells are close to zero, in the pooled sample there do have a quite large part of firm-worker cells with mean residual significantly deviated from 0, though the levels of most deviations are moderate. This result further supports our hypothesis that firm wage premiums can potentially differ across occupations, which we will formally test in Section 7.

7 Extensive Analyses

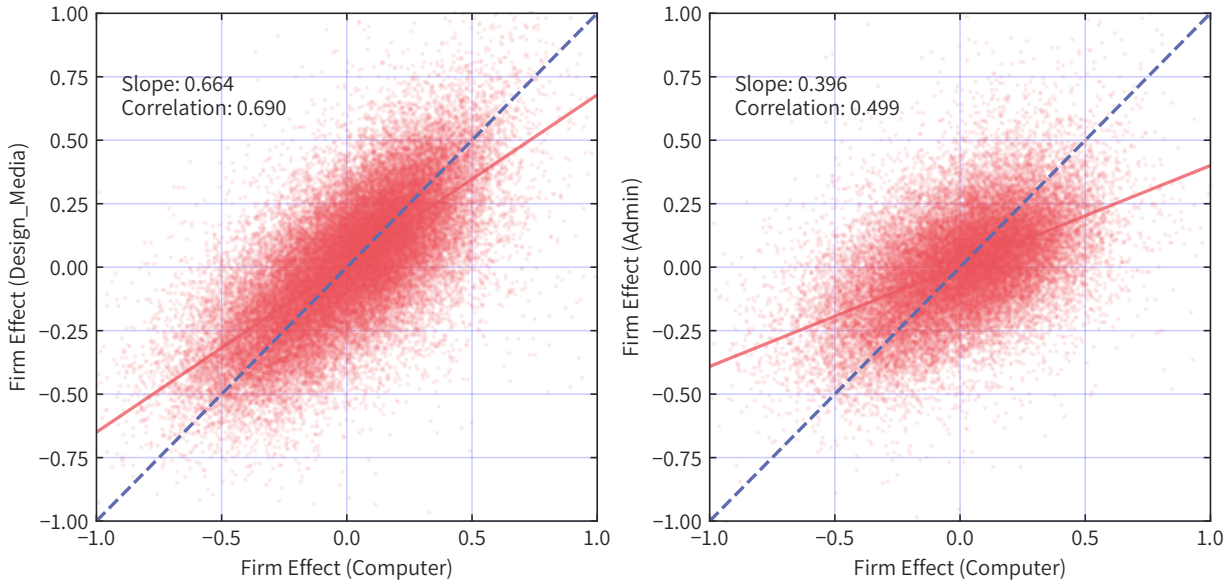
In this section, we extend our baseline econometric model in Section 4 to more general specifications and try an alternative approach to estimate the high-dimensional wage regression. We use these specifications and approach to study the occupational-specificity of the firm premiums and skill prices (Section 7.1), the potential complementarity between firm effect and job effect (Section 7.2), and the differences in wage components across occupations (Section 7.2). The aim is both to check the validity of our econometric model and to learn more about the posted wage dispersion. Finally we also show some results for the trend and determinants of the posted wage dispersion in our data (Section 7.4). All the new estimations in this section are conducted directly on the Pooled sample.

7.1 Occupational-Specific Firm Premiums and Skill Prices

In this subsection, we test whether our baseline econometric model Equation (1) is mis-specified because of the existence of the heterogeneity of firm pay policies and skill prices across different occupations. We first do a simple test on if a firm pays the similar wage premiums for different occupations within the firm by using the occupational level estimations in Section 6. To be specific, we find the overlapped firms in two individual major occupations and plot the two separately estimated (demeaned) levels of firm fixed effect in Figure 3. For both subfigures in Figure 3, the x-axis is the estimated firm fixed effects in the Computer occupation sample, and the y-axis is the estimated firm fixed effects in the Design & Media occupation and the Admin occupation, respectively. While in both cases there is a strong positive relationship indicating that firms' posted wage policies are in general consistent across occupations, the slopes of a linear regression are less than one which suggests that firms incline to pay a smaller degree of wage premiums (and discounts) in the Design & Media and Admin occupations than the Computer occupation. Moreover, the level of the deviation from the firm effects in Computer occupation is more significant in the low-skilled Admin occupation than the medium-skilled Design & Media occupation. In particular, while for the pair between Design & Media occupation and Computer occupation the slop and the correlation is both close to 0.7, for the pair between Admin occupation and Computer occupation the slop is 0.4 and the correlation is

0.5. In other words, the firm wage policy for an Administrative vacancy is more likely to be irrelevant to the same firm's wage policy for a Computer vacancy. As such divergence can be simply stemmed from the difference levels of observation sample size and measurement error, we repeat this analysis by estimating the firm fixed effects with the sample of firms that have more than ten vacancies in both occupations of each pair Figure E5, and find the similar results. Therefore, this preliminary check confirms the possibility that firms pay different levels of wage premiums for different occupations, and suggests that this differences will be especially large when two occupations have very different skill or wage levels.

Figure 3: Variation of Firm Effects Across Occupations



Notes. Firm effects are estimated using the specification of X in Section 6.1. We then find the set of firms that have both estimated firm effects in each two pairs of major occupations. A linear regression is then estimated on the demeaned firm effects for each pair. The red line shows the slop of the regression and the blue dash line is the 45 degree line.

Next, we then formally test two extensive specifications of the wage regression model in Equation (1),

$$\ln w_{i,j,o,t} = X_i\beta + \psi^o_j + \iota_t + \epsilon_i \tag{6}$$

,

$$\ln w_{i,j,o,t} = \sum_o \mathbb{1}_{[i\in o]} X_i\beta_o + \psi_j + \iota_t + \epsilon_i \tag{7}$$

, where $o \equiv o(i)$ denotes the major occupation where the job vacancy is classified in. In Equation (6), we allow for the occupational-specific firm premiums ψ^o_i, and in Equation (7), we allow for the occupational-specific skill prices β_o. In addition, we also compare Equation (6) with a specification that simply adds an occupation fixed effect to Equation (1), $\ln w_i = X_i\beta + \psi_j + o_i + \iota_t + \epsilon_i$, which allows for a fixed occupational wage premium invariant across firms, unlike the more flexible wage premiums in Equation (6). We use these three

specifications to estimate our Pooled sample, and the results are shown in Table 6. The first benchmark column are the results of the baseline specification in Section 6. The third column shows that after allowing for occupation-specific firm pay policies, the variance share of job effect declines to 38 percent and variance share of firm effect and firm-job sorting now increase to 18 percent and 20 percent, respectively. This large change in posted wage components suggest that both firm wage policies and the firm-job sorting will be more important drivers of the posted wage variances if you allow for firm varying wage policies across jobs of different occupations. Moreover, the results in third columns is significantly different from the second column where only firm-invariant occupation wage premiums are allowed, indicating that different firms could have different wage policies on one certain occupation. In addition, when we plot the residual mean distribution generated from the specification of Equation (6) across the job-firm decile cells as we do in Section 6.4, the extent of the deviations of residual mean from 0 now decrease a low (see Figure E9), suggesting that this might be the better specification of the posted wage generation. Finally, in the fourth columns we test the specification Equation (7) that allows for occupational specific skill prices and find there is only limited increase (2 percent points) in the job effect and no discernible impact on firm effect or firm-job sorting. This insignificance might be due to the fact that a major part of skills and tasks are occupational-specific and thus does not overlap across different occupations at all.

Table 6: Posted Wage Variance Decomposition Under Different Specifications

	Benchmark		$\psi_j \equiv \hat{\psi}_j + \hat{o}_i$		$\psi_j \equiv \hat{\psi}^o_j$		$\theta_i \equiv X\hat{\beta}_o$	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.362	-	.360	-	.361	-
$\text{Var}(\theta_i)$	.163	.450	.141	.391	.136	.378	.170	.470
$\text{Var}(\epsilon_i)$	.098	.272	.096	.265	.088	.245	.092	.255
$\text{Var}(\psi_j)$	.049	.136	.056	.156	.065	.182	.049	.136
$2\,\text{Cov}(\theta_i, \psi_j)$	.051	.142	.068	.188	.070	.196	.050	.139
Obs	3998840		3998840		3926231		3998840	
Firm	86165		86165		300079		86165	

Notes. The benchmark results is the decomposition results for the Pooled sample in Table 4, but the results may be slightly different due to the matrix simulation procedure used in the estimation. The second and third board columns shows two different specifications which allow for occupational-specific firm pay policies. And the final board column shows the specification of occupational-specific skill prices.

One possible reason behind the fact that a firm vary its wage policies across jobs of different occupations can be that firms only pay high wage premiums for some core jobs (e.g. IT engineers for an IT firm) within the firm. And the further reason behind the wage premiums can be either complementarity in production function, rent sharing, or efficiency wage to avoid higher turnover cost, as suggested in Bloesch et al. (2021). On the other hand, those auxiliary jobs could have less complementarity to those core jobs, less rent sharing, and less turnover cost, and thus firm would not pay an equally high wage premium for those jobs. If such idea holds true, it also explains our earlier finding that high-skilled occupations also has larger posted

wage variance due to firm effect and firm-job sorting than low-skilled occupations. Before we test this relationship between the importance of the firm pay policies and firm-job sorting with the wage or skill level across occupations, we first introduce, in next subsection, an alternative way to estimate our posted wage model, which will also offer some convenience for our investigation on the occupational heterogeneity of posted wage components.

7.2 A Shortcut and Linearity

In this subsection, we propose a shortcut to estimate the major components of posted wage variance, namely the job effect, firm effect, and sorting between job and firm, by clustering job vacancies and/or firms into different types. The key idea is motivated by the estimation strategies for employer-employee data proposed in Bonhomme et al. (2019, 2020), where they first use K-Means clustering algorithm to classify firms into a number of discrete types based on the information of within-firm wage distribution, and then estimate the worker types as correlated random effects in a statistical model where conditional wage distributions for job movers and job stayers are explicitly modeled. Here, rather than estimating the statistical model with correlated random effects, we exploit the embedding space estimated from the word-embedding model in Section 5.2 to directly classify the job vacancies into different types of jobs. We suggest that our job clustering algorithm is essentially an unsupervised classification of different jobs into arbitrary number of occupations based on the differences in the job texts, i.e. how the employers describe their jobs in the nature language. One perhaps surprising result in Bonhomme et al. (2019, 2020) is that they show that the estimation under even fairly low-dimensional firm categories from a simple classification method yields quite close results to the estimation under full high-dimensional firm dummies with further bias correlation. We will test to what extent the similar results hold for our job clustering approach. We also use this alternative estimation approach as a robustness test for our dimensional reduction algorithm in Section 5.3 used to conduct our baseline estimation. Finally, Bonhomme et al. (2019) uses their approach to test the linearity assumption of the AKM framework and find supporting evidences, and we will conduct the similar tests for our posted wage regression model.

We now explain the details of the shortcut way of estimation. In Bonhomme et al. (2019), they illustrate that one can exploit the information of the within-firm wage distribution and use a simple weighted K-Means algorithm to identify the latent firm types in the employer-employee data. In particular, the K-Means algorithm does the following likelihood maximization,

$$\min_{\mathfrak{k}_1,\ldots,\mathfrak{k}_J,H_1,\ldots,H_{\mathfrak{K}}} \sum_{j=1}^{J} n_j \int \left(\widehat{F}_j(w) - H_{\mathfrak{k}_j}(w)\right)^2 d\mu(w) \tag{8}$$

, where $\mathfrak{k}_j$ denotes the firm type that firm j is assigned, $\mathfrak{K}$ is the pre-determined number of the firm types, n_j is the number of workers (or job vacancies in our context) that the firm has in the dataset, $\widehat{F}_j$ is the empirical wage distribution function within the firm j, $H_{\mathfrak{k}_j}$ is the wage distribution of the cluster $\mathfrak{k}_j$, $\mu(\cdot)$ is a measure of the wage distribution. In practice, Bonhomme et al. (2019, 2020) use 20 percentiles of the log wage distribution as the distribution functions, and set $\mathfrak{K} = 10$. This is the first step in Bonhomme et al. (2019, 2020), and in the second step, the authors use the estimated firm classes to further estimate a correlated random effect model

in order to obtain the worker classes and the conditional wage distributions. Instead of their second step, here we propose a simple and context-specific way to estimate the job classes in our job vacancy data, which shares a very similar idea to the first step in Equation (8). The key observation here is that in Section 5.2 our word-embedding model has already map each selected features in the vocabulary V' into a high dimensional space, $\mathbf{u}_k$ with $k \in V'$, and thus we can do a simple linear transformation for all the selected features to get a high-dimensional representation for each job vacancy. In fact, we can just sum up all the selected features in a job vacancy i with vocabulary V_i', i.e.

$$\mathbf{z}_i = \sum_{k \in V_i'} \mathbf{u}_k = (z_{i1}, \ldots, z_{iH}) \tag{9}$$

, where $\mathbf{z}_i$ is now the high-dimensional vector of the job vacancy i in the embedding space. Then we can apply a K-Means algorithm to all z_i to obtain arbitrary numbers of clusters of job vacancies. In fact, if we denote the job classes for all i as $\mathfrak{l}_i$, and the total number of job classes as $\mathfrak{L}$, we can then use the following minimization,

$$\min_{\mathfrak{l}_1,\ldots,\mathfrak{l}_I, G_1,\ldots,G_{\mathfrak{L}}} \sum_{i=1}^{I} \sum_{h=1}^{H} \left(z_{ih} - G_{\mathfrak{l}_i}(h)\right)^2 \tag{10}$$

, where $G_{\mathfrak{l}}$ is the value function of the dimension h for the job class $\mathfrak{l}$. Using both Equation (8) and Equation (10), we can thus have two separate steps to estimate both the firm classes and the job classes, and while the former is using the wage information, the latter is using the job text information. One interesting interpretation of our job classification in Equation (10) is that we are actually doing occupation classification without any target categories. In other words, we cluster different job posts by checking if the employers use the similar types of words and terms, and this imitates how people distinguish jobs: they put jobs with similar job tasks into a bundle and call it an occupation. Thus we can also use Equation (10) to generate arbitrary number of occupations from our data or any other job vacancy data, and the generated occupations can be more granular than any human defined occupation categories.

Because there is no perfect way to define $\mathfrak{K}$ and $\mathfrak{L}$, in practice we choose a set of different numbers, and use the estimated firm and job classes to estimate our posted wage regression with a new specification $X = \{\text{EXP}, \text{EDU}, \mathfrak{l}\}$, and/or with the firm fixed effect now replaced with a fixed effect of the firm class $\mathfrak{k}_j$. Figure 4 plots the estimated posted wage components in the Pooled sample when using only job clusters, using only firm clusters, and using both clusters. We also plot the baseline results in Section 6.2 in dotted line as the benchmark. Figure 4 shows that when we replace our job characteristics variables with only 10 job classes, the estimated shares due to job effect is about 7 percent points less than the benchmark result, and the estimated shares due to firm effect and residual wage are higher than benchmark one. If we increase in number of job classes that the job vacancies can be clustered, the job effect share increases continuously, and the firm effect shares and residual wage shares decline correspondingly, and with 320 job clusters, the results are now quite close to the benchmark shares. This indicates that similarly to our algorithms in Section 5, our job clustering algorithm in Equation (10) are also able to capture the within-occupation job skill and task variations

as long as we allow for enough types of jobs in the labor market. In contrast, for the cases where firms are replaced by the firm classes, we observe that even with a quite low number of firm clusters, say 10 or 20, the estimated firm effect is already quite close to the benchmark level, and further increasing the number of firm classes only make the results slightly more close to the benchmark value. This feature is consistent with the finding in Bonhomme et al. (2019, 2020). Therefore, it seems that while the heterogeneity of firm pay policies can be approximated by rather few firm types, the heterogeneity of job characteristics can only be well approximated by significantly larger number os job types. Also we find that when we using the firm clusters instead of firm fixed effect, the estimated shares due to sorting is larger than the benchmark level, again consistent with the results in Bonhomme et al. (2020). Finally, when we use both job and firm clusters, it shows a mixed result of the two cases.

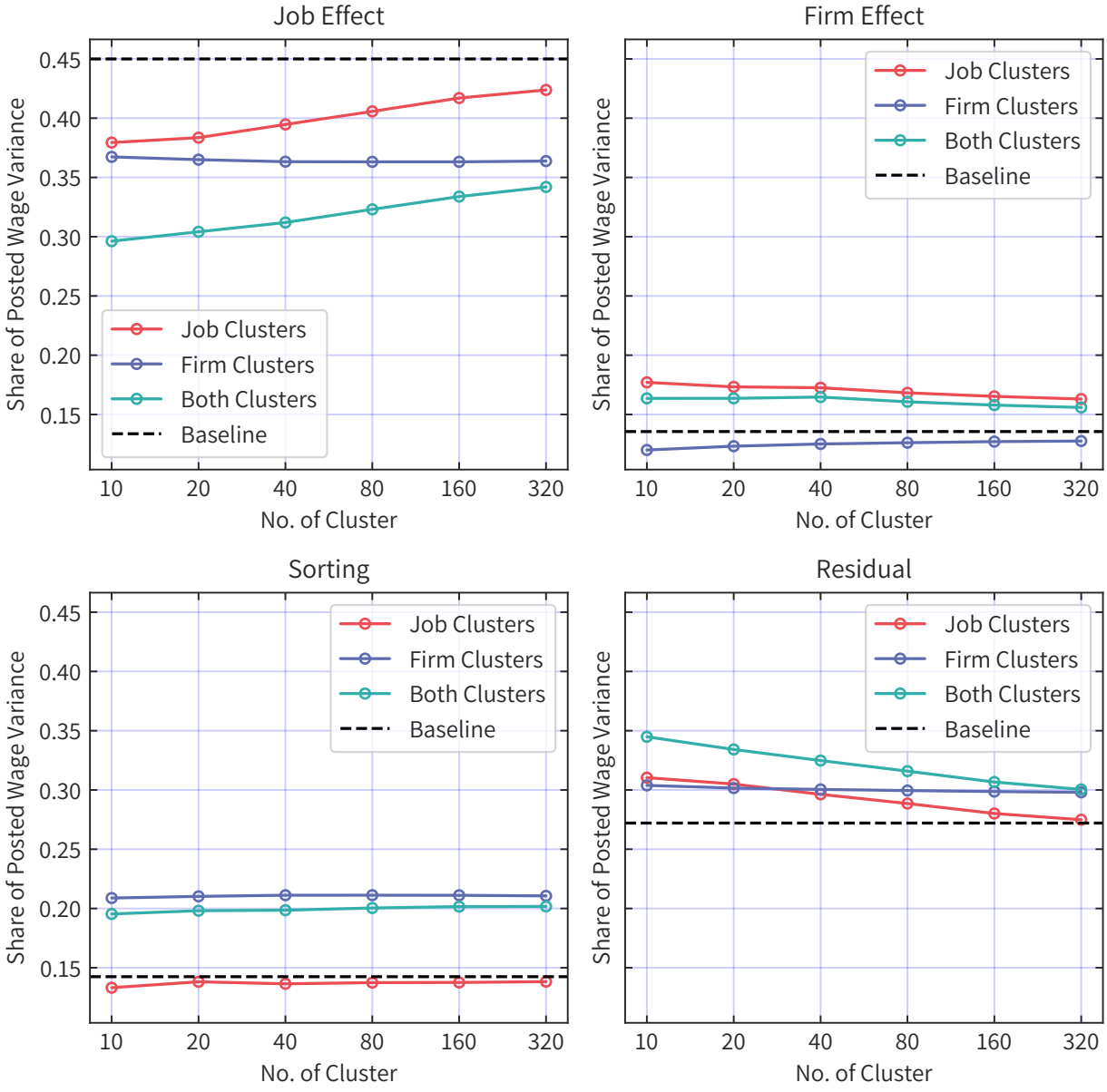
With the low dimensional firm and worker clusters in hand, we can now follow the analysis in Bonhomme et al. (2019) to examine the validity of the separable additivity assumption in our posted wage regression. We first plot the job composition by firm types and the mean log posted wage by ten firm types and five job types in Figure 5, and both firm and job types are ranked based on their estimated values. The left panel of the job composition depicts clear pattern of firm-job sorting. The highest wage-premium firms have about half of their posts belong to the highest skilled jobs, while the lowest pay-premium firms post less 5 percent of such high-skilled jobs but post more than 40 percent of the lowest type of jobs. The right panel of the mean log-wage of the job posts illustrates that while there is rather limited evidences of complementarity for the three job types with highest job values since the wage lines are generally in parallel and thus increases in mean wage along with the firm ranks for different types of workers are similar. However there are some evidences of a lack of complementarity for the two job types with the lowest job values, especially for the lowest type of job.[50] We observe the similar features when we allow for significantly more job types, as we show in Figure E8. Next, given that we can now have a posted wage regression specification with job controls and firm controls being replaced by low-dimensional job and firm classes, we can allow for more flexible interaction terms in the regression, similar to the ψ_i^o term in Equation (6). When we compare the results from a posted wage regression with interaction term between job and firm classes to the results from a simple linear specification of these two classes, we find very limited decreases in adjusted R-squared, indicating a small role of firm-job complementarity in our data, which is again consistent with the results from administrative employer-employee data in Bonhomme et al. (2019).

7.3 Posted Wage Decomposition Across Occupations

In our main results in Section 6.2, we show that among the three selected major occupations, the higher skilled occupation in general has also larger variances for all three main components of the posted wage differentials, namely, the job effect, firm effect, and firm-job sorting, and

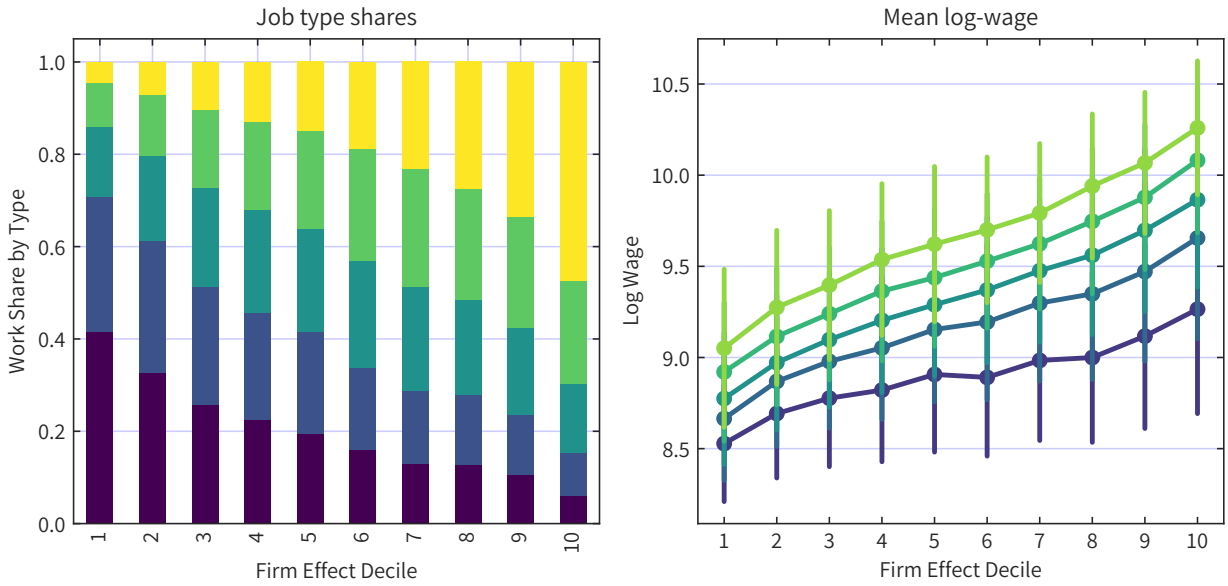
[50] This results is opposite to the results in Bonhomme et al. (2019), where the authors find no evidence of complementarity for most worker classes except the worst worker class which shows some stronger complementarity between worker type and firm type. However it is not straightforward to compare their results to ours as given the limited scope of the labor market in our data, our worst job class does not necessarily matches with the worst worker class in their data, and in our data there is way stronger variations across firm types for a certain type of worker.

Figure 4: Estimated Wage Components Using Job and Firm Clusters



Notes. The baseline values are from the results in Section 6.2. The methods to generate the job and firm clusters are described in the text.

Figure 5: Job Composition and Mean Wage By Firm Clusters



Notes. To obtain the firm effect deciles and five job types in the figure, we first use Equation (8) and Equation (10) to identify ten firm clusters and ten job clusters, and then use them to estimate the posted wage regression. We then use the estimated results to rank the ten firm clusters to be the ten firm effect decile and further bundle the estimated ten job clusters into five job types. The vertical lines in the mean log-wage figure plot the standard deviation of the posted log wage for the firm-job pairs.

thus generating a substantially larger overall wage variances than the lower skilled occupation. However, the limited number of the occupations in this comparison prevents any useful statistical inference. In this subsection, we examine how the composition of the posted wage dispersion differs across different occupations or types of jobs by classifying our job vacancies into more granular occupations.

In particular, we use two methods to assign the job vacancies into more specific occupations or job types. The first method is to match the job vacancies to the 5- or 6-digit occupation categories in the U.S. Standard Occupational Classification (SOC) 2018. This assignment, which is described in Appendix A.3, uses an algorithm that mixes a simple dictionary method and an supervised classification method, and eventually gives us 37 minor occupations in our entire data. The second method is to reply on the job clustering algorithm introduced in Section 7.2. In practice, we generates 320 job clusters, which can be regarded as 320 highly granular occupations. The posted wage components of the minor occupations generated by the first method are estimated through the baseline specification in Section 6, and the posted wage components of the job clusters generated by the second method are estimated through the shortcut way in Section 7.2. Figure 6 show the results for these two methods, where x-axis for each occupation is its mean wage. We find very similar patterns of the composition of the posted wage variance under both ways of occupation classification. Specifically, all three main components of posted wage variance are increased in occupations with higher posted wages, which presumably means high skilled occupations. The positive correlation is most significantly for the job effect, i.e. the job skill and task differences, but less steep for the firm effect and the firm-job sorting. When we instead check the shares of these components, the job effect and firm-job sorting still increase in the occupation mean-wage, but the wage variance share due to firm effect now decreases in the skill level, indicating a less significant positive correlation comparing to other two effects. Also we note that there are non monotonicity in the firm-job effect in that the positive correlation is actually as strong as the job effect for the occupations below 90 percentile, and that the occupations above 90 percentile somehow show a negative relationship. We suggest that these positive correlations further confirm our main findings in Section 6 that high skilled occupations with high posted wage dispersions are largely due to the facts that these occupations have more specific skills which generate not only more job differences but also more assortative matching between job qualities and firm wage policies.

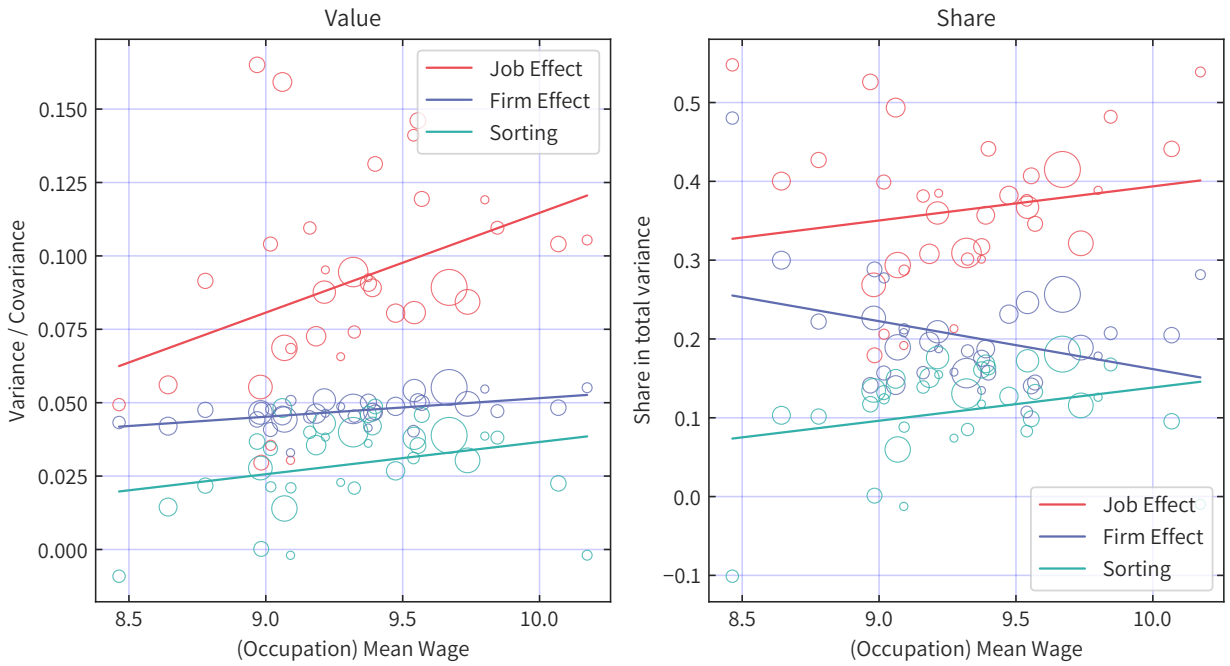
7.4 Posted Wage Inequality Trend

While in above all analyses we consider our data as cross-sectional, in our final analysis we examine the trend of the posted wage inequality in our Chinese job vacancy data, along with the potential drivers of any observed trend. The analysis here will be more tentative given the limitation of the data in the chronological sense: the trend of wage inequality might involve the sample changes over time during the development of the job board. However, we still believe that there are some informative insights that can be drawn since this sub-labor-market is probably the entire job market that certain types of job seekers encounter and is subject to rapid technological changes over the observation period.

In order to alleviate the impact of sample size on the estimation results, we split our entire observation period into three sub-periods (2014-2016, 2017-2018, 2019-2020) with roughly

Figure 6: Posted Wage Components Across Occupations

(a) Occupation as SOC Minor Occupations

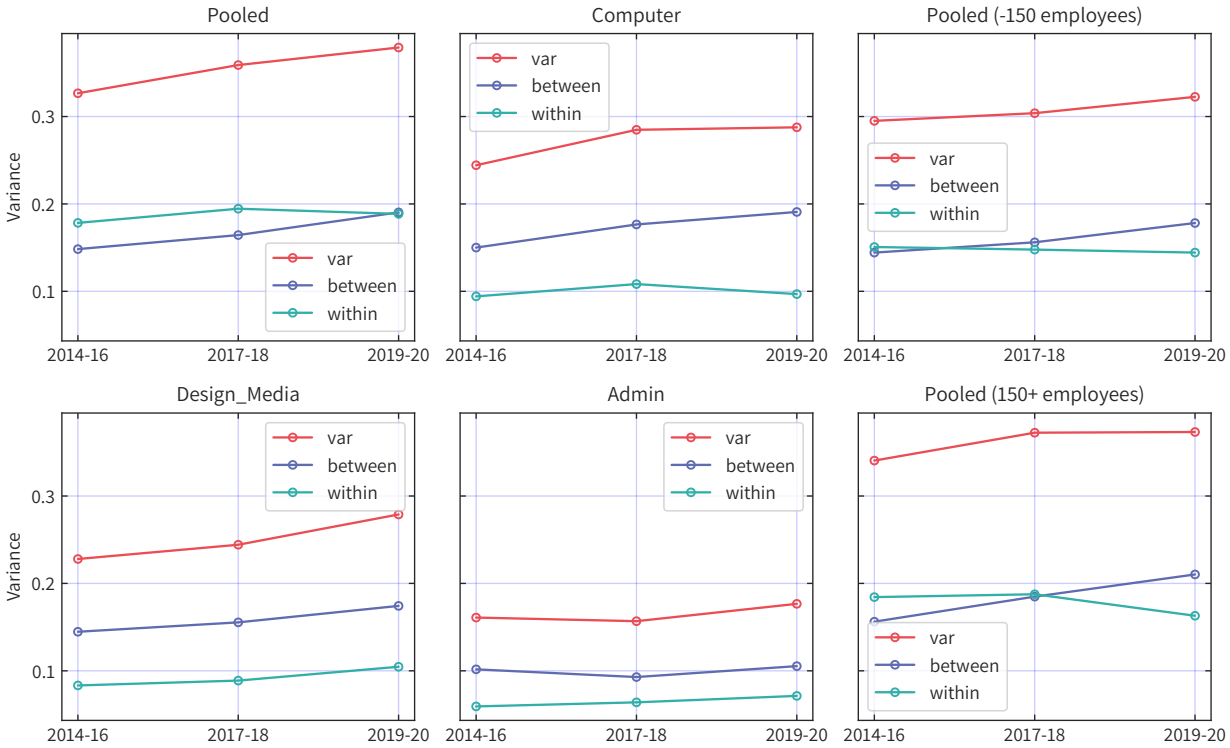


(b) Occupation as Job Clusters

Notes. The figures depicts the three main components of the posted wage dispersions across different occupations along with their mean wage levels. In the Panel (a), the occupations are 37 5- or 6-digit occupations in the U.S. Standard Occupational Classification (SOC) 2018, which are assigned to the job vacancies in our data using the method in Appendix A.3. In the Panel (b), the occupations are 320 job clusters that are classified from the our Pooled data using the method in Section 7.2. The size of the burbles represents the number of the job vacancies in each occupation. The lines show the linear regression models fitting the data for each main component.

similar number of job vacancies. We plot the overall posted wage variance for these three periods in Figure 7, where we also plot a simple decomposition of the between- and within-firm components. To resolving the issue of changing sample composition across periods, in addition to the Pooled sample, we also check the trends for the three selected major occupations as well as two sub-samples of firms with above or below 150 employed workers. Although the exact timing and extent vary across different samples, in all samples we see an increase in the posted wage variance over time, and for most samples this is mainly from an increase in the between-firm posted wage variance. Hence, there are increased posted wage inequality in our data, and this inequality is largely because the pay levels across firms diverge over time.

Figure 7: Posted Wage Variance Trends



Notes. ...

To identify the deep drivers of this increase in posted wage inequality, we then estimate our baseline model in Section 6 for these three sub-periods respectively. The results are shown in Table 7. While the variances of all three main components of the posted wage dispersion increase over time, their extents vary significantly. In particular, comparing between the first period to the last period, the variance of job effect and firm effect increased by about 0.1, while the variance of firm-job sorting increased by more than 0.3, doubling the value and accounting for two-third of the entire increase in the posted wage variance. Therefore, the most important driver of the increased posted wage inequality in our data is not more different jobs or more divergent firm pay policies, though they did occur, but increased sorting between high policy firms and high quality job posts. We then further take advantage of our framework and

decompose the source of this increased sorting between firms and jobs. The panel B shows that two-thirds of increased sorting is from the extensive margin of job effect and the other one-third is from the intensive margin. Moreover, panel C makes it clear that the main drivers of the increased sorting due to the extensive margin are those those specific skills and tasks, though there is also some increase from the medium-specific education-related skills and tasks. As a result, our examination here suggests that specific skills and tasks are not only the most important components of job heterogeneity that account for the posted wage inequality, but also a major contributor of the increased inequality in the posted wage in our data.

Table 7: Posted Wage Variance Decomposition By Periods

	2014-2016		2017-2018		2019-2020	
	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.326	-	.357	-	.377	-
Panel A: $X = \{\textbf{EDU}, \textbf{EXP}, \Xi_2, \ldots, \Xi_8\}$						
$\text{Var}(\theta_i)$	.149	.455	.163	.457	.157	.417
$\text{Var}(\epsilon_i)$	.096	.294	.092	.258	.094	.249
$\text{Var}(\psi_j)$	.048	.148	.050	.141	.059	.157
$2\,\text{Cov}(\theta_i, \psi_j)$	.033	.103	.051	.144	.067	.177
Panel B: Decompose θ **Terms**						
$\text{Var}(X_{int})$	.039	.121	.043	.120	.041	.109
$\text{Var}(X_{ext})$	.069	.212	.071	.198	.068	.180
$2\,\text{Cov}(X_{int}, X_{ext})$	.040	.123	.049	.139	.048	.128
$2\,\text{Cov}(X_{int}, \psi_j)$	.011	.035	.018	.051	.022	.059
$2\,\text{Cov}(X_{ext}, \psi_j)$	.022	.067	.033	.093	.044	.118
Panel C: Further Decompose X_{ext} **Terms**						
$\text{Var}(\Xi_g)$	.001	.003	.001	.002	.001	.002
$\text{Var}(\Xi_m)$	.005	.016	.006	.017	.006	.015
$\text{Var}(\Xi_s)$	.039	.120	.039	.109	.037	.098
$2\,\text{Cov}(\Xi_g, \Xi_m)$	.002	.006	.002	.005	.002	.004
$2\,\text{Cov}(\Xi_g, \Xi_s)$	.007	.021	.006	.016	.006	.015
$2\,\text{Cov}(\Xi_m, \Xi_s)$	.015	.046	.018	.049	.017	.045
$2\,\text{Cov}(\Xi_g, X_{int})$	.004	.011	.004	.010	.004	.010
$2\,\text{Cov}(\Xi_m, X_{int})$	.009	.027	.011	.032	.011	.028
$2\,\text{Cov}(\Xi_s, X_{int})$	.028	.085	.034	.096	.034	.090
$2\,\text{Cov}(\Xi_g, \psi_j)$	.002	.005	.002	.006	.003	.008
$2\,\text{Cov}(\Xi_m, \psi_j)$	.007	.020	.010	.027	.011	.030
$2\,\text{Cov}(\Xi_s, \psi_j)$	.014	.043	.022	.060	.030	.080
Obs	930149		1494468		1565866	
Firm	41750		62907		53662	

Notes. The results are derived by estimating our baseline specification on the pooled samples of different periods. See the notes in Table 4.

8 Conclusion

In this paper we develop a new method to study the components of wage inequality in the labor market. This method relies on vacancy data and machine learning algorithms and can work as an alternative to the popular method in the literature which uses two-way fixed effects and employer-employee panel data. Applying the method to the vacancy data of a Chinese job board, we find that at least in this high-end labor submarket in China, the compositions of posted wage inequality is consistent with other findings in the labor markets of the U.S. and European countries. More importantly during the analysis process, we unmask the most granular details of job characteristics and find a data-driven skill and task structure featured by different levels of specificity. We find that those occupational specific skills and tasks are the most important part of job heterogeneities that can account for the posted wage inequalities and especially the sorting between firms and jobs or workers.

There are two caveats on our approach and results that are worth mentioning. First, as we have argued earlier, online vacancy data does not cover the entire labor market. A typical online vacancy data is inclined to those young, educated, and internet-related jobs and workers. Firms may not post all their jobs on the internet and the vacancy posting frequency could be potentially different from real job compositions within the firm. Also, the posted wages are always the entry wage and lack the information of within-firm wage changes, wage bargaining and other firm-level wage determinants. To what extent do these issues matter is an empirical question worth future investigation. The second caution is that throughout our analysis we examine the wage inequality in monthly pay rather than in an efficient unit level of hourly wage. In fact in most cases precise information about working hour is not available in the online vacancy data and thus such examination is prohibited. One might suggest that this would result overestimated labor market inequality if higher posted wages are in fact fully compensated by the difference in different working hours. Here we argue three points that could potentially alleviate this concern. First, there will often be additional wage for overtime work that are not accounted in the posted wage. Second, the variations in working hours are rather limited comparing to the variations in posted wages. Finally, labor market inequality is often more reasonable to be considered on the total compensation level because firms are likely to provide wage and working-time as an indivisible package.

In terms of the future work, one important task is to validate to what extent are the results of our new approach be consistent or different from the results of using administrative employer-employee data and AKM approach. One straightforward way to test for this is to find a country with both types of data to be available and then conduct both analysis and compare the results. Also given the fact that our online vacancy data used in this paper is limited to a labor submarket rife with IT-related firms, we expect to see if the similar results on the compositions of posted wage inequality can be obtained when applying to the vacancy data of other labor markets, though in those cases some adaptions and adjustments in the practical details of machine learning algorithms might be necessary.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Andrews, M. J., L. Gill, T. Schank, and R. Upward (2008). High wage workers and low wage firms: negative assortative matching or limited mobility bias? *Journal of the Royal Statistical Society: Series A (Statistics in Society) 171*(3), 673–697.

Atalay, E., P. Phongthiengtham, S. Sotelo, and D. Tannenbaum (2020). The evolution of work in the united states. *American Economic Journal: Applied Economics 12*(2), 1–34.

Autor, D. H. and M. J. Handel (2013). Putting tasks to the test: Human capital, job tasks, and wages. *Journal of labor Economics 31*(S1), S59–S96.

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

Barth, E., A. Bryson, J. C. Davis, and R. Freeman (2016). It's where you work: Increases in the dispersion of earnings across establishments and individuals in the united states. *Journal of Labor Economics 34*(S2), S67–S97.

Becker, G. S. (1964). *Human Capital*. University of Chicago Press.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Bloesch, J., B. Larsen, and B. Taska (2021). Which workers earn more at productive firms? position specific skills and individual worker hold-up power. *Working Paper*.

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (2020). How much should we trust estimates of firm effects and worker sorting? Technical report, National Bureau of Economic Research.

Bonhomme, S., T. Lamadon, and E. Manresa (2019). A distributional framework for matched employer employee data. *Econometrica 87*(3), 699–739.

Braxton, J. C. and B. Taska (2020). Technological change and the consequences of job loss. *Forthcoming, American Economic Review*.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics 36*(S1), S13–S70.

Card, D., A. R. Cardoso, and P. Kline (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly journal of economics 131*(2), 633–686.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics 128*(3), 967–1015.

Dauth, W., S. Findeisen, E. Moretti, and J. Suedekum (2022). Matching in cities. *Journal of the European Economic Association 20*(4), 1478–1521.

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

Deming, D. J. and K. Noray (2020). Earnings dynamics, changing job skills, and stem careers. *The Quarterly Journal of Economics 135*(4), 1965–2005.

Di Addario, S., P. Kline, R. Saggio, and M. Sølvsten (2022). It ain't where you're from, it's where you're at: hiring origins, firm heterogeneity, and wages. *Journal of Econometrics*.

Frank, M. R., D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming, M. Feldman, M. Groh, J. Lobo, E. Moro, et al. (2019). Toward understanding the impact of artificial

intelligence on labor. *Proceedings of the National Academy of Sciences 116*(14), 6531–6539.
Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.
Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica 87*(4), 1307–1340.
Hershbein, B. and L. B. Kahn (2018). Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review 108*(7), 1737–72.
Hou, S. and L. Milsom (2021). The butcher, the brewer, or the baker: The role of occupations in explaining wage inequality.
Katz, L. F. (1986). Efficiency wage theories: A partial evaluation. *NBER macroeconomics annual 1*, 235–276.
Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.
Krueger, A. B. and L. H. Summers (1988). Efficiency wages and the inter-industry wage structure. *Econometrica: Journal of the Econometric Society*, 259–293.
Kuhn, P. and K. Shen (2013). Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics 128*(1), 287–336.
Lachowska, M., A. Mas, R. Saggio, and S. A. Woodbury (2022). Wage posting or wage bargaining? a test using dual jobholders. *Journal of Labor Economics 40*(S1), S469–S493.
Lise, J. and F. Postel-Vinay (2020). Multidimensional skills, sorting, and human capital accumulation. *American Economic Review 110*(8), 2328–76.
Marinescu, I. and R. Wolthoff (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics 38*(2), 535–568.
Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy 66*(4), 281–302.
Mortensen, D. (2005). *Wage dispersion: why are similar workers paid differently?* MIT press.
Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.
Sanders, C. and C. Taber (2012). Life-cycle wage growth and heterogeneous human capital. *Annu. Rev. Econ. 4*(1), 399–425.
Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of economic literature 31*(2), 831–880.
Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. Von Wachter (2019). Firming up inequality. *The Quarterly journal of economics 134*(1), 1–50.
Spitz-Oener, A. (2006). Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of labor economics 24*(2), 235–270.
Torres, S., P. Portugal, J. T. Addison, and P. Guimaraes (2018). The sources of wage variation and the direction of assortative matching: Evidence from a three-way high-dimensional fixed effects regression model. *Labour Economics 54*, 47–60.
Yamaguchi, S. (2012). Tasks and heterogeneous human capital. *Journal of Labor Economics 30*(1), 1–53.
Zhu, X. (2022). Post compensation inequality. *Working Paper*.

Appendices

A Data Collection And Processing

A.1 Data Collection

We set up a scraper which scraped all the vacancy data from the website of lagou.com in 2020. Because each vacancy that has been posted in the lagou.com website is attached with a unique ID, we were able to access to the information of the historical vacancies. Given the fact that at the end of the year 2020 new vacancy posts are typically assigned with an ID slightly larger than 8,000,000, we set up our scraper to try scraping all the vacancies with an ID between 0 and 8,000,000.

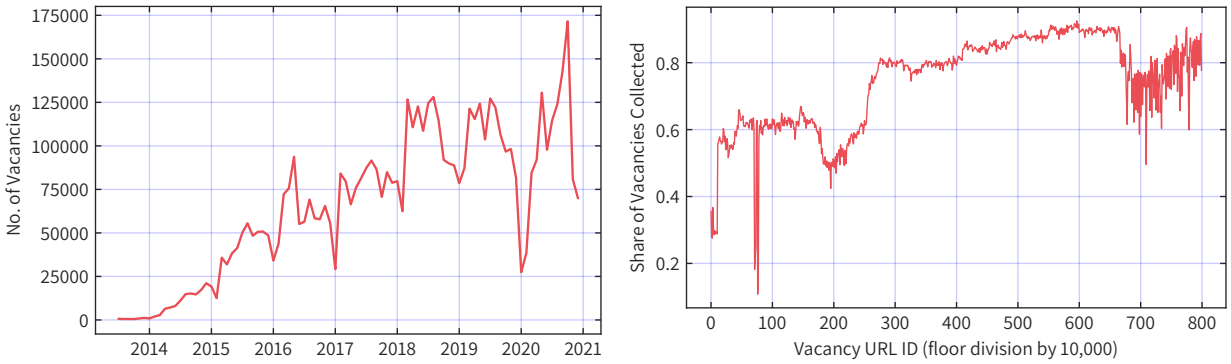
Despite a part of the vacancies that had been deleted from the website at the time our scraper accessed, we successfully collect a majority (about 75%) of all the historical vacancies that were still observable. Figure A1b plots the share of successfully collected vacancies for each 10,000 chunks of the total 8 million vacancies ordered by the ID. It shows that in general we scraped a consistent share of vacancies across all the IDs. In particular, for the vacancies ID between 0 and 3,000,000, we collect over 60% of all the vacancies and for the vacancies ID between 3,000,000 and 8,000,000 we collect over 80% of the vacancies, and within the unsuccessfully corrected vacancies over 20% are invalid vacancies that we have removed when scraping the data. Although we have no information on the deleted vacancies, we think those are more likely to be invalid or repeated vacancies and does not systematically bias any main results in our paper.

While the vacancy ID is only roughly correlated with time of posting, we can directly observe the posted time for each vacancy along with other information. Figure A1a plots the time trend of the monthly number of posted vacancies that we successfully collect. The monthly amount of vacancies increase over time which represents the growing popularity of the website. In particular, the average monthly amount of vacancies collected in year 2014 is about 12,000, and it grows to around 70,000 in 2016 and 2017, and around 100,000 between 2018 and 2020. Within a typical year, the number of posted vacancies is higher in the first half of the year and plummets in the end of the year, and this trend is consistent with other Chinese job vacancy data that target more general labor market (see e.g. He et al. (2021)).

A.2 Occupation, Sampling, and Statistics

Occupation. One empirical problem in our data is that there are no ready-for-use occupation categories for the job vacancies, as like many other online vacancy data. Although our machine learning method introduced in Section 5 does not rely on using occupation dummies, pre-classified occupations will help us to conduct our analysis on the board occupational level. Moreover, we argue that the procedure of occupation classification, under whatever methods, actually shares the same objective with our main approach, namely mapping individual jobs from a high-dimensional skill and task space to a low dimensional space. However, unlike

Figure A1: Trends on Collected Vacancies



(a) # of Vacancies by Posted Month **(b)** % of Vacancies by Post ID Chunks

Notes. The sample here is the collected vacancies removing about 15% invalid posts that have either signals of being test posts, abnormal wages, lack of key information, or too less content in job descriptions. This sample is further trimmed to obtain the sample used in the analysis conducted in the main text.

our data-driven approach, occupation classification relies on pre-specified rules to determine a bunch of subsets in the skill and task space, and thus ignore any within-occupation skill and task variations.

Here we briefly explain some key points of our original method of occupation classification, which combines a dictionary approach and a supervised classification approach. The details of the procedures and the comparison between our approach and other alternative approaches used in the literature are described in Appendix A.3. Whether through human classifying or machine learning methods, the task of occupation classification is to learn some information about a job, in our case from the job title and job description of a vacancy, and then label it to one of the pre-determined set of occupation categories based on that information. Given that our data contains limited scope of jobs comparing to the whole labor market, we first reduce the target occupation categories to a set of 55 6-digit ("minor") occupation categories within 8 2- or 3-digit ("major") occupation categories in the U.S. Standard Occupational Classification (SOC) 2018.[51] Next we prepare a dictionary by selecting multiple keywords for each of those selected 6-digit occupations according to their occupation descriptions in the SOC. The rule here is to select specific phrases or compound phrases so that the chances that these keywords

[51] We use the U.S. SOC because there is no well-designed official occupation classification for the Chinese labor market and the Chinese IT industry closely follows the technological trend in the U.S. market. This reduction relies on some human inspection on the vacancy data and the official occupation classification and thus might be, to some degree, arbitrary in the occupation choices, but it can largely increase the accuracy of the occupation assignment under even very simple classification algorithms. Within this selected set, major occupations vary in the number of minor occupations selected. For the Computer occupation, we include all 6-digit occupations in the SOC, while for other major occupations, the selected 6-digit occupations are rather limited compared with the full lists in the SOC. Also in practice, we further add 8 more 2- or 3-digit major occupation with no detailed 6-digit minor occupations appended to form an "Other" major category which is used to help increase accuracy of the classification and the vacancies classified to this major category, which have a fairly small share in the whole data, will be removed from the final pooled sample.

Figure A2: A Sample Vacancy From ByteDance

Notes. The style of the web page changes over time and this is a screenshot token in 2020 December. Some contents of vacancies (the part of job tasks, requirements, and benefits in left white space) are not always tidy as we have shown in this sample.

appear in the non-targeted occupations due to the multiple meanings of natural language are low.[52] Perhaps not surprisingly, phrases selected following this rule are basically specific skills and task contents that only used in that specific minor occupation. During this procedure, we further combine some 6-digit occupations when it turns out hard to find exclusive keywords to distinguish these occupations, reducing the 55 6-digit occupations to 34 minor occupation categories. With the dictionary in hand, we then check for each vacancy to see if its job title and job description contains these keywords. If a vacancy is matched with only one minor occupation, we regard it as a success of our dictionary method, label it with that matched occupation, and assign it to the "training" sample. If a vacancy has no match or multiple matches, we regard it as a failure and assign it to the "unknown" sample. In the next step we use our "training" sample to train a Naive Bayes classifier, which takes the vectorized text of job titles and job descriptions of a vacancy as input to predict the probabilities that this vacancy belongs to each of the minor occupations. We then apply the trained classifier to the "unknown" sample and assign those vacancies with the most likely occupation predicted. Finally, we also apply the trained classifier back to our "training sample" to rectify the potential misalignment under my dictionary method.

In summary, our occupation classification approach uses terms of specific skills and tasks to first identify the correct occupations for a subgroup of vacancies, and then uses this subgroup to learn the occurrence probability of all skills and tasks terms (along with some other terms) of a vacancy conditional on the vacancy belonging to that occupation. In other words, our algorithm relies on the perspective that occupation categories are different bundles of skills required and tasks conducted on the job. In some sense, this way is even more natural than strictly sticking with the guidelines in the official classification documents because it directly follows a general understanding of various occupations on the labor market, where such understanding may vary across different firms and evolve over time.[53] However, as we have mentioned earlier, these occupation categories can only represent the differences between different centroids of the subsets in the skill and task space, i.e. between-occupation skill and task variations, and thus do not contain any information about within-occupation skill and task variations. We will show later in our analysis that although the occupation dummies generated here can account for a large part of the skill and task variations across different vacancies in our data, the full-scale skill and task variables generated by our approach in Section 5 make it clear that the within-occupation skill and task variation is also an important part for the posted wage variation.

Sampling. To remove invalid vacancies and to reduce measurement errors in the vacancy data, we first drop all vacancies that are not full-time jobs, have outlier wages, or have job descriptions with less than 20 words.[54] We also drop all the vacancies posted in the website

[52] We use the corresponding Chinese translation of the English phrases, which sometimes requires to transform those phrases to the Chinese terms that are specific to the Chinese labor market context.

[53] Spitz-Oener (2006) shows that the compositions and levels of the tasks indicated by the occupation actually change over time under technological or organizational changes, and there are large variations within the same occupation for different workers in different firms and positions.

[54] To be specific, we remove the vacancies with a wage lower bound larger than 100,000CNY or smaller than 1,000CNY, and the vacancies with a wage upper bound larger than 200,000CNY or smaller than 2,000CNY. The words in the job description are counted either as Chinese characters or English words. Given the large size of

launch year 2013 from our sample due to the fact that both the sample size and the share of successfully scraped vacancies are substantially smaller than later years. We further trim our sample by dropping the vacancies from firms that have less than 10 posts and from all the locations that have less than 1000 vacancies over the observation periods. This trimming removes firms and locations with limited samples and thus both reduces the potentially invalid posts and reduces the measurement errors in our data. But it also largely reduce the proportion of small firms and small cities in our sample, resulting the majority of the firms in our sample to be middle or large size firms in large cities. Moreover, we identify the duplicated vacancies that have exactly the same job descriptions and education and experience requirements, and only keep the one with the highest wage posted.[55] Finally, we also remove a small share of vacancies with only English job descriptions that are mainly posted by multinational firms in order to focus our textual analysis on Chinese.

Summary Statistics. Table A1 shows the summary statistics both for the pooled sample and for three selected major occupations. In total our final sample contains around 4 million posted vacancies from over 86 thousand firms. Under our occupation classification, this includes 33 percent vacancies in Computer occupations, 14 percent in Design & Media occupations, 29 percent in Business Operation occupations, 5 percent in Financial & Legal occupations, 11 percent in Sales occupations, and 7 percent in Administrative occupations. The numbers of firms that post vacancies in each major occupation are between 70 percent to 90 percent of the total number of firms in the pooled sample, except for Financial & Legal occupations (50 percent). In fact over 40 thousand firms in our data post vacancies in more than four major occupations, although on average firms have fewer vacancies posted in Sales and Admin occupations (5-8 vacancies) comparing to Computer and Business Operation occupations (14-17 vacancies). Hence, a majority of the firms in our sample post vacancies in multiple occupations, which allows us to study both the firm level pay differences and the potential pay differences across different occupations within the firm. Also, the average number of words in a vacancy is quite similar across different board occupations, suggesting that firms do not behave very differently on their information closure.

As we have explained earlier, the information on firm size and education and experience requirement shows that our vacancy data inclines to a young and high-end part of the labor market. Most firms in our data are middle to large sized, evenly distributed across four size categories: 15 to 50 employees, 50 to 150 employees, 150-500 employees and more than 500 employees. In comparison, firms with less than 15 employees accounts for only 3 percent, mainly due to our sample trimming strategy which cuts off all firms with less than 10 vacancy posts. The fact that firm size distributions are close across different occupations again suggests that we have the same set of firms that post jobs in different occupations. In terms of required education, among all the vacancies, 33 percent requires some college degree, 54 percent requires bachelor degree, 1 percent requires post-graduate degrees, and 12 percent has no

our dataset, our results are not sensitive to any of the thresholds selected here.

[55] We use this keeping strategy to avoid the case that firms post the original vacancy with wage too low to attract any fitted workers and have to repost the same vacancy but with a raised wage which is now more close to the market level. However, this strategy will also remove the case that the firm simply repost the same job with an inflated wage.

Table A1: Summary Statistics

	Pooled	**Major Occupation**					
	-	Computer	Design_ Media	Business_ Operations	Financial_ Legal	Sales	Admin
Vacancy #	3,999,005	1,330,001	561,236	1,162,404	214,661	452,771	277,932
- share	1.00	.33	.14	.29	.05	.11	.07
Avg # Words	108.91	104.26	103.05	115.60	110.69	120.31	95.09
Wage (1k CNY):							
- Mean	13.64	17.38	10.68	14.19	11.95	10.21	6.32
- SD	9.24	9.79	6.31	9.52	9.19	6.53	3.90
Firm:							
- #	86,330	67,369	68,092	78,244	41,285	58,847	59,016
- Avg Posts	46.32	19.74	8.24	14.86	5.20	7.69	4.71
- Median Posts	20.0	9.0	4.0	6.0	2.0	3.0	2.0
Firm Size (share):							
- -15	.03	.03	.05	.02	.02	.03	.03
- 15-50	.18	.17	.25	.16	.15	.19	.20
- 50-150	.23	.21	.26	.22	.22	.23	.26
- 150-500	.21	.21	.21	.22	.23	.20	.23
- 500-2000	.15	.16	.12	.16	.18	.15	.14
- 2000+	.20	.23	.11	.22	.21	.19	.13
Education (share):							
- Vocational College	.33	.24	.38	.29	.27	.51	.52
- Bachelor	.54	.66	.47	.61	.63	.22	.24
- Master/Doctor	.01	.02	.00	.01	.03	.00	.00
- Not Specified	.12	.08	.15	.09	.07	.27	.23
Experience (share):							
- 0	.22	.12	.21	.16	.25	.48	.50
- 1-3	.37	.33	.48	.37	.36	.31	.38
- 3-5	.31	.41	.25	.33	.26	.16	.10
- 5-10	.11	.14	.05	.14	.13	.05	.03

Notes. From the raw data, we drop all vacancies that fit either of the following conditions: not full-time jobs, having outlier wages, having job descriptions with fewer than 20 words, posted at year 2013, posted by firms with less than 10 posts, with work locations that have less than 1000 vacancies, and non-Chinese posts. The average number of words are the number of Chinese characters or English words in the job descriptions. The posted wage is calculated as the mean of the wage lower bound and wage upper bound documented in the vacancy. Vocational school in China means a 2- or 3-years college curriculum which focuses on vocational training comparing to academic training and does not offer Bachelor degree. Not specified education can have different meanings on different cases but generally would indicate a lower bound of education level down to high school or vocational college.

requirement on education.[56] The high requirement on education level is due to both the nature of online job market and the large demand on cognitive-intensive jobs in the IT-producing and IT-using industries. In terms of required experience, close to 70 percent of the vacancies require 1 to 5 year experience, 22 percent do not require any experience, and 11 percent require 5 to 10 years experience.

Different from vacancy text length or firm size, education and experience requirements and posted wage vary substantially across different major occupations. In particular, Computer occupations vacancies have the highest average posted wage at CNY 17.4 thousand per month.[57] In comparison, the average posted wage of administrative vacancies is only around one-third of this number, CNY 6.3 thousand. Monthly wage in other occupations locate in between CNY 10 thousand to 14 thousand. This difference in posted wage goes hand in hand with education and experience requirements. While over 60% of the vacancies in Computer, Business Operation, and Financial & Legal occupations require bachelor degree or graduate degree, only around 20% of Sales and Administrative vacancies require an undergraduate or above. Those occupations requiring a higher education level also more often require higher than three year experience, while those occupations requiring lower education levels usually require 0 or 1 to 3 years work experience. This may indicate potential complementarity between college education and on-the-job training or learning by doing, and, if training or learning on the job develops within-occupation skill variations, complementarity between formal education and specific skills or tasks required on the vacancies. Given this distinction in the posted wage and education and experience requirements, we thus select Computer occupations, Design & Media occupations, and Administrative occupations as the representative high-, middle- and low-level occupations and show their results in the following analysis. However, all of our qualitative results hold if we pick say Business Operation occupations as middle-level and/or Sales occupations as low-skill occupations.

A.3 Details on Occupation Classification

In this section we explain the choices and the methods we use to assign the major and minor occupation for all the vacancies in our data.

There are two major steps of classifying the occupations for any vacancy data. The first step is that we need to decide that to which occupation code and in which level do we match our vacancies. Here, we decide to match our vacancy data to the official Standard Occupational Classification (SOC) 2018 designed by U.S. Bureau of Labor Statistics for two reasons. First, the U.S. SOC category has been widely used and studied in the labor literature and equipped with detailed task and skill descriptions and variables that can be used to compare with our own measure (after taking average on the occupational level). Second, because our data mainly contains IT jobs and other jobs in IT firms in recent years, it requires a recently updated oc-

[56] No specified requirement on education can have different meanings depending on different cases. But in general this indicates that the firm will have a lower requirement on the formal education level than the normal case and in most of the cases this means the lower bound can go down to high school or vocational college degree.

[57] The average wage is calculated as the mean of the lower bound and higher bound of the posted wage range. This mean wage of Computer occupations translates to 31,600 US dollars annual earning by using a currency ratio of 1USD:6.6CNY and then multiplying with 12 (months), and is three times over the Chinese GDP per capita in 2020 (10,500USD).

cupation classification to obtain a good match. Due to the fact that Chinese IT market has been advanced fast in recent years and largely followed the technological and organizational innovation in the global leading US IT market, we think the SOC 2018 would be a good fit to our data here.[58]

The second choice in the first step is selecting the occupation classification level. Ideally we want to match with the finest occupation level in SOC, which is the 6-digit occupations, so that we can use it to form the most accurate control of the heterogeneous skills and tasks between different jobs. However, Turrell et al. (2019) documents a potential tradeoff between the accuracy and the granularity in applying machine learning algorithms to assign job vacancies to the occupations codes. In particular, they argue that if matching with too granular occupation classification, the machine learning algorithm that based on job information from job title and job description text would find it difficult to accurately assign vacancies to the correct occupation. As a result, we decide to classify to the 3-digit level of the U.K. SOC.[59] We suggest that this result is mainly due to two reasons. First, adding more granular occupations as matching targets adds the possibility of the repetition of keywords across occupations which represents different meanings, and thus increase the difficulty of classifying occupations based on the job texts by any machine learning algorithms that only consider the occurrences of the keywords. Second, at the most granular level, i.e. the 6-digit SOC, some occupation categories might not be well-defined and easily distinguished from other occupations even from a theoretical perspective— it might not easier even for the worker themselves to distinguish the similar occupation categories. This conceptual problem in occupation classification design is easy to understand in a multi-dimensional task framework, where occupations are defined as the different compositions of the multi-dimensional tasks. In such a framework, the most granular occupation is often defined as working on one specific task, or on an easy-to-recognized specific composition of tasks. However in many real world cases, the typical job that one works on can range within a set of composition of these specific tasks, and those who work on close shares of tasks would find it difficult to classify into each single one. One example is that while in the U.S. SOC 2010, the "15-1130 Software Developers and Programmers" is further divided by "15-1131 Computer Programmers", "15-1132 Software Developers, Applications", "15-1133 Software Developers, Systems Software", and "15-1134 Web Developers", the two items "15-1132 Software Developers, Applications" and "15-1133 Software Developers, Systems Software" are combined as "15-1252 Software Developers" in U.S. SOC 2018, probably due to the fact that these two occupations share very similar tasks. Considering these two problems and the feature of our data, in this paper we choose the occupation matching targets to be a limited set of SOC 6-digit occupations with some rearrangements to combining not well-defined occupations. In particular, rather than mapping to a whole set of all occupations in the SOC, we limit our target occupations to be six major occupations (Computer, Art & Design & Media, Business Operations, Financial & Legal & Educational, Sales, and Administrative Occupations) that constitute the bulk of our vacancy data and one other occupations that we use to

[58] In comparison, the official occupation classification in China is not open to public access and largely outdated comparing to the fast development in the Chinese labor market, especially for ICT industries.

[59] The commonly used U.S. vacancy data from Burning Glass Technology has their vacancies data equipped with an occupation classification at 6-digit level of U.S. SOC. However, they do not make their machine learning algorithms public and thus one cannot tell the accuracy of their occupation assignment.

classify any other occupations.[60] This limitation requires some preliminary check on what kind of the job the data contains, but it can significantly simplify the classification. For each major occupation, we select the relevant SOC 6-digit occupations to be the minor occupations and in some cases combining several SOC 6-digit occupations into one to make classification easy. Again the selection on the 6-digit occupations and the bundles requires the understanding of the data and is subject to potential bias. However, our machine learning algorithm introduced later would automatically refine any of these problems because in nature it will be a task-based classification. Finally, we add one new minor occupation, "product manager" into the major occupation "Business Operations", which appears to be a new occupation in our data but has no corresponding category in the 2018 SOC.[61] The set of all minor occupations are shown in Table A2.

After deciding the target for matching, the second task is to use the information in the job vacancies to match the most suitable occupations codes for each job vacancy. Prior literature (e.g. Turrell et al. (2019) and Atalay et al. (2020)) measure the similarity between each pair of a vacancy and an official occupation category and select the most similar pair as the assignment. To be specific, one typically first represent the texts of job title and job description in each vacancy and official classification documents as a numerical array and then calculate the cosine similarity between the arrays.[62] While this method is relatively simple and can be easily conducted for any vacancy data, the disadvantage of the method is that the texts of SOC occupation descriptions and sample job titles often is very limited and thus sometimes not contain enough information to distinguish different occupations. Also, these official descriptions are written by official analyst but not replacing the real words that will be used in the real job vacancies. These problems are especially severe in our case because the English description and job titles after translation is often not the similar Chinese words used in the Chinese labor market and thus does little help to distinguish the occupations of vacancies.

To overcome this problem, in this paper we use a simple dictionary method to select a

[60] By selecting these 6 major occupations, we do not include any management occupations (11-0000 Management Occupations in the SOC) although manager occupations can indicate skills and tasks important for wage determination. This is because the management occupations usually contain both some occupation-specific tasks and some general management tasks, which would usually dampen the accuracy of machine learning algorithm. This is also because often it's hard to tell the distinction between a management job and non-management job as there is no strict threshold of the share of management tasks beyond which a job will be recognized as a management job. Finally, the word manager or manage translated in Chinese is often used in non-management occupations and thus would likely to mislead the occupation classification. Note although we do not assign any vacancies to management occupations, we partly control its explanatory power on wage through our measure on experience. And eventually in our textual analysis on the job description, we would explicitly examine the importance of the management tasks and skills.

[61] This "product manager" is likely a new occupation that have been updated in the 2018 SOC. Actually the updating of SOC designs is lagged behind the real labor market, especially for the sectors with the rapid technological changes. For example, in US SOC 2018, "15-1253 Software Quality Assurance Analysts and Testers" and "15-1255 Web and Digital Interface Designers" have been newly added into "15-1250 Software and Web Developers, Programmers, and Testers", although these two occupations have been commonly recognized in the labor market years before 2018.

[62] The methods of transforming raw text to a numerical array usually includes bag-of-words (BoW), term frequency-inverse document frequency (tf-idf) and n-grams. For details of these methods one can refer to Gentzkow et al. (2019). Atalay et al. (2020) first runs a word embedding model, Word2Vec, to represent each words as a vector in a hidden feature space, and then add the vectors of all words in a vacancy to construct the vacancy-level array in the same latent feature space.

learning sample and then do supervised machine learning on this sample so that we can both classify the occupations for the remaining sample and refine the result from our simple dictionary matching. In particular, we construct a dictionary that for each minor occupation I prepare several exclusive words or phrases that are either job titles or specific skills or tasks in Chinese that correspond to the terms in the SOC documents. Then for each vacancy we check if its job title or job description contains these keywords or not. If there is only one match, we directly assign the matched minor occupation to this vacancy and classify it as the learning sample. If there is no match or multiple matches, we classify it as the unknown sample. Because our man-made dictionary is likely not perfect, we would likely to have wrongly assigned vacancies in our learning sample, but by restricting the keywords to be highly specific, we ensure that the majority of the learning sample is correctly assigned. Next we use bag-of-words (BoW) method to transform the job text of vacancies $\mathbf{D}$ to a matrix of token counts $\mathbf{C}$ and apply a naive Bayes (NB) classifier to our learning sample. Each vacancy is represented by a row in the token matrix, $\mathbf{c}_i$, and each entry in this row, $c_{ik}, k \in K$, means the counts of the occurrence of token k from the entire token vocabulary K in the vacancy i. The details of this construction of $\mathbf{C}$ can be found in Appendix B.1.

The NB classifier is the most common and simple supervised classification algorithm and works quite well in many real-world situations in spite of its over-simplified assumptions. It is a generic model that assumes hypothetical distributions that generates the data and thus following the Bayes' theorem the possibility of a vacancy belonging to a minor occupation o given its token vector $\mathbf{c}_i$ is

$$P(o \mid \mathbf{c}_i) = \frac{P(\mathbf{c}_i \mid o) P(o)}{P(\mathbf{c}_i)} = \frac{\prod_j P(c_{ik} \mid o) P(o)}{P(\mathbf{c}_i)}$$

, where the second equation is from the naive conditional independence assumption across tokens. The different naive Bayes classifiers differ by the assumptions on the distribution of $P(c_{ik} \mid o)$, and we follow the custom to use a multinomial version of NB classifier which is the typical one used in text classification. The probability $P(c_{ik} \mid o)$ can then be easily estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting

$$P(c_{ik} \mid o) = \frac{\sum_i c_{ik} + \alpha}{\sum_i \sum_k c_{ik} + \alpha K}$$

, where smoothing parameter α is often set to 1.

The estimated multinomial NB classifier is then used to classify the occupations for the unknown sample and also reapplied to classify the occupations for the learning sample. The latter process is done because in nature our classifier assigns the occupations by looking at how likely the tokens, which are mainly tasks and skills, occur given that it belongs to this occupation, and thus by applying the classifier back to our learning sample we can rectify the potential misassignment by the dictionary approach. This reassignment is shown in A3 from where we can see that most of the reassignments occur across the minor occupations within the major occupations. This means the confusing mainly exist across minor occupations because they share similar tasks and skills and indicate that our classifier works quit well.

We need to note that one might find our method not easily to be generalized to the whole

labor market. This is because one needs to select the keywords for the dictionary used in the first step to pick the learning sample and thus the whole procedure is not fully automated but involve applying human knowledge. Hence, our method is currently more suitable for vacancies data with limited amount of occupations so that the researchers can easily construct the dictionary. However, in general we think our strategy have the potentiality to be improved and applied to the more general labor market. In particular one might find a way to automate the procedure of finding the unique keywords in the dictionary from the official descriptions. Or one might find an alternative way to obtain the sample vacancies for each occupation. The core advantage of our strategy is that in the second step, we can use simple supervised machine learning algorithms to learn from our data in hand and then classify the rest of the vacancies. The learning sample need not be 100% correct because we can apply the supervised machine learning back to itself to rectify mistakes. By using the supervised machine learning algorithms, we think the accuracy of our strategy will be largely better than the alternative methods that use cosine similarity. And this method would be more solid for the case of matching non-English vacancy data to English occupation classifications, like our paper.

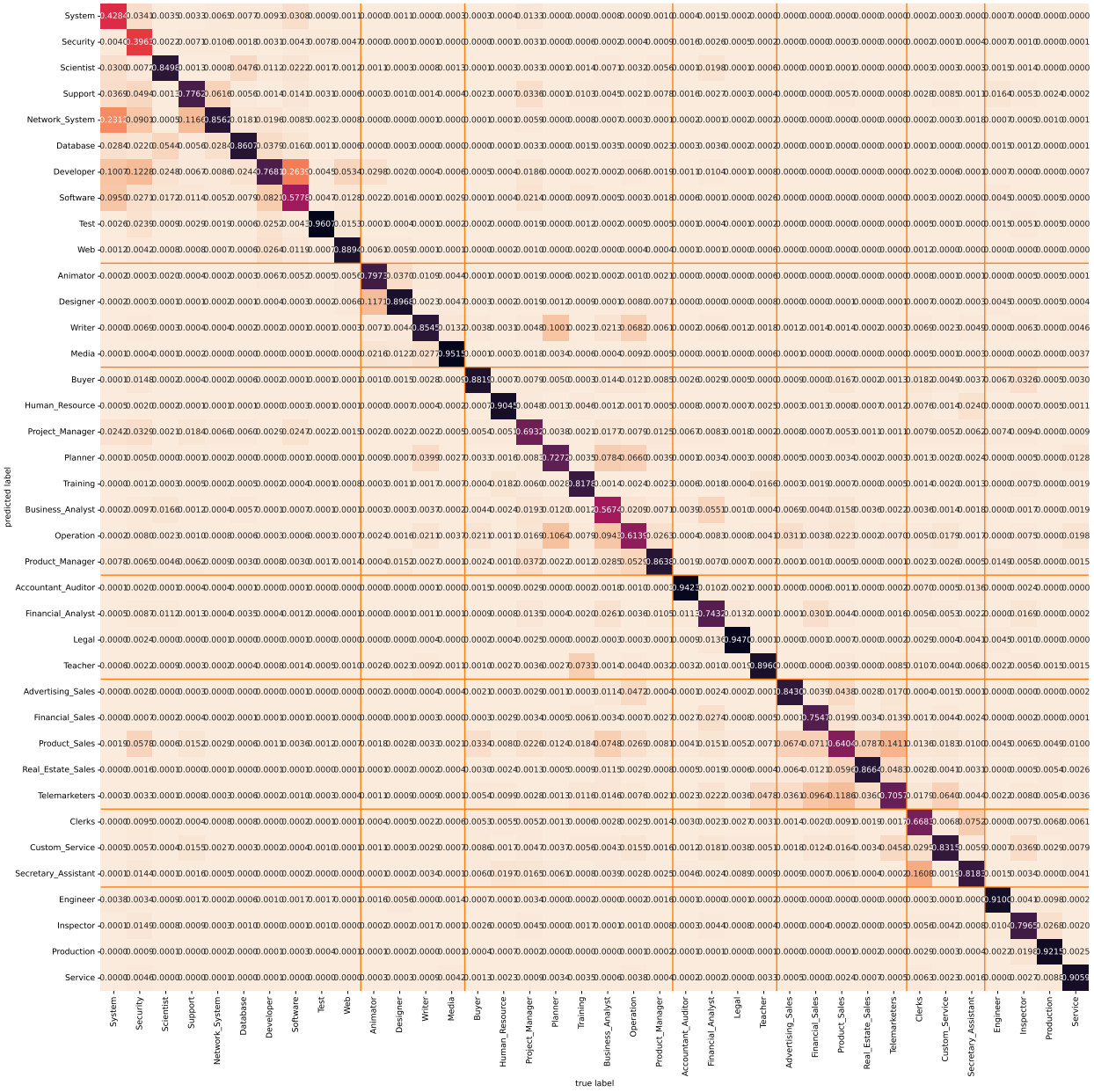
Table A2: Occupations And Keywords Selected

SOC Major	**SOC Minor (6-digit)**	**Keywords Used For Assignment (Translation from Chinese)**
15-1200 Computer Occupations	15-1211 Computer Systems Analysts	"Systems Analysis", "Systems Architect", "Systems Engineer"
	15-1212 Information Security Analysts	"Information Security", "Network Security", "System Security"
	15-1221 Computer and Information Research Scientists 15-2051 Data Scientists	"Data Mining", "Algorithm", "Machine Learning", "Deep Learning", "Image Processing", "Image Recognition", "Voice Recognition", "Computer Vision", "Natural Language Processing"
	15-1231 Computer Network Support Specialists 15-1232 Computer User Support Specialists	"IT Support", "Support Engineer", "Network Technician", "Network Support", "Pre-Sales Engineer", "After Sales Engineer"
	15-1241 Computer Network Architects 15-1244 Network and Computer Systems Administrators	"Network Engineering", "Network Architecture", "Network Management", "System Administration", "System Operations and Maintenance", "Operations and Maintenance Engineer"
	15-1242 Database Administrators 15-1243 Database Architects	"Data Engineer", "Data Architecture", "Database Engineering", "Database Architecture", "Database Administration", "Database Development"
	15-1251 Computer Programmers	"Development Engineer", "Programmer", "IT Engineer"
	15-1252 Software Developers	"Software Engineer", "Software Development", "Software Architect", "Application Development"
	15-1253 Software Quality Assurance Analysts and Testers	"Test Engineer"
	15-1254 Web Developers	"Frontend", "Web"
27-0000 Arts, Design, Entertainment, Sports, and Media Occupations	27-1013 Fine Artists, Including Painters, Sculptors, and Illustrators 27-1014 Special Effects Artists and Animators	"3D", "2D", "Original Painting", "Animation", "Painter", "Artwork", "Fine Art"
	27-1021 Commercial and Industrial Designers 27-1024 Graphic Designers	"Designer", "Graphic Design", "UI", "Drafting"
	27-3041 Editors 27-3043 Writers and Authors	"Editor", "Copywriter", "Editor", "Writer", "Lead Writer", "Screenwriter"
	27-4011 Audio and Video Technicians 27-4021 Photographers	"Photography", "Videography", "Editing", "Video Production"
13-1000 Business Operations	13-1022 Wholesale and Retail Buyers, Except Farm Products 13-1023 Purchasing Agents, Except Wholesale, Retail, and Farm Products	"Trade", "Import/Export", "Foreign Trade", "Purchasing", "Buyer"
	13-1071 Human Resources Specialists	"Personnel", "Human Resources", "HR"
	13-1081 Logisticians 13-1082 Project Management Specialists	"Project Management", "Process Management", "Logistics Management", "Logistics Planning"
	13-1121 Meeting, Convention, and Event Planners	"Event Planning", "Meeting Planning", "Event Operations"
	13-1151 Training and Development Specialists	"Trainer", "Training Instructor"
	13-1161 Market Research Analysts and Marketing Specialists	"Business Analysis", "Business Analysis", "Strategic Analysis", "Marketing Strategy", "Market Analysis"

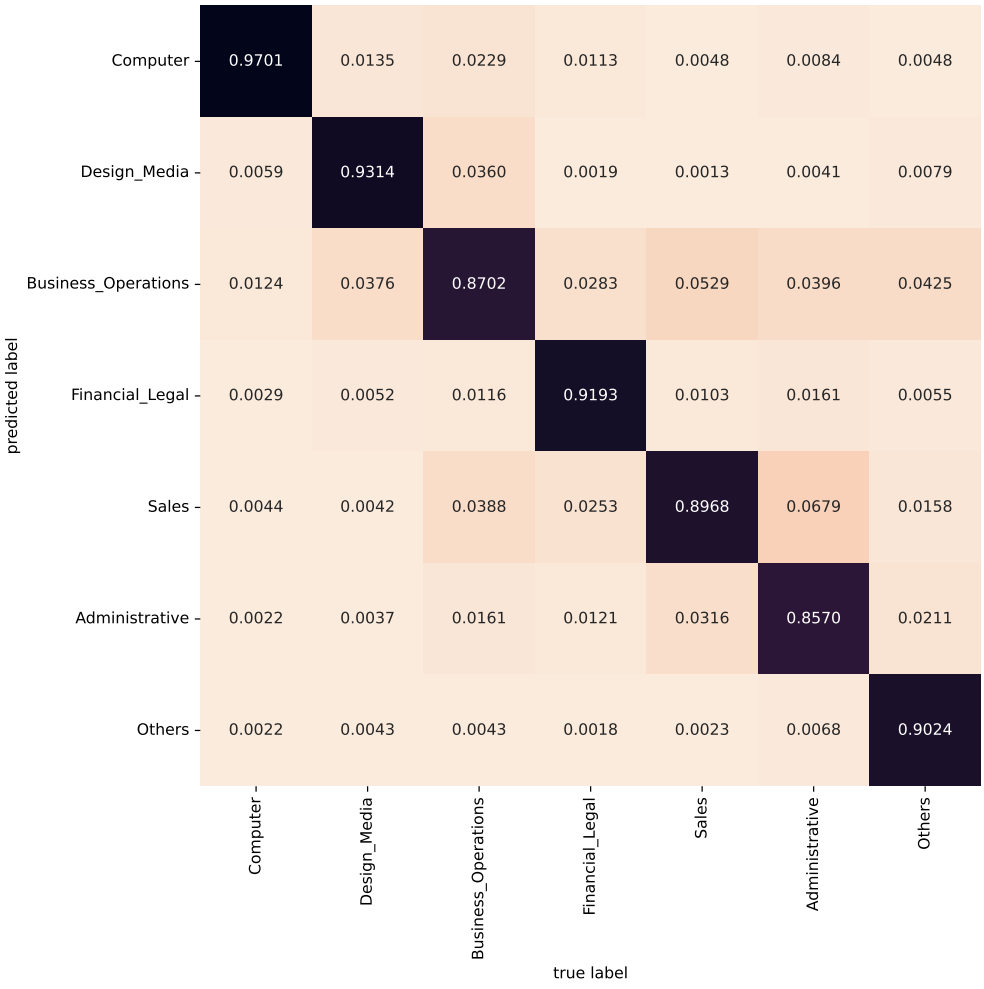
	13-1190 Miscellaneous Business Operations Specialists 13-1??? Advertising, Promotions, Marketing Specialists	"Product Operation", "User Operation", "Promotion Operation", "Advertising and Marketing"
	13-1??? Product Manager	"Product Manager", "Product Design", "Product Planning"
13-2000 Financial Specialists; 23-0000 Legal Occupations; 25-0000 Educational Instruction Occupations	13-2011 Accountants and Auditors	"Accounting", "Audit", "Finance", "Tax"
	13-2041 Credit Analysts 13-2051 Financial and Investment Analysts 13-2054 Financial Risk Specialists	"Credit Analysis", "Credit Assessment", "Risk Control", "Risk Management", "Investment Manager", "Investment Analysis", "Industry Research", "Industry Analysis", "Securities Analysis"
	23-1011 Lawyers 23-2011 Paralegals and Legal Assistants	"Lawyer", "Legal", "Law"
	25-2011 Preschool Teachers, Except Special Education 25-3011 Adult Basic Education, Adult Secondary Education, and English as a Second Language Instructors	"Teacher", "Assistant Teacher", "Teacher", "Kindergarten Teacher"
41-0000 Sales and Related Occupations	41-3011 Advertising Sales Agents	"Advertising Sales"
	41-3021 Insurance Sales Agents 41-3031 Securities, Commodities, and Financial Services Sales Agents 13-2052 Personal Financial Advisors	"Investment Advisor", "Financial Advisor", "Financial Manager", "Financial Planning", "Financial Sales", "Insurance Sales"
	41-4011 Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products 41-4012 Sales Representatives, Wholesale and Manufacturing, Except Technical and Scientific Products	"Sales Representative", "Account Representative", "Sales Specialist", "Commercial Specialist", "Channel Sales"
	41-9021 Real Estate Brokers 41-9022 Real Estate Sales Agents	"Real Estate Consultant", "Real Estate Agent", "Real Estate Agent", "Real Estate Sales", "Real Estate Sales"
	41-9041 Telemarketers Solicit donations or orders for goods or services over the telephone	"Telemarketing"
43-0000 Office and Administrative	43-4171 Receptionists and Information Clerks 43-9061 Office Clerks, General	"Clerk", "Receptionist"
	43-4051 Customer Service Representatives	"Customer Service"
	43-6011 Executive Secretaries and Executive Administrative Assistants 43-6014 Secretaries and Administrative Assistants, Except Legal, Medical, and Executive	"Secretarial", "Administrative", "Clerical"
Others (Droped in Analysis)	17-2000 Engineers 17-3000 Drafters, Engineering Technicians, and Mapping Technicians	"Mechanical Engineer", "Process Engineer", "Equipment Engineer"
	19-4000 Life, Physical, and Social Science Technicians	"Quality Inspection", "Quality Testing", "Environmental Testing", "Equipment Testing", "Food Testing", "Communication Testing", "Chemical Testing", "Non-Destructive Testing"
	51-0000 Production Occupations	"General Laborer", "Operator", "Welder"
	35-0000 Food Preparation and Serving Related Occupations 39-0000 Personal Care and Service Occupations 41-2000 Retail Sales Workers 53-0000 Transportation and Material Moving Occupations	"Receptionist", "Delivery Person", "Courier", "Rider", "Beautician", "Driver", "Cook", "Sales Clerk", "Salesman", "Swimmer", "Taster", "Anchor", "Florist"

Figure A3: Confusion Matrix of Occupation Assignment

(a) Minor Occupations



(b) Major Occupations



B Vectorization, Word Embedding, And Dimensional Reducing

B.1 Vectorization

In this section we explain our procedure of transforming the raw text of our vacancies **D** into the numerical token matrix **C** that are used in the machine learning algorithms. For all three machine learning methods that use **C**, namely the naive Bayes classifier for occupation classification, the lasso regression for feature selection, and the word embedding (Word2Vec) for feature clustering, there are several differences in the detailed choices but the general steps are exactly the same.

The first step is to select the individual documents $\{\mathbf{D}_i\}$ which is used to construct the individual numerical vector $\{\mathbf{c}_i\}$ (the rows of **C**). In the occupation classification and the lasso feature selection, an individual document is simply a vacancy. For the occupation classification, the $\mathbf{D}_i$ is the combined text of job title and job description. For the Lasso regression, the $\mathbf{D}_i$ is the combined text of job description and job benefits. In the word embedding model Word2Vec, the documents are not vacancies but the sentences in the job description and job benefits.

The second step is tokenization, i.e. breaking up raw text data **D** into short strings, and constructing the vocabulary set V, i.e. selecting the K standardized tokens (or in general features) that form the columns of **C**. In the textual analysis with English text the tokens are usually words obtained by splitting on spaces. But in Chinese a sentence is usually formed by multiple words which are present in a single sequence of characters without any spaces. To tokenizing the Chinese words, we use an open sourced Chinese tokenizer package, jieba.[63] The major advantage of jieba is that it is able to recognize Chinese compound words as well as to automatically tokenize both Chinese words and English words contained in one sentence.[64] We also add a list of IT words, education words, compensation words and etc. to the jieba tokenizer to enhance its performance. To reduce the dimension of the token/feature space, a lower bound of the occurrence of the tokens is often set to remove the words are too rare and do not convey much meaning. We set a lower bound of 10, so we only collect the tokens occurs over than 10 times. Also, after tokenizing the words from the text, a "stop words" list is often used to remove the words are very common and/or meaningless in the text. To do this, we use a commonly-used Chinese stop words list and in addition we use regular expression to remove all the tokens that are pure Chinese or English numbers or just one Chinese characters.[65] Finally, we remove all firms name from the segmented tokens because in the textual analysis firm names will be able to predict the posted wage through firm effects and thus disturb our examination on the skills, tasks and compensations. The remaining tokens then form the vocabulary V and

[63] Jieba is one of the most popular Chinese tokenizers that are open sourced. See the detailed information of its Chinese text segmentation functions in https://github.com/fxsjy/jieba.

[64] Because jieba does not automatically standardize the English words, we first lowercase all the English words before feeding our text to jieba tokenizer. We do not do any stemming or lemmatization for the English words because they are mostly technical words. There is no need to do stemming or lemmatization for the Chinese words because there is also no concept of a stem in Chinese at all.

[65] The common Chinese stop words list is taken from https://github.com/Alir3z4/python-stop-words. The numbers and single characters are removed because they often contain ambigious information.

thus all the features in **C**.

The final step is to select method of encoding for each entry $c_{i,k}$ in **C**. For the occupation classification, we use the most common way of encoding, bag-of-word, i.e. c_{ik} are the number of times token k occurs in document i, which is classical when using the multinomial naive Bayes classifiers. For the feature selection, we encode c_{ik} as an indicator of the presence of token k in document i, which is the simplest way to interpret the lasso regression. Using alternative methods like "term frequency-inverse document frequency" (tf-idf) would not affect our results qualitatively. Although these encoding methods are extremely simple and totally ignore the order of words that represents high-dimensional structure of the text, we find these simple methods are powerful enough to study the information embedded in the job texts.

B.2 Word Embedding

The model of word embedding with continuous bag-of-words (CBOW) architecture assumes the following process.

First, following exactly the same procedure as Appendix B.1, we construct the vocabulary set V from our raw vacancy documents **D** which contains K unique words or phrases.[66] Then we partition the entire corpus **D** into sentences, and each sentence is represented by a sequence of words denoted by $\{w_1, w_2, \ldots\}$, where each w_i is a word in V. Accordingly, we define a context of a word w_i in a certain sequence as a set of its adjacent words, $O = \{w_{i-m}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+m}\}$, i.e. a subset of $2m$ words in the same sequence that locate within a m-word window of w_i.

The basic idea of the estimation is to find two mapping U and W. The first function U maps any word w_i into a real vector in the hidden embedding space with pre-determined size H. The second function W maps the transferred vectors of a context, $U(O)$, to a conditional probability distribution: $\hat{P}\left(w_j \mid O\right) = W\left(U\left(w_{i-m}\right), \ldots, U(w_{i-1}), U(w_{i+1}), \ldots\right)$. And these two mapping are chosen by matching the estimated conditional probability with the conditional probabilities observed in the corpus through maximum likelihood procedure.

In practice, each word w_i is represented as a one-hot encoded vector, $\mathbf{x}_i \in \mathbb{R}^{|V|}$, i.e. an indicator vector of length K. Accordingly, a context is then denoted as $\{\mathbf{x}_{i-m}, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+m}\}$. Also, the two mapping is created as two matrices, input word matrix $\mathbf{U} \in \mathbb{R}^{H \times K}$ and output word matrix $\mathbf{W} \in \mathbb{R}^{K \times H}$. Although the **U** is still the mapping from word to the hidden embedding space, the **W** matrix here does not directly map $\mathbf{U}(O)$ to the conditional probability. In particular, we first calculate an averaged vector that represents the context in the latent layer, i.e.

$$\hat{\mathbf{u}} = \frac{\mathbf{U}(\mathbf{x}_{i-m}) + \ldots + \mathbf{U}(\mathbf{x}_{i-1}) + \mathbf{U}(\mathbf{x}_{i+1}) + \ldots + \mathbf{U}(\mathbf{x}_{i+m})}{2m} \in \mathbb{R}^H$$

. Then we use **W** to transfer this vector into a score vector $\hat{\mathbf{v}} = \mathbf{W}\hat{\mathbf{u}} \in \mathbb{R}^K$. Finally, we pass this score vector to a softmax operator to obtain the output vector $\hat{\mathbf{y}} \in \mathbb{R}^K$, with each element

[66] In general this vocabulary V need not be exactly the same as the one used for occupation classification or feature selection. But our selection and cleaning of the gathered tokens potentially also helps for the CBOW word embedding model so we use the same vocabulary here.

calculated by

$$\hat{y_k} = \frac{\exp(\hat{v}_k)}{\sum_{j=1}^{K} \exp(\hat{v}_j)}$$

.[67] This output vector $\hat{\mathbf{y}}$ is our estimation of the conditional probability distribution $\hat{P}(w \mid O)$, and $\mathbf{U}$ and $\mathbf{W}$ are found by maximizing the objective function $\sum_{k=1}^{K} \log(\hat{y}_k)$.[68]

In our computation we follow the literature to choose the primary parameters of our word embedding model. In particular, we set the window size of preceding and succeeding context words to be five ($m = 5$), and the dimension size of the hidden embedding space to be 100 ($H = 100$).

B.3 Dimension Reduction

Here we explain the procedure of the PLS dimension reduction. To easy notation, we denote our target variable log wage as $\mathbf{Y} \in \mathbb{R}^{N \times 1}$ and our predictive token matrix simply as $\mathbf{C} \in \mathbb{R}^{N \times K}$ (note in practice we go through each $\mathbf{C}'_{\mathbf{p}} \in \mathbb{R}^{N \times |V_p|}$). Our aim is to seek a representation of $\mathbf{C}$ in the lower dimensional space, $\boldsymbol{\Xi} \in \mathbb{R}^{N \times Q}$, where Q is the predetermined number of the components.

To obtain the first component, we first find a weight vector $\omega_1 \in \mathbb{R}^K$ that maximize the covariance between projected $\mathbf{C}$ and the target log wage $\mathbf{Y}$, $\mathrm{Cov}(\mathbf{C}\omega_1, \mathbf{Y})$. This can be achieved by finding the first left singular vectors of the cross-covariance matrix $\mathbf{C}^T\mathbf{Y}$, i.e. computing the singular value decomposition of $\mathbf{C}^T\mathbf{Y}$ and retain the singular vector with the biggest singular values. Then the first component is simply obtained as the projection $\xi_1 = \mathbf{C}\omega_1$. To calculate the second and following components, we take orthogonalization for both $\mathbf{C}$ and $\mathbf{Y}$ with respect to ξ_1, i.e. finding a loading vector $\gamma_1 \in \mathbb{R}^K$ and a loading value $\delta_1 \in \mathbb{R}$ that minimize the norm between $\xi_1\gamma_1^T$ and $\mathbf{C}$ and the norm between $\xi_1\delta_1$ and $\mathbf{Y}$ respectively, and replacing the original $\mathbf{C}$ and $\mathbf{Y}$ by the errors of their approximation respectively. We then take the orthogonalized value back to above procedure and iterate the whole process to obtain all remaining components $\xi_2, \ldots, \xi_Q$. In the end we gather all the components $\xi_1, \ldots, \xi_Q$ to form $\boldsymbol{\Xi}$ which is the demanded projection matrix of $\mathbf{C}$, and $\mathbf{C} = \boldsymbol{\Xi}\boldsymbol{\Gamma}^T + \mathbf{E}$ where $\boldsymbol{\Gamma}$ consists of the loading vectors $\gamma_1, \ldots, \gamma_Q$, and $\mathbf{E}$ are the error terms.

[67] This softmax function is equivalent to the multinomial logit model in discrete choice problems.

[68] In practice updating the two matrix and calculating the objective function is computational expensive due to the large size of the latent layer and thus a technique called negative-sampling is often used as a more efficient way of deriving word embeddings.

C Within- And Between-Group Posted Wage Differentials

In this section, we study the within and between posted wage differentials for two types of group: firm and occupation. The distinction of within- and between-firm wage differentials has be studied in the recent literature of wage dispersion using AKM approach and employer-employee data (Barth et al., 2016; Song et al., 2019). Here we show the results in our job vacancy data in China to further confirm that the results of wage dispersion estimated using job vacancy data are consistent with the results using dominant administrative data. We also estimate our baseline model with a specification where we only put education, experience, and minor occupation dummies into X and compare the results to the ones in our baseline model where the full controls of job skills and tasks are included, so that we can examine to what extent does the within-occupation skill and task heterogeneity account for the posted wage differentials.

Following Song et al. (2019), we can rewrite the variance decomposition in (2) into

$$\text{var}(\ln w_i) = \underbrace{\text{var}\left(\theta_i - \bar{\theta}_j\right) + \text{var}(\epsilon_i)}_{\text{Within-firm component}} + \underbrace{\text{var}\left(\bar{\theta}_j\right) + 2\,\text{cov}\left(\bar{\theta}_j, \psi_j\right) + \text{var}\left(\psi_j\right)}_{\text{Between-firm component}} \tag{11}$$

, so that the total wage variance is divided into within- and between-firm components. The within-firm component includes the variance of the deviation of each job's value from the firm average level, $\text{var}\left(\theta_i - \bar{\theta}_j\right)$, and the variance of wage residual, $\text{var}(\epsilon_i)$. The between-firm component contains the variance of the average valuation of each firm's jobs, $\text{var}\left(\bar{\theta}_i\right)$, along with the variance of firm pay premium and covariance of sorting between job and firm effects.[69] In other words, this further decomposition divides the job variance into with-firm and between-firm job parts. The results of this slightly more granular variance decomposition under three specifications that control for different sets of job characteristics are shown in Table C1. Specifically, in the specification of Panel A, we only include the education and experience in X, in the specification of Panel B, we also add our minor (5- or 6-digit) occupation dummies, and the specification of Panel C is exactly the same as the one of the baseline specification described in Section 6.1. All the results shown in Table C1 are plug-in estimates without any bias corrections, but as we have discussed in Section 6.4 that the finite sample bias would have no impact on the estimates of the job effect and job-firm sorting but only result overestimated the firm effect of limited extent when the numbers of jobs per firm in the sample are not too low.

Our main focus here is the pooled sample as we want to compare our results with the results in previous studies using national census data in the U.S. In our baseline specification where all detailed job skills and tasks are controlled (Panel C), the within-firm and between-firm wage variances account for 57 percent and 43 percent of the total wage, respectively, and the total 45 percent posted wage variance accounted by the job effect are now divided into 30 percent from within-firm part and 15 percent from between-firm part. These results are consistent to the results in Song et al. (2019) where the authors show that on the U.S. labor market between 2007 and 2013 (the period in their results closest to our observation period), total within-

[69] Note that $\text{Cov}\left(\bar{\theta}_i, \psi_j\right)$ here is exactly the same as $\text{Cov}\left(\theta_i, \psi_j\right)$ in (2) as ψ_j does not vary across jobs within firm.

Table C1: Posted Wage Variance Decomposition

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.360	-	.279	-	.251	-	.164	-
Panel A: X={EDU, EXP}								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
Within-Firm:								
$\text{Var}(\theta_i - \bar{\theta}_j)$	.072	.199	.037	.133	.036	.144	.033	.204
$\text{Var}(\epsilon_i)$	.132	.367	.089	.318	.078	.310	.061	.371
Between-Firm:								
$\text{Var}(\bar{\theta}_j)$	.030	.084	.015	.055	.017	.068	.017	.102
$\text{Var}(\psi_j)$	.076	.212	.102	.365	.086	.342	.041	.253
$2\,\text{Cov}(\bar{\theta}_j, \psi_j)$	.049	.137	.036	.130	.034	.136	.011	.069
Panel B: X={EDU, EXP, OCC}								
$\text{Var}(\theta_i)$	.146	.407	.065	.232	.061	.243	.052	.320
Within-Firm:								
$\text{Var}(\theta_i - \bar{\theta}_j)$	.103	.286	.049	.176	.040	.159	.035	.214
$\text{Var}(\epsilon_i)$	.101	.280	.077	.275	.074	.295	.059	.361
Between-Firm:								
$\text{Var}(\bar{\theta}_j)$	.044	.121	.016	.057	.021	.085	.017	.107
$\text{Var}(\psi_j)$	.064	.179	.096	.344	.079	.314	.040	.245
$2\,\text{Cov}(\bar{\theta}_j, \psi_j)$	.048	.134	.041	.148	.037	.148	.012	.074
Panel C: X={EDU, EXP, $\Xi_2, \ldots, \Xi_8$}								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.331	.067	.408
Within-Firm:								
$\text{Var}(\theta_i - \bar{\theta}_j)$	.108	.298	.055	.197	.050	.197	.044	.272
$\text{Var}(\epsilon_i)$	.096	.267	.071	.252	.065	.255	.050	.304
Between-Firm:								
$\text{Var}(\bar{\theta}_j)$	.055	.152	.027	.095	.034	.133	.022	.137
$\text{Var}(\psi_j)$	.051	.141	.074	.263	.062	.243	.035	.216
$2\,\text{Cov}(\bar{\theta}_j, \psi_j)$	.051	.142	.054	.193	.043	.171	.012	.072
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Notes. In the specification of panel A, X only contains education and experience dummies, while in the specification of panel B, X also includes minor (close to 6-digit) occupation dummies. The specification of panel C is exactly the same as the one for generating the baseline results in Table 4. The variance and covariance terms related to year dummies have been subtracted from the total variance of log wage, and thus the sum of all with-firm and between-firm components would be equal to the total variance of log wage. Job effect $\text{Var}(\theta_i)$ is the sum of with-firm job variations $\text{Var}\left(\theta_i - \bar{\theta}_j\right)$ and between firm job variations $\text{Var}\left(\bar{\theta}_j\right)$. For each major occupation sample estimated, we drop vacancies belong to the firms that have less than two vacancy posts in this major occupation. Different from the results in the main text, all estimated variances and covariances here are plug-in estimates without bias corrections. But as we have discussed in Section 6.4, the bias corrections will only have non-negligible impact on the estimated variance of the firm fixed effects when the sample numbers per firm is small.

firm variances account for 60 percent and total between-firm variances account for 40 percent of the overall wage variances, and within-firm worker effect accounts for 38 percent of the total wage variance while the between-firm worker effect accounts for 13 percent.[70] Another useful comparison is to compare our specification of only education and experience controls in Panel A with the results in Barth et al. (2016), where they are using another source of U.S. earning data and conduct an estimation without worker fixed effects but only controls for workers' schooling, potential experience, and some other demographics. The estimated wage variance share in Barth et al. (2016) in their last period 2007 is 12 percent due to within-firm worker differences and 6 percent due to between-firm worker differences. The corresponding shares in our specification in Panel A is 20 percent and 8 percent. The large accountability of education and experience in our data is perhaps because the experience documented in the job vacancy is a better measure of the expertise than the potential wage often used in the Mincer-style wage regression. Overall, we suggest that these comparisons indicate that the estimated between-firm and within-firm wage components from our job vacancy data in China are generally consistent to the results in the previous studies where administrative earning data in the U.S. is used.

Next, we discuss the differences in the results of different specifications. By examining the results across three specifications, it is easy to observe that with adding more granular job controls, the share of posted wage variances accounted by job effect increases and the share accounted by firm effect decrease. In particular, after adding 5- or 6-digit occupation dummies, the share of job effect increases from 28 percent to 41 percent, while the share of firm effect decreases from 21 percent to 18 percent. Adding detailed job skill and task controls further increases the job effect share to 45 percent and reduces the firm effect share to 14 percent. Two implications can be derived from these results. First, without correctly controlling for the unobserved worker or job characteristics, the importance of the differences in firm wage policies will be significantly overestimated. In other words, a large part of positive sorting between firm and job or worker are based on unobserved job or worker characteristics.[71] Second, comparing to the specification that control for granular occupation classifications, our baseline specification with full job skill and task controls show that there are important job heterogeneity even within detailed occupations that matter for wage differentials. In particular, the comparison between the three specifications in Table C1 suggests that the within-occupation skill and task variations account for more than one quarter of the job variations conditional on education and experience, and this number is likely to be a lower bound given that (i) our occupation classification are generated from skill and task clustering algorithms and thus probably more well-assigned than the occupation information in most datasets, (ii) that our dimensional reduction algorithms used to generate full job controls inevitable bring some in-

[70] Their estimation captures more wage variances accounted by within-firm variances and less by residual terms than our results, indicating that there could be more measurement error in our data or that the real wage dispersion in the labor market includes also a part of worker effect different from the job effect captured in our estimation. The latter possibility are for future studies in the cases where linked employer-employee-vacancy data is available.

[71] The similar overestimation can be found by comparing the results in Barth et al. (2016) and Song et al. (2019). The estimated firm effect share in Barth et al. (2016) with only controlling for education and potential experience is about 36 percent whereas the estimates in Song et al. (2019) using the AKM approach is less than 10 percent.

formation loss, and (iii) some part of the within-occupation skill and task variations might have already been explained by the education and experiences variables which work as a proxy of detailed skills and tasks. Moreover, the comparison between Panel B and Panel C show that the within-occupation job heterogeneity seems to be more important for between-firm job differences than for within-firm job differences, indicating that the within-occupation skill and task differences could be mainly a firm-level feature. Another interesting and related feature in Table C1 is that the estimates in individual occupations show more significant increase in the job effect shares comparing the pooled sample, which implies that within-occupation skill and task differences are more important for considering the wage differences in a given board occupation than the wage dispersions in the whole labor market.[72]

In summary, our examination here shows that the estimated within- and between-firm wage differentials from our Chinese job vacancy data are roughly consistent with the results in previous studies that use administrative employer-employee data in the U.S. Despite different types of data (posted job v.s. worker in administrative data) and different labor markets (China v.s. U.S.), the compositions of the wage inequality components illustrate surprisingly similarity and the estimated within-firm job or worker variances always doubles the estimated between-firm job or worker variances. In addition, we compare different specifications to quantify the importance of within-occupation skill and task variations and find that even after controlling for granular occupation categories, within-occupation job differences can explain about 5 percent of the posted wage variances, which is more than one third of the wage variations accounted by the between-occupation job differences. We suggest that this is only a preliminary quantitive exercise on the importance of the within-occupation skill and task variations and is likely to be a lower bound. It will be interesting to compare this result to the results of a similar investigation with administrative employer-employee data.

[72] Another way to interpret this result is that while the board occupation categories do a good job in separating different types of skills and tasks and thus can account a large amount of wage differentials, the more granular occupation categories are somehow not same useful in distinguish the skill and task differences that determine wage dispersions within board occupations.

D Connections To Deming & Kahn (2018)

Deming and Kahn (2018) is one of the pioneer works that use the online job vacancy data to study how different types of skills and tasks can play a role in wage inequality. In particular, they link the U.S. national job vacancy data from Burning Glass Technology to the U.S. labor census data at granular location-occupation level, and find that the heterogeneity in the demands for cognitive and social skills, which are proxied by the shares of vacancies documenting certain keywords related to those skills, have statistically significant predict power on wage differences across location-occupation cells. Their primary focus on the two types of skills, cognitive skills and social skills, is due to "their prominence in the literature linking technological change to wage inequality", and they claim that while other types of skills can have important explanatory power in wage regressions, they "do not have a general framework for analyzing them". Given that one of our main results is that those general skills, whether cognitive or social or noncognitive, hold relatively little impact in wage dispersions, and that our framework consider not human-selected skill categories but the entire set of skills and tasks documented in the job texts, one may wonder how does our result here get square with their result or how should we interpret these different findings in different methods. In this section, we examine these problems by replicating the analysis of Deming and Kahn (2018) in our data and discussing the potential issues existed in this type of studies.

In order to replicate the estimations in Deming and Kahn (2018), we first select the Chinese keywords that are corresponding to the English skill keywords used in Deming and Kahn (2018), and then use them to construct the indicator variables for cognitive skills and social skills of each job vacancy by checking if the job texts contain any words in the keyword list. The keywords we used are listed in Table D1, where we also show what exactly terms in our Lasso-selected vocabulary set contain those keywords. For cognitive skills, we find that many linked terms belong to specific skills (e.g. industry analysis) and relatively less to general skills (e.g. problem solving), while for social skills, we find more terms in general skill sets (e.g. communication) but only a few terms are specific (e.g. cooperation projects). This result indicates, perhaps intuitively, that cognitive skills are more likely to be specific while social skills are inclined to be general ones. There are two important notes here. First, while Deming and Kahn (2018) and many other recent studies in the literature of skills and tasks often either implicitly regard cognitive skills and social skills as general skills or avoid discussing the specificity at all, here we show that the specificity can be different across different types of skills. Thus only considering board and abstract skill category and using certain keywords to generating desired variables may risk in misleading on what are the these skills and on how workers acquire their skills and why firm demand different types of skills. Second, comparing to our Lasso-selected vocabulary, there are many terms in the job vacancy texts that are also related to cognitive and social skills but are omitted under the keyword approach (e.g. project plan or business negotiation). As a result, the keyword approach used in Deming and Kahn (2018) may be biased if firms with different wages also vary in how they use the terms of skills.

The results of regressing posted log wage on these two skill variables along with a bunch of different sets of controls are shown in Table D2. In column (1), where the regression controls for education, experience, and occupation dummies, we find the significant and positive relationship between wage and both cognitive and social skill indicator variables, and the size of the coefficients are quite close to the ones found in the baseline estimations in Deming and

Kahn (2018). This successful replication stands as an interesting robustness check that the main result in Table D2 is valid even when we use the vacancy data in a developing country and estimate the wage regression directly on the posted wage. However, when we further add the interaction of the cognitive skill and social skill into the regression, we find significant and negative coefficients for the interactive variable, which is inverse to the results in Deming and Kahn (2018). Therefore, somehow the cognitive skills and the social skills do not have a complementary relationship in our data. In the rest of the columns, we add our skill and task variables $\Xi_2, \ldots, \Xi_8$ complied from the entire job texts to investigate if controlling other skills and tasks in the job will change the results.[73] The results show that, while still being significantly positive, the coefficients of both the cognitive and social skill variables decreases substantially after controlling for other skills and tasks, and the defines are most significant when we adding the controls for those specific skills and tasks (Ξ_s). This indicates there are potential unobserved bias when in the estimation of Deming and Kahn (2018) although the bias may not change the results qualitatively . Also we find that the negative coefficients of the interactive variables become larger after controlling for all types of skills and tasks documented in the job.

To sum up, we follow the methods in Deming and Kahn (2018) to construct variables of cognitive skills and social skills, and our posted wage regressions show that these variables can predict posted wage differences, though adding the full controls on other skills and tasks documented in the job texts substantially reduce their predictive powers. However, despite this significantly positive correlation, the cluster of the general skills in our main analysis, which incorporates many of the keywords of cognitive and social skills here, turns out to be a rather unimportant driver of the total wage variations. Thus the perhaps most important takeaway here is that, given the high-dimension of skills and tasks embedded in the job text data, a significant correlation between the posted wage and certain terms in the job texts does not necessarily mean that those terms and the skills they represent are eventually important for wage determination and wage inequality.

[73] $\Xi_2, \ldots, \Xi_8$ are re-constructed after removing the keywords of cognitive and social skills used above from the job vacancy texts and thus different from the ones used in the main text.

Table D1: Keywords of Job Skills

Job Skills	Keywords and Phrases	
	Deming & Kahn (2018)	Chinese Correspondents
Cognitive	Problem solving, research, analytical, critical thinking, math, statistics	解决, 问题, 研究, 分析, 批判, 思考, 数学, 统计
Social	Communication, teamwork, collaboration, negotiation, presentation	交流, 沟通, 讨论, 演示, 展示, 合作, 团队, 协作
	Matched Keywords and Phrases in V'	
	V_g, V_e	$V_{s1}, \ldots, V_{s5}$
Cognitive	分析判断(analysis & judgment); 思考(reflections); 独立思考(independent thinking); 解决问题(problem solving); 数学(mathematics); 研究生(graduate students); 研究者(researchers); 统计学(statistics); 认真思考(think carefully)	统计(statistics); 统计分析(statistical analysis); 问题解答(question answers); 商业分析(business analysis); 行业研究(industry research); 业务分析(business analysis); 关键问题(key issues); 分析(analysis); 分析报告(analysis report); 功能分析(functional analysis); 可行性研究(feasibility study); 解决(solutions); 解决方案(solutions); 问题(question); 市场分析(market analysis); 数据分析(data analysis); 深入分析(in-depth analysis); 深入研究(in-depth research); 研究(research); 兼容性问题(compatibility issues); 定位问题(positioning issues); 疑难问题(difficult questions); 系统分析(system analysis); 面向对象分析(object-oriented analysis)
Social	交流(communication); 人际沟通(interpersonal communication); 协作(collaboration); 合作(cooperation); 团队(team); 团队精神(team spirit); 沟通(communication); 沟通交流(communication);学术交流(academic exchange)	合作项目(cooperation projects); 沟通了解(communication & understanding); 合作方(partners)

Table D2: Wage Regression With Skill Indicators (Pooled Sample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Cognitive	.045	.054	.027	.047	.013	.032	.011	.033
	(.000)	(.001)	(.000)	(.001)	(.000)	(.001)	(.000)	(.001)
Social	.035	.041	.030	.045	.020	.033	.025	.041
	(.001)	(.001)	(.001)	(.001)	(.000)	(.001)	(.001)	(.001)
Both required		-.012		-.026		-.024		-.029
		(.001)		(.001)		(.001)		(.001)
Ξ_g, Ξ_m			✓	✓			✓	✓
Ξ_s					✓	✓	✓	✓
Education FE	✓	✓	✓	✓	✓	✓	✓	✓
Experience FE	✓	✓	✓	✓	✓	✓	✓	✓
Occupation FE	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Adj. R^2	.582	.582	.604	.604	.636	.636	.641	.641

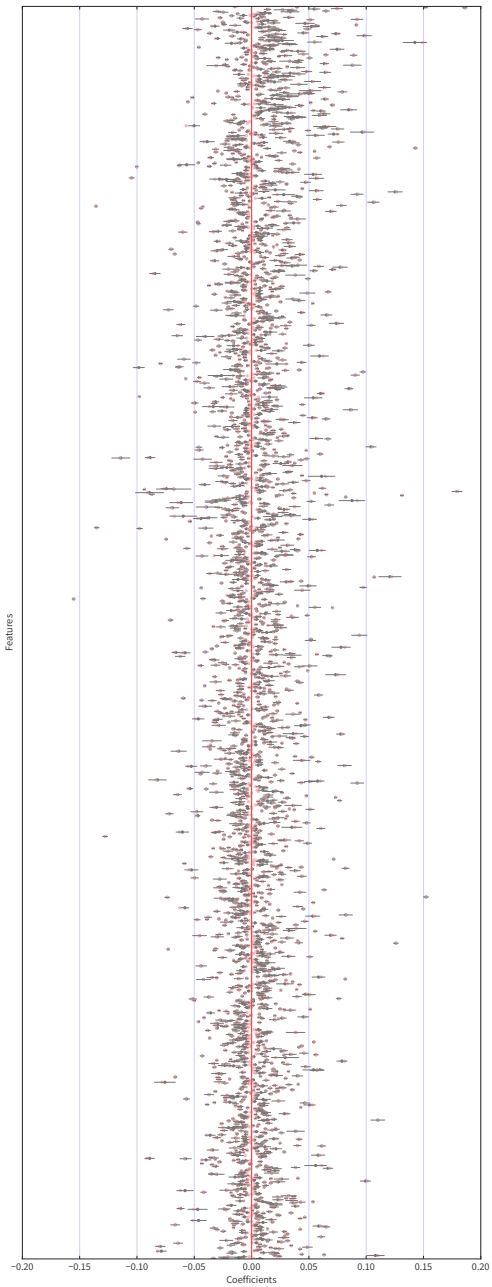
Notes. The construction of cognitive and social indicator variables use the corresponding Chinese keywords of the keywords in Deming and Kahn (2018), which is shown in Table D1. "Both required" is the interaction of these two variables. The Ξ_g, Ξ_m, and Ξ_s are re-generated from the corresponding vocabularies with those cognitive and social keywords removed, and thus different from the ones used in the main analysis.

E Additional Tables And Figures

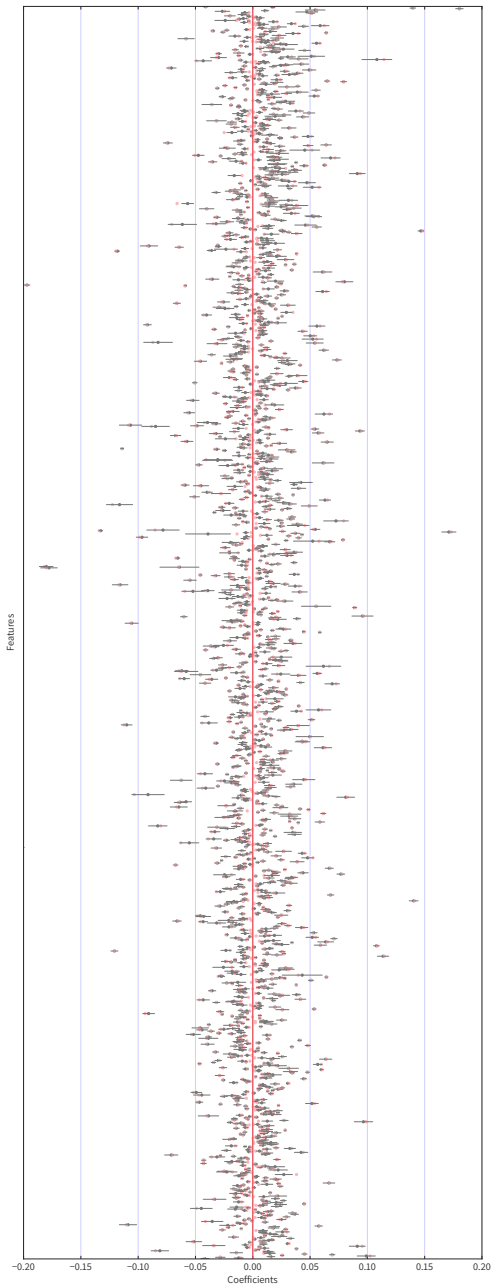
E.1 Additional Figures for Results of Machine learning Algorithms in Section 5

Figure E1: Subsampling on Lasso Non-Zero Coefficients

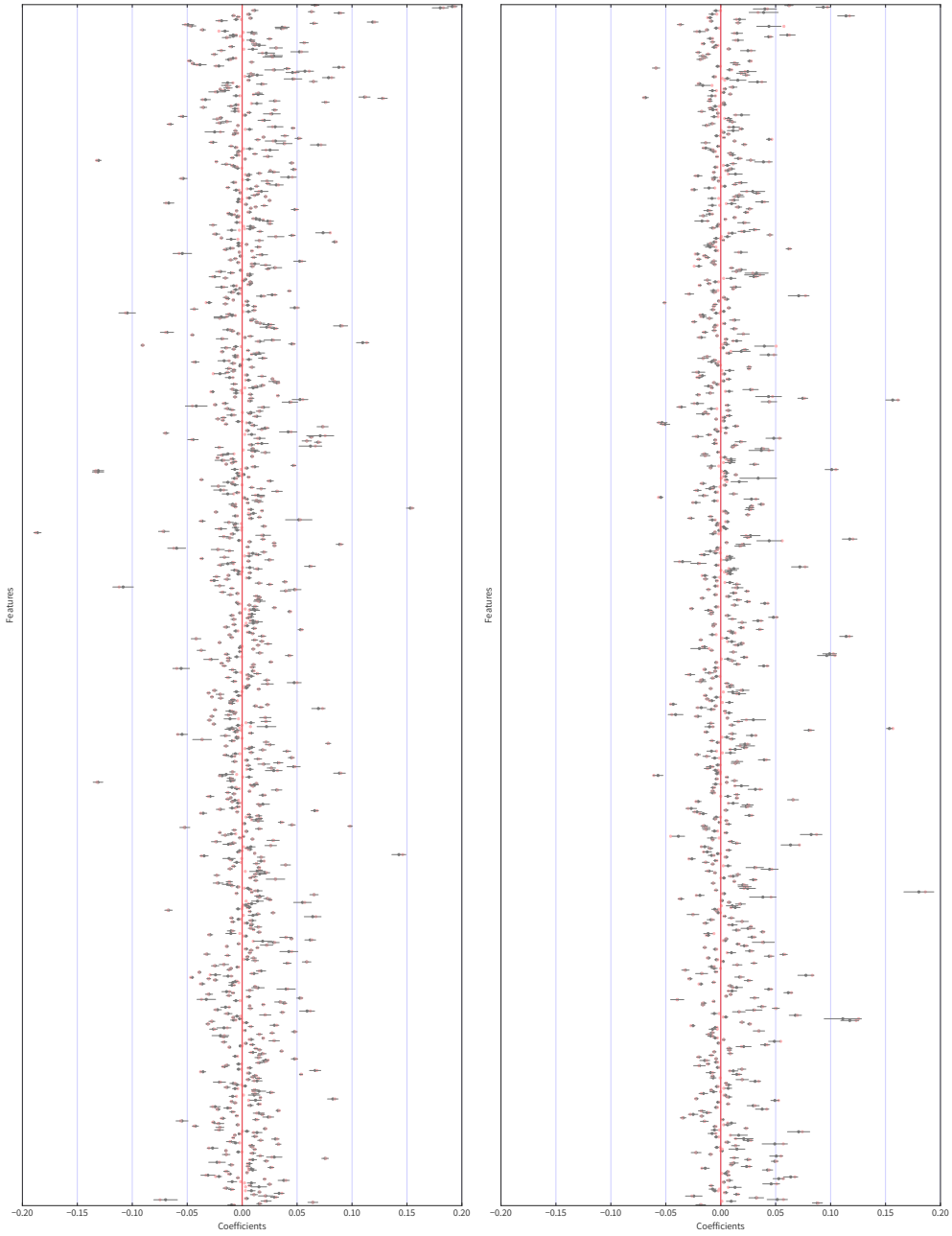
(a) Pooled



(b) Computer



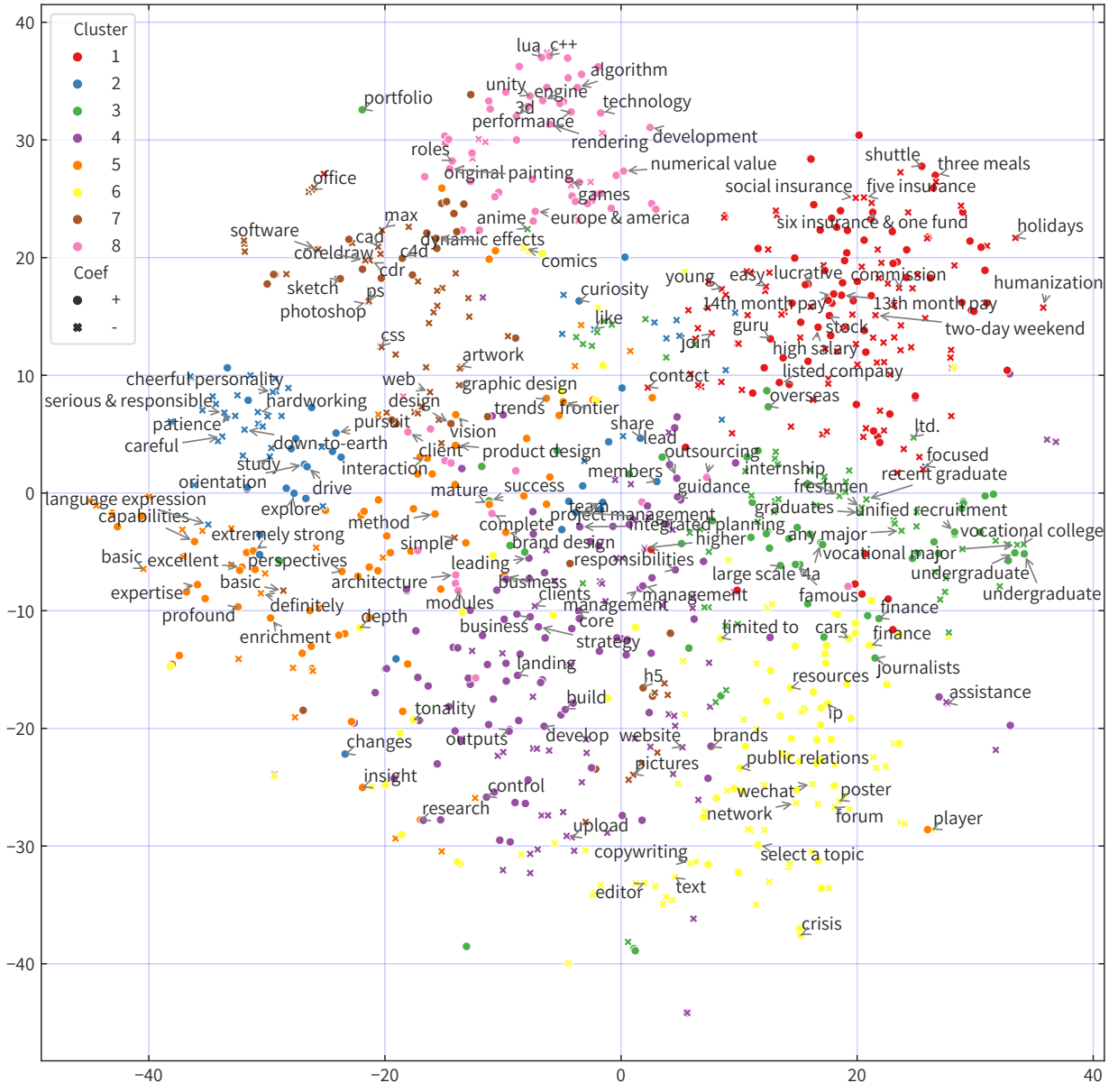
(c) Design_Media **(d)** Administrative



Notes. The figure plots the standard deviation for all nonzero coefficients under subsampling. The corrected standard deviation for each feature is calculated as $\text{sd}(\hat{\zeta})\sqrt{1/10}$ (because ten splits) under the assumption that the estimator's rate of convergence to be $\sqrt{N}$. Changing the number of splits or the rate of convergence does not qualitatively change the results.

Figure E2: Feature Clustering

(a) Design_Media



(b) Administrative

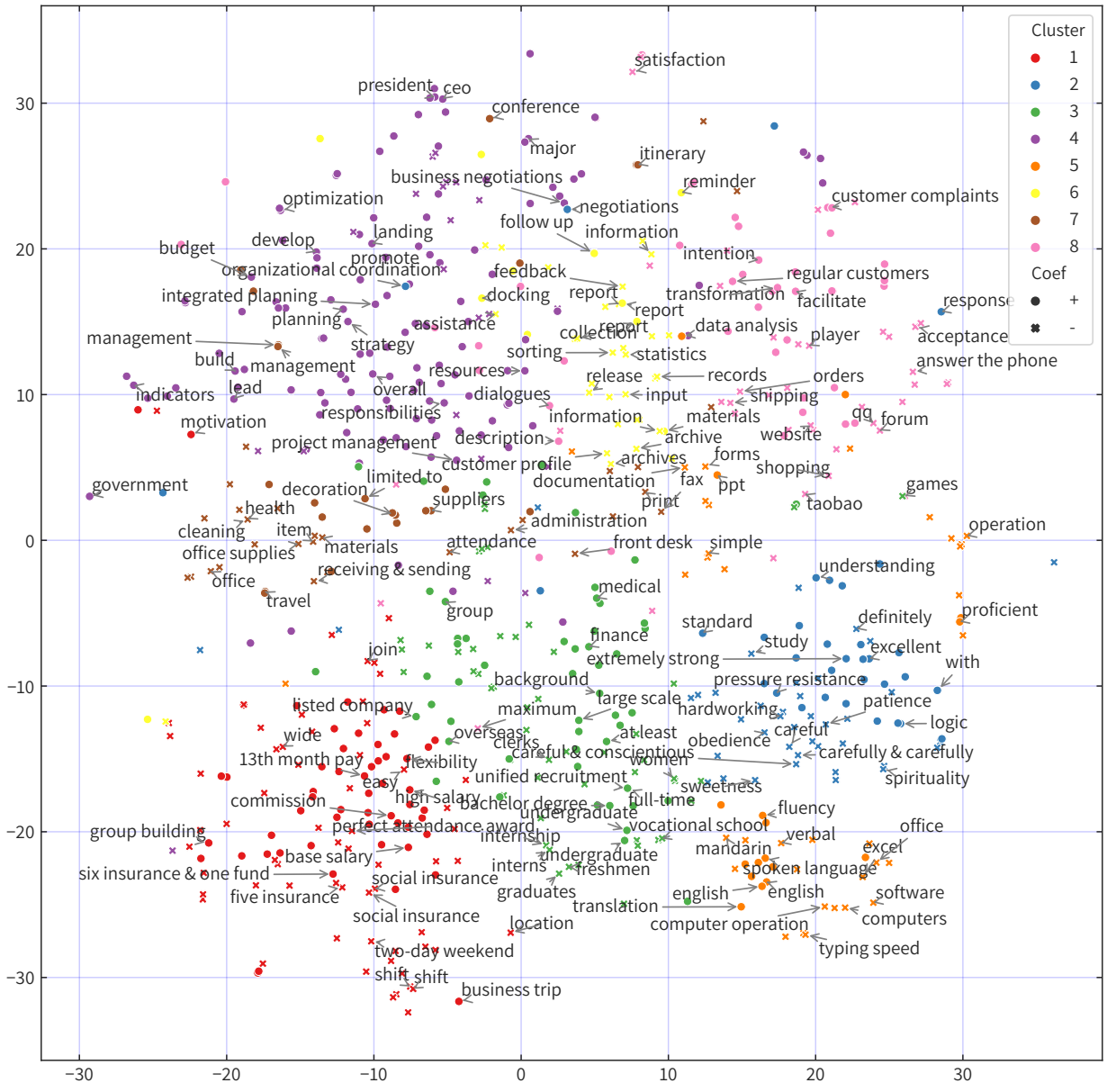
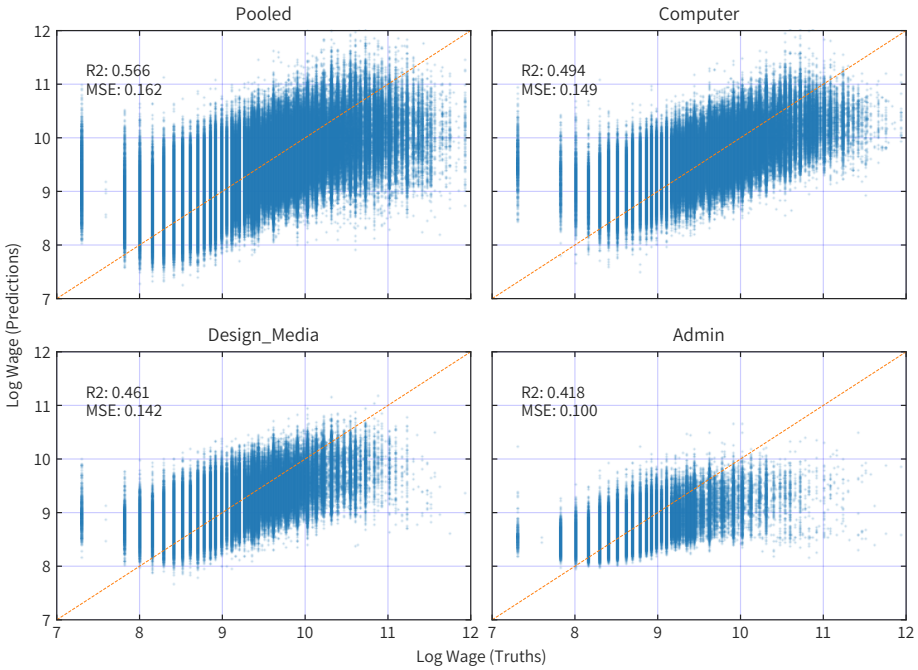


Figure E3: Model Predictions

(a) Lasso Prediction on Posted Wage



(b) OLS Prediction on Posted Wage Using $\{\Xi_{p=1}^{P}\}$

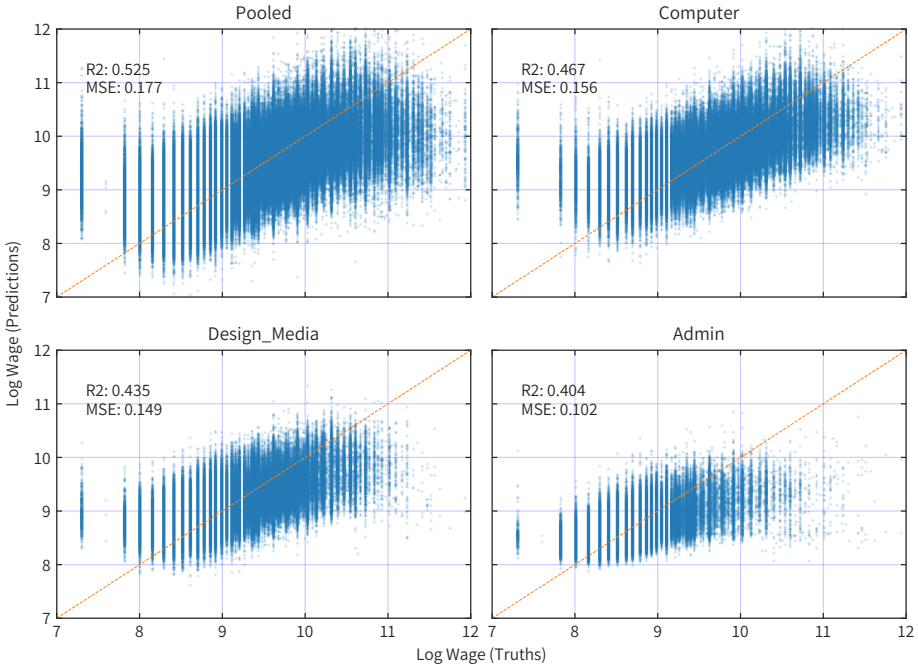
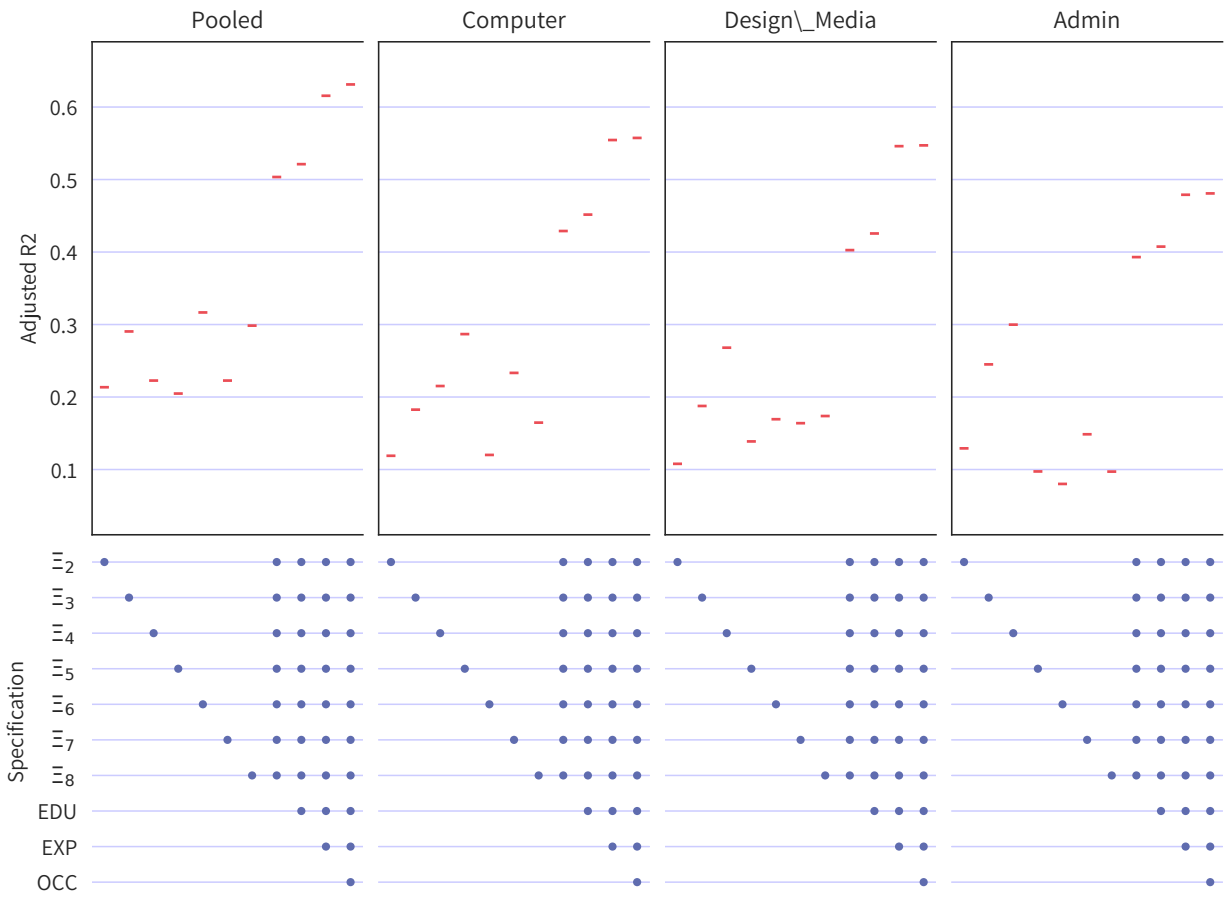


Figure E4: Prediction Power of Skill and Task Clusters in Posted Wage Regression



E.2 Additional Figures and Tables on the Robustness Checks in Section 6

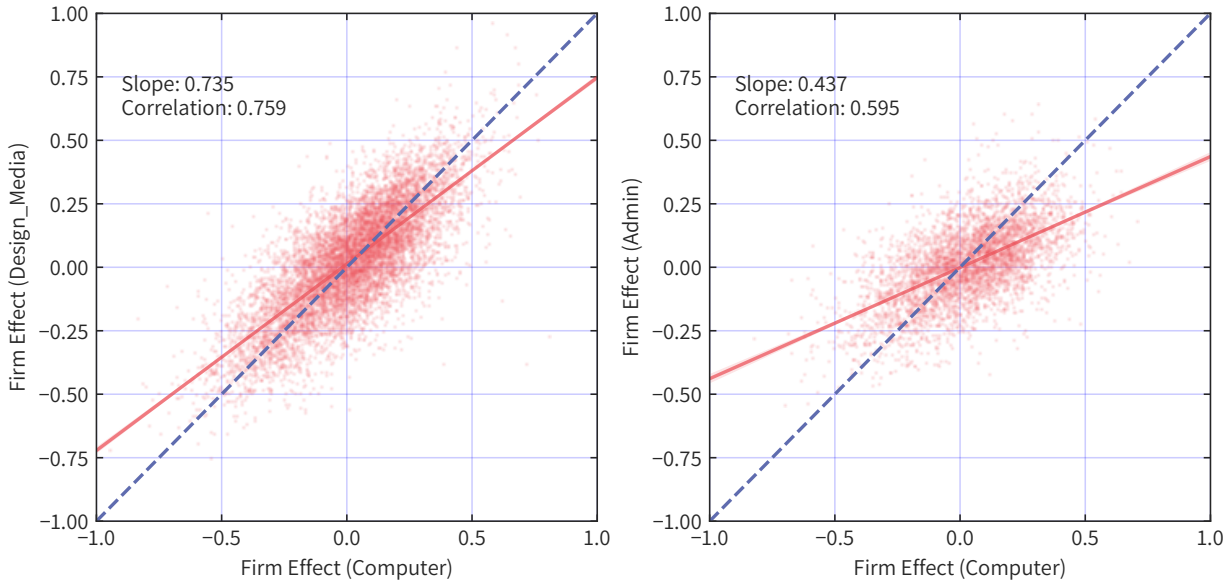
Table E1: Bias Correction on Posted Wage Variance ($X = \{\text{EDU, EXP}\}$)

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.360	-	.279	-	.251	-	.164	-
Panel A: Plug-In								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
$\text{Var}(\epsilon_i)$	.132	.367	.089	.318	.078	.310	.061	.371
$\text{Var}(\psi_j)$	.076	.212	.102	.365	.086	.342	.041	.253
$2\,\text{Cov}(\theta_j, \psi_j)$	.049	.137	.036	.130	.034	.136	.011	.069
Panel B: Homoscedasticity Correction								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
$\text{Var}(\epsilon_i)$	.135	.376	.093	.334	.087	.345	.072	.441
$\text{Var}(\psi_j)$	.073	.204	.097	.349	.077	.307	.030	.183
$2\,\text{Cov}(\theta_j, \psi_j)$	.049	.137	.036	.130	.034	.136	.011	.069
Panel C: KSS (Leave-Out) Correction								
$\text{Var}(\theta_i)$	.102	.283	.052	.188	.053	.212	.050	.307
$\text{Var}(\epsilon_i)$	.135	.374	.093	.332	.085	.339	.071	.431
$\text{Var}(\psi_j)$	.074	.205	.098	.350	.079	.314	.032	.193
$2\,\text{Cov}(\theta_j, \psi_j)$	.049	.138	.036	.130	.034	.136	.011	.069
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Table E2: Bias Correction on Posted Wage Variance ($X = \{X_e, \tilde{\Xi}\}$)

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.281	-	.253	-	.164	-
Panel A: Plug-In								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.331	.067	.408
$\text{Var}(\epsilon_i)$	.096	.267	.071	.252	.065	.255	.050	.304
$\text{Var}(\psi_j)$	.051	.141	.074	.263	.062	.243	.035	.216
$2\,\text{Cov}(\theta_i, \psi_j)$	.051	.142	.054	.193	.043	.171	.012	.072
Panel B: Homoscedasticity Correction								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.331	.067	.409
$\text{Var}(\epsilon_i)$	.099	.273	.074	.264	.072	.284	.059	.361
$\text{Var}(\psi_j)$	.049	.135	.070	.251	.054	.214	.026	.159
$2\,\text{Cov}(\theta_i, \psi_j)$	.051	.142	.055	.194	.043	.171	.012	.070
Panel C: KSS (Leave-Out) Correction								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.331	.067	.407
$\text{Var}(\epsilon_i)$	.098	.272	.074	.264	.071	.279	.058	.352
$\text{Var}(\psi_j)$	.049	.136	.071	.251	.056	.219	.028	.168
$2\,\text{Cov}(\theta_i, \psi_j)$	.052	.142	.054	.194	.044	.173	.012	.073
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Figure E5: Variation of Firm Effects Across Occupations



Notes. Same plot as Figure 3 except now the firm fixed effects are estimated by the firms with more than ten job posts in both occupations of each pair.

Table E3: Variance Decomposition Conditional on EXP=0

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.305	-	.407	-	.226	-	.097	-
Panel A: $X = \{\textbf{EDU}, \textbf{EXP}, \Xi_2, \ldots, \Xi_8\}$								
$\text{Var}(\theta_i)$	.079	.258	.069	.169	.036	.159	.014	.146
$\text{Var}(\epsilon_i)$	.115	.377	.111	.273	.084	.372	.049	.512
$\text{Var}(\psi_j)$	.068	.222	.138	.339	.075	.333	.029	.298
$2\,\text{Cov}(\theta_i, \psi_j)$	.044	.143	.089	.219	.033	.145	.005	.047
Panel B: Decompose θ **Terms**								
$\text{Var}(X_{int})$	.000	.000	.000	.000	.000	.000	.000	.000
$\text{Var}(X_{ext})$	.079	.258	.069	.169	.036	.159	.014	.146
$2\,\text{Cov}(X_{int}, X_{ext})$	.000	.000	.000	.000	.000	.000	.000	.000
$2\,\text{Cov}(X_{int}, \psi_j)$	.000	.000	.000	.000	.000	.000	.000	.000
$2\,\text{Cov}(X_{ext}, \psi_j)$	.044	.143	.089	.219	.033	.145	.005	.047
Panel C: Further Decompose X_{ext} **Terms**								
$\text{Var}(\Xi_g)$	.001	.004	.001	.003	.001	.005	.000	.002
$\text{Var}(\Xi_m)$	.005	.018	.010	.024	.004	.016	.003	.031
$\text{Var}(\Xi_s)$	.047	.153	.036	.087	.021	.094	.007	.068
$2\,\text{Cov}(\Xi_g, \Xi_m)$	.001	.004	.001	.004	.001	.002	.000	.004
$2\,\text{Cov}(\Xi_g, \Xi_s)$	.006	.021	.003	.008	.003	.012	.001	.009
$2\,\text{Cov}(\Xi_m, \Xi_s)$	.018	.058	.017	.043	.007	.032	.003	.032
$2\,\text{Cov}(\Xi_g, X_{int})$	.000	.000	.000	.000	.000	.000	.000	.000
$2\,\text{Cov}(\Xi_m, X_{int})$	.000	.000	.000	.000	.000	.000	.000	.000
$2\,\text{Cov}(\Xi_s, X_{int})$	.000	.000	.000	.000	.000	.000	.000	.000
$2\,\text{Cov}(\Xi_g, \psi_j)$	.003	.010	.005	.013	.002	.008	.000	.002
$2\,\text{Cov}(\Xi_m, \psi_j)$	.008	.027	.024	.060	.006	.029	.002	.022
$2\,\text{Cov}(\Xi_s, \psi_j)$	.032	.106	.059	.146	.024	.108	.002	.023
Obs	858147		144122		104960		120241	
Firm	66010		20060		19946		24807	

Table E4: Variance Decomposition If $\Xi_m \equiv \{\text{EDU}, \Xi_3, \Xi_4\}$

	Pooled		Computer		Design_Media		Admin	
	Comp.	Share	Comp.	Share	Comp.	Share	Comp.	Share
$\text{Var}(\ln w)$	.362	-	.281	-	.253	-	.164	-
Panel A: $X = \{\textbf{EDU}, \textbf{EXP}, \Xi_2, \ldots, \Xi_8\}$								
$\text{Var}(\theta_i)$	.163	.450	.082	.291	.084	.330	.067	.409
$\text{Var}(\epsilon_i)$	.098	.272	.074	.264	.071	.279	.058	.353
$\text{Var}(\psi_j)$	.049	.136	.071	.251	.056	.219	.027	.168
$2\,\text{Cov}(\theta_i, \psi_j)$	.052	.142	.054	.193	.043	.170	.012	.072
Panel B: Decompose θ **Terms**								
$\text{Var}(X_{int})$	.042	.115	.028	.099	.030	.119	.016	.096
$\text{Var}(X_{ext})$	.072	.199	.035	.126	.030	.117	.030	.184
$2\,\text{Cov}(X_{int}, X_{ext})$	.049	.136	.019	.067	.024	.094	.021	.129
$2\,\text{Cov}(X_{int}, \psi_j)$	.017	.048	.017	.060	.018	.072	.004	.025
$2\,\text{Cov}(X_{ext}, \psi_j)$	.034	.094	.037	.133	.025	.099	.008	.047
Panel C: Further Decompose X_{ext} **Terms**								
$\text{Var}(\Xi_g)$	.001	.003	.000	.001	.000	.001	.000	.002
$\text{Var}(\Xi_m)$	.017	.048	.007	.026	.006	.025	.018	.109
$\text{Var}(\Xi_s)$	.022	.062	.014	.051	.011	.045	.003	.019
$2\,\text{Cov}(\Xi_g, \Xi_m)$	.004	.010	.001	.003	.001	.004	.002	.011
$2\,\text{Cov}(\Xi_g, \Xi_s)$	.005	.012	.001	.005	.001	.004	.001	.003
$2\,\text{Cov}(\Xi_m, \Xi_s)$	.023	.064	.011	.039	.009	.037	.007	.041
$2\,\text{Cov}(\Xi_g, X_{int})$	.004	.011	.001	.004	.001	.005	.001	.006
$2\,\text{Cov}(\Xi_m, X_{int})$	.020	.054	.006	.022	.011	.042	.017	.102
$2\,\text{Cov}(\Xi_s, X_{int})$	.026	.071	.011	.041	.012	.047	.003	.020
$2\,\text{Cov}(\Xi_g, \psi_j)$	.002	.007	.002	.007	.001	.005	.000	.001
$2\,\text{Cov}(\Xi_m, \psi_j)$	.014	.040	.015	.052	.012	.048	.007	.040
$2\,\text{Cov}(\Xi_s, \psi_j)$	.017	.048	.021	.075	.012	.046	.001	.007
Obs	3998840		1325260		548808		260364	
Firm	86165		62628		55664		41448	

Figure E6: Work Types and Posted Wage by Firm Types

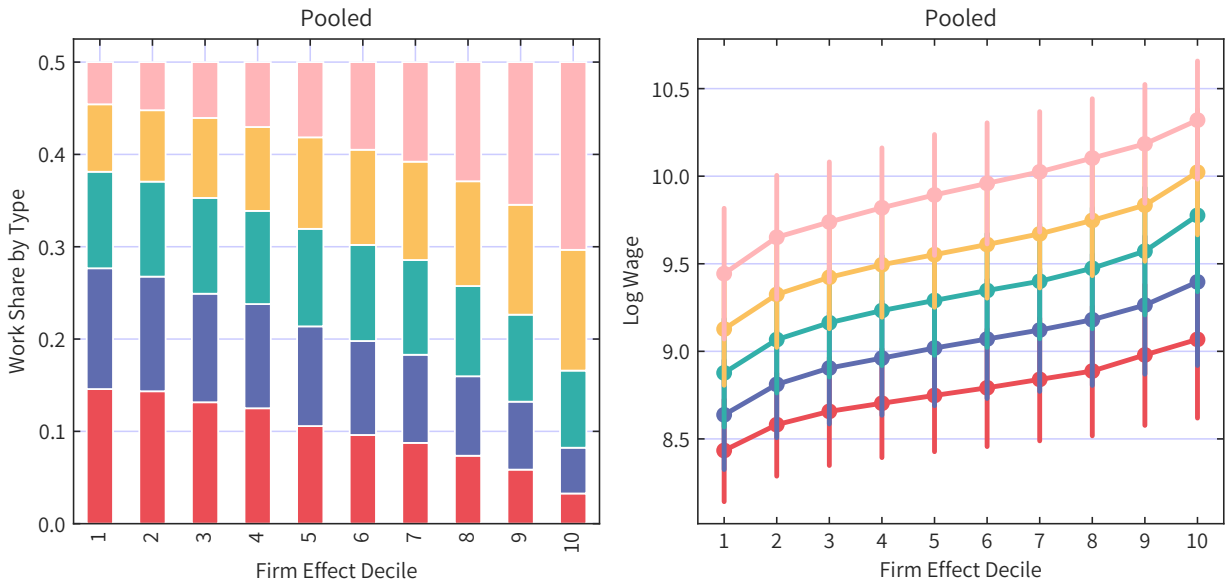


Figure E7: Mean Residual for Work-Firm cells

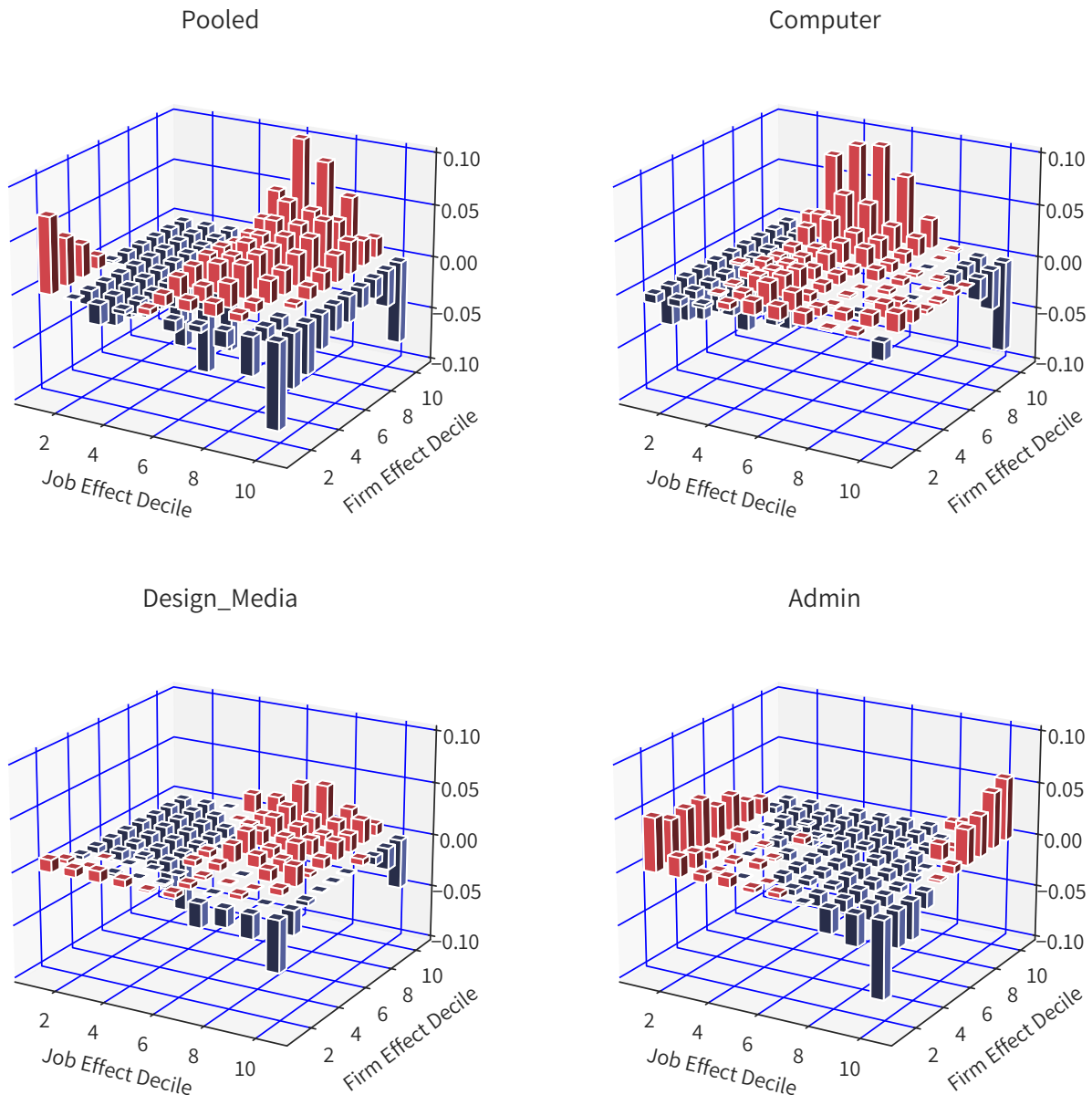


Table E5: Firm Fixed Effect and Firm Characteristics

	Pooled			Computer			Design_Media			Admin		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
fsize.15-50	.019**	.018**	.023**	.012	.011	.014+	.049**	.035**	.045**	-.032	-.039	-.034
	(.004)	(.004)	(.003)	(.010)	(.009)	(.008)	(.011)	(.010)	(.008)	(.038)	(.034)	(.033)
fsize.50-150	.044**	.038**	.050**	.043**	.034**	.032**	.083**	.058**	.073**	-.023	-.038	-.035
	(.004)	(.004)	(.003)	(.010)	(.009)	(.007)	(.010)	(.010)	(.008)	(.038)	(.034)	(.033)
fsize.150-500	.069**	.059**	.068**	.079**	.053**	.043**	.127**	.087**	.094**	-.009	-.032	-.032
	(.004)	(.004)	(.003)	(.010)	(.009)	(.008)	(.011)	(.010)	(.009)	(.038)	(.034)	(.033)
fsize.500-2000	.099**	.081**	.086**	.119**	.070**	.053**	.176**	.121**	.120**	.015	-.014	-.019
	(.005)	(.004)	(.004)	(.011)	(.009)	(.008)	(.012)	(.011)	(.009)	(.038)	(.035)	(.033)
fsize.2000+	.125**	.105**	.121**	.154**	.077**	.065**	.213**	.140**	.134**	.028	-.005	-.006
	(.005)	(.005)	(.004)	(.011)	(.010)	(.008)	(.013)	(.012)	(.010)	(.038)	(.035)	(.034)
Job Effect ($\tilde{\theta}$)		.284**	.200**		.793**	.622**		.479**	.395**		.262**	.171**
		(.004)	(.003)		(.009)	(.008)		(.010)	(.009)		(.020)	(.018)
const	.148**	-1.101**	-.630**	-.176**	-3.946**	-3.018**	.157**	-1.931**	-1.488**	.175**	-.919**	-.468**
	(.003)	(.016)	(.015)	(.010)	(.042)	(.037)	(.010)	(.046)	(.040)	(.038)	(.079)	(.073)
Location FE			✓			✓			✓			✓
Adj. R^2	.017	.096	.381	.025	.243	.515	.053	.190	.473	.014	.062	.292
No. Obs	84023	84023	84023	30658	30658	30658	13871	13871	13871	5592	5592	5592

E.3 Additional Figures and Tables on the Additional Analysis in Section 7

Figure E8: Work Types and Posted Wage by Firm Types

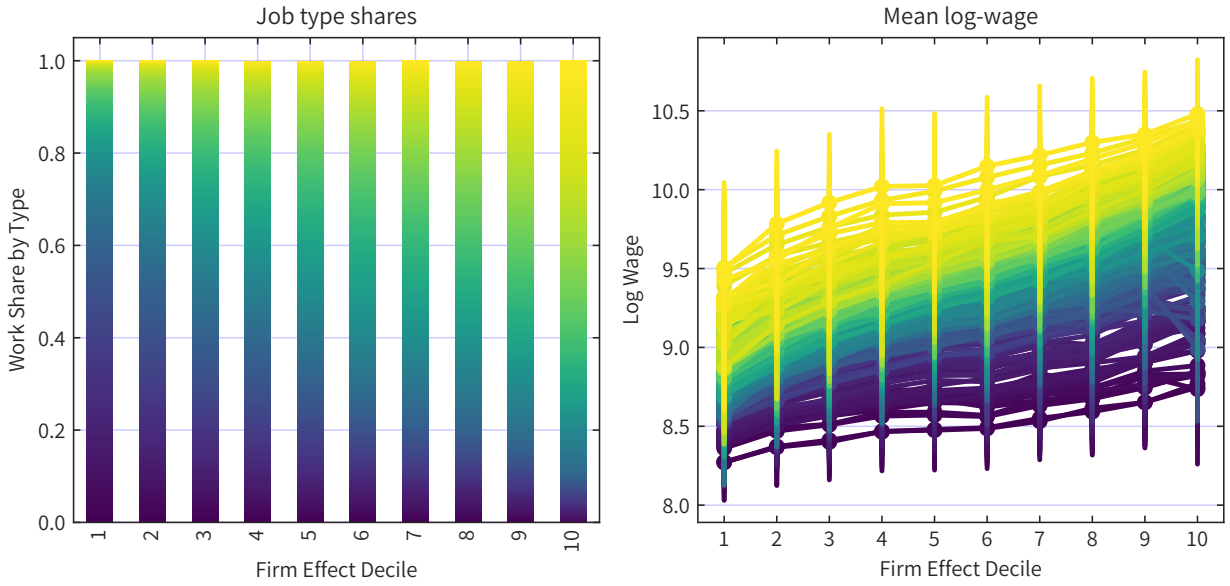
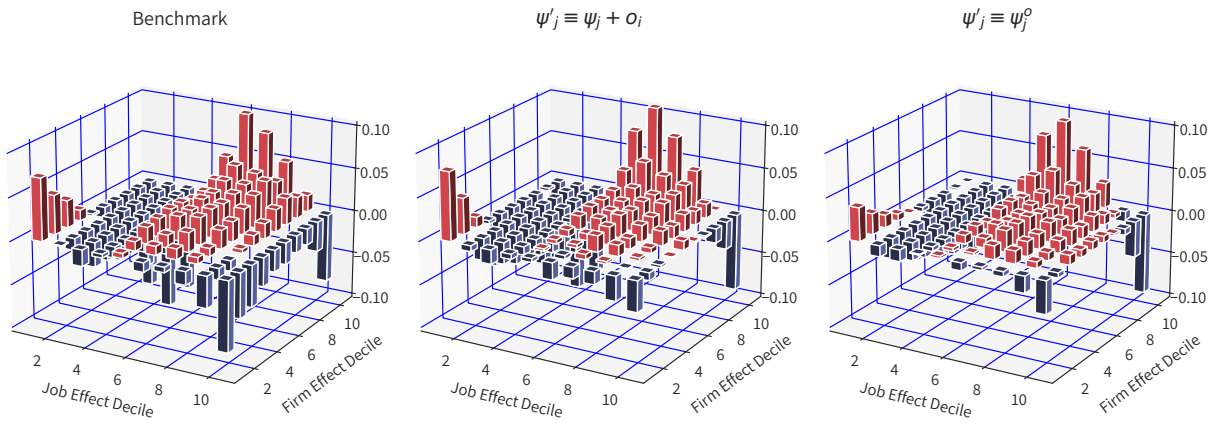


Figure E9: Mean Residual for Work-Firm cells



Appendix Reference

Atalay, E., P. Phongthiengtham, S. Sotelo, and D. Tannenbaum (2020). The evolution of work in the united states. *American Economic Journal: Applied Economics 12*(2), 1–34.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.

He, C., K. Mau, and M. Xu (2021). Trade shocks and firms hiring decisions: Evidence from vacancy postings of chinese firms in the trade war. *Labour Economics*, 102021.

Turrell, A., B. J. Speigner, J. Djumalieva, D. Copple, and J. Thurgood (2019). Transforming naturally occurring text data into economic statistics: The case of online job vacancy postings. Technical report, National Bureau of Economic Research.