The Geography of Job Tasks

Enghin Atalay Sebastian Sotelo Daniel Tannenbaum*

January 27, 2021

Abstract

We present new facts about the geography of work using online job ads and introduce new measures of job tasks, technology requirements, and the degree of specialization within firms or occupations. We show that (i) the intensity of interactive and analytic tasks, (ii) technological requirements, and (iii) task specialization all increase with city size. The gradient for tasks and technologies is steeper for jobs requiring a college degree. We show that these facts help account for the urban wage premium, both in aggregate and across skill groups.

JEL Codes: J20, J24, R12

*Atalay: Research Department, Federal Reserve Bank of Philadelphia. Sotelo: Department of Economics, University of Michigan-Ann Arbor. Tannenbaum: Department of Economics, University of Nebraska-Lincoln. We thank Reid Gahan and Ryan Kobler for outstanding research assistance. We thank seminar participants at the annual meetings of the Society of Labor Economists and the Urban Economics Association, the UNL labor reading group, and Gregor Schubert for valuable feedback. This work is supported in part by grant #92-18-05 from the Russell Sage Foundation. Sotelo thanks the IES at Princeton for its hospitality during part of this research. The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

1 Introduction

Rural-urban inequality in the United States has been widening since the 1980s, as economic growth has increasingly concentrated in large cities (Moretti, 2013; Giannone, 2019; Eckert, Ganapati, and Walsh, 2019). Behind these broad geographic trends, workers at different education levels have fared differently: the urban wage premium, which was equally steep for college and non-college educated workers in the 1970s and 80s, has flattened for non-college workers since the 2000s (Baum-Snow, Freedman, and Pavan, 2018; Autor, 2019).

Understanding the forces behind the productivity gains in cities and the erosion of these gains for workers without a college degree is a critical input in the design of policies ranging from local taxation to incentives for worker mobility. But while economists have studied the many ways productivity relates to location, prior research has been limited in its ability to characterize the geography of work by both the available data sources and their measures of job tasks and technologies.

One key measurement challenge is that job tasks and technologies are typically measured at the occupation level, using nationwide surveys such as O*NET. Using an occupation-level dataset, it is not possible to measure the extent to which the content of occupations varies across markets. In this paper we fill this measurement gap. We study the geography of job tasks and technology requirements in the U.S., using a novel approach to measurement and a new data source: the text of online job ads, published between 2012 and 2017. Our task and technology measures are not fixed at the occupation level, and allow for heterogeneity within and across regions.

We take two approaches to task measurement. The first approach, following our prior work on newspaper job postings (Atalay, Phongthiengtham, Sotelo, and Tannenbaum, 2018, 2020), maps words in job descriptions into routine and non-routine task categories. Our second approach uses tools from natural language processing to define tasks as verb-noun pairs in the job descriptions, thus imposing fewer ex ante restrictions on the classification of tasks. This second approach departs from the existing task literature, where it is common to select a subset of survey questions from O*NET and then classify these items into economically meaningful task categories (Autor, 2013). There are two key advantages to our more granular approach to task measurement: first, it reduces the amount of researcher discretion in classifying tasks, and second, it allows us to measure how *specialized* jobs are – i.e., how far apart workers are in task space, within firms or occupations.

We validate our data and task measures in several ways. We show that online vacancies by sector are aligned with those measured in the Job Openings and Labor Turnover Survey (JOLTS). The education requirements in job ads, moreover, are highly correlated with the

education of employed workers in the American Community Survey (ACS). Turning to our new task measures, we first show that when we create occupation-level task measures from the job ad text that correspond to O*NET task categories, these measures are highly correlated with O*NET importance scales. We also show our measures of job tasks account for variation in average wages – even within occupations. Hence, we capture occupational characteristics beyond what is available in standard data sources, and these characteristics are also reflected in market wages as observed in the ACS. These validation exercises are presented in Appendix A.

Our main empirical analysis presents several new facts on the geography of work in the United States. We first show that analytic and interactive tasks have a steep positive gradient in market size. For example, relative to the bottom population decile, commuting zones in the top population deciles have 0.20 to 0.30 standard deviations higher intensity of non-routine analytic and interactive tasks. This gradient remains significant even after conditioning on narrowly defined occupation categories (six-digit SOC codes), and hence, is not merely driven by the composition of occupations across markets. Exploiting the richness of our data, we decompose interactive tasks into those capturing interactions outside the firm and those capturing interactions within the firm. We find that the market size gradient is positive for both external interactive tasks and internal interactive tasks and that this relationship is more pronounced for jobs requiring a college degree.[1]

We next consider whether the technological requirements of jobs, as measured by the appearance of any of O*NET's "Hot Technologies," are more likely to be mentioned in job descriptions in larger markets, and how this gradient differs for high and low-skilled jobs. We find a positive gradient of technological requirements with respect to market size. The proportion of job ads in the first decile commuting zone that mentions any of the technologies is 7 percent and it is approximately 2.5 times higher in the seventh through tenth deciles. About 15 percent of the gradient remains after conditioning on six-digit occupational categories. Moreover, the technology gradient is only present for jobs requiring a college degree, and vanishes for jobs requiring only high school. This provides evidence that technologies are an important mechanism behind the flattened urban wage premium for non-college workers (Autor, 2019). Aligned with this interpretation, the technologies with the steepest gradient for college degree-holders involve computer programming (e.g., Python, Javascript, and Linux), while for high school degree-holders they involve data entry and word processing (e.g., Microsoft Excel, Microsoft Outlook, Microsoft Word) and social media (e.g., LinkedIn,

[1]We also show that jobs that are jointly intensive in interactive and analytic tasks are overrepresented in large markets. Thus, the increasing aggregate importance of social and analytic tasks since the 1980s (Deming, 2017) is mirrored by the differential task content between rural and urban labor markets.

Facebook).[2]

Our paper also introduces a novel approach for measuring the degree of specialization using the content of job descriptions. Specifically, we first extract verb-noun pairs from the job description text to measure tasks at a more granular level.[3] We then measure the degree of specialization between two jobs as the cosine dissimilarity between vectors representing their task contents. This approach is similar to that of Hoberg and Phillips (2016) in their measurement of product differentiation in firms. We show that task specialization is increasing in market size and that this relationship is stronger for non-tradeable sector firms.

Our paper contributes to the literature that studies the geography of job tasks and technologies (e.g., Frank, Sun, Cebrian, Youn, and Rahwan, 2018). Our measures capture heterogeneity within occupations and across geography, which has been unavailable in prior work. Part of the literature on tasks also exploits job vacancy postings across different labor markets (Hershbein and Kahn, 2018; Deming and Kahn, 2018; Hemelt, Hershbein, Le, Martin, and Stange, 2020). We contribute to this literature by introducing a new approach to task measurement, which uses natural language processing and requires fewer restrictions ex ante relative to widely used O*NET scales and categories. We emphasize, moreover, the role of market size and specialization.

Recent work explores interactions between workers as a source of agglomeration, both theoretically (Davis and Dingel, 2019) and empirically (Bacolod, Blum, and Strange, 2009; Michaels, Rauch, and Redding, 2018). Our contribution with job vacancy data is to characterize the spatial distribution of job tasks, and to provide new evidence informing several key mechanisms underlying urban productivity gains. We find that urban jobs are more intensive in interactive and analytic tasks, and in their technological requirements, and this rural-to-urban gradient is much stronger for jobs requiring a college degree. We also show within-occupation task and technological differences are strongly predictive of the within-occupation spatial wage premium, even after controlling for educational differences of workers across geography. Notably, while this association is strong for white collar and service occupations, it is weak for blue collar occupations. These findings suggest that the returns to

[2]These results complement an expanding literature studying the spatial distribution of technology adoption. Eckert, Ganapati, and Walsh (2019) emphasize the impact of cheaper ICTs on services that agglomerate in large cities and focus on the creation and communication of information. Bloom, Hassan, Kalyani, Lerner, and Tahoun (2020) examine where new technologies develop and how they diffuse. In prior work, we study technology adoption within occupations over the twentieth century (Atalay, Phongthiengtham, Sotelo, and Tannenbaum, 2018).

[3]We build on Michaels, Rauch, and Redding (2018), who use verbs from job descriptions in the Dictionary of Occupational Titles and then categorize them based on Roget's Thesaurus. In this paper, we extract verb-nouns from the job description text and use them to measure how task contents vary across markets, even within occupations.

employment in large cities differ by skill group. The findings also suggest that both the task content and the technology requirements of occupations shift from rural to urban markets, which may limit the mobility of workers, even within the same occupation.

We also contribute to the literature that relates the division of labor to the extent of the market (Young, 1928; Stigler, 1951; Kim, 1989; Becker and Murphy, 1992). Recent work finds greater occupational diversity in cities. Duranton and Jayet (2011), for example, show that specialist occupations are overrepresented in large cities in France, and Tian (2019) uses Brazilian data to show that firms demand a greater number of distinct occupations in larger markets, which helps account for the wage-market size premium. Our contribution to this literature is to measure the degree of specialization directly, going beyond occupations, and to show that job tasks account for a substantial portion of the urban wage premium.

2 Data and Measurement

Our data source is a comprehensive database of online job ads, posted between January 2012 and March 2017, which we purchased from Economic Modeling Specialists International (EMSI). This dataset is similar to Burning Glass, which has been used in recent work to study the labor market (Hershbein and Kahn, 2018; Deming and Kahn, 2018; Modestino, Ballance, and Shoag, 2020). Like Burning Glass, our data are proprietary and assembled using web crawlers that extract job vacancy postings from all major online job boards, and de-duplicates postings that appear across boards. A virtue of the EMSI data is that it contains all of the original job ad text. To reduce computational time, we use a five percent random sample of the data, 7.2 million ads.[4]

In addition to the full text content of each ad, the data include additional information that EMSI extracts from the postings, including the educational requirement of the job, the firm name (which we use to create firm identifiers), the firm's industry (six-digit NAICS), the occupation code (six-digit SOC), and the job location (county FIPS code). We map FIPS codes to commuting zones (CZs) following Autor, Dorn, and Hanson (2019). We adopt the CZ as our geographic unit of analysis and refer to CZs throughout as local labor markets. Appendix A.1 provides descriptive statistics for the CZs in the sample, including population and number of ads by CZ employment decile. We exclude ads with fewer than the 1st and more than the 95th percentile word count.[5] We also exclude ads in Hawaii and Alaska, and

[4]EMSI is our preferred data source because it contains the complete job description text, which is ideal for extracting job tasks and measuring specialization. By contrast, Burning Glass provides a combination of tasks, skills, and technologies. Nevertheless, we reproduce our main results using Burning Glass data and report them in Appendix C.5. Reassuringly, our results are similar with this alternate data source.

[5]Dropping extremely short ads removes those that are unlikely to have meaningful task information,

exclude ads with missing occupation or FIPS codes. For the firm-level analysis, we drop ads with missing firm names or industry codes, and also drop staffing firms, which are flagged in the data, since these firms act as intermediaries between the worker and firm hiring the worker. These restrictions leave us with a sample of 6.3 million ads for the occupational analysis and 5.6 million ads for the firm-level analysis. The sample restrictions are detailed in Appendix A.2.

For the several exercises requiring wages at the occupation level, and for the construction of employment weights, we use the 2010-2017 ACS (Ruggles, Flood, Goeken, Grover, Meyer, Pacas, and Sobek, 2020), and restrict the sample to full-time, full-year workers, defined as working at least 40 weeks in the past year and 35 or more hours per week. We apply a chain-weighted price deflator for personal consumption expenditures to wages before averaging at the four-digit SOC. We link job ads data to the ACS by four-digit SOC and CZ.

We assess the representativeness of the online ads data in Appendix A.3, comparing our data to the Job Openings and Labor Turnover Survey (JOLTS) dataset. We find broad concurrence in the vacancy shares across industries, suggesting that online vacancies measure a fairly representative cross-section of total vacancies. Certain industries, such as Manufacturing, Finance and Insurance, and Education have higher representation in EMSI compared to JOLTS, while others such as Health and Social Assistance, Government, and Accommodation and Food have higher representation in JOLTS.

2.1 Measuring Tasks: Classification and Extraction

We extract job tasks from the job descriptions using two approaches. Our first approach follows our earlier work (Atalay, Phongthiengtham, Sotelo, and Tannenbaum, 2018, 2020) and maps keywords in the job descriptions to task categories. Our main analysis maps words into five task categories – non-routine interactive, non-routine analytic, non-routine manual, routine cognitive, routine manual – following the categorization of Spitz-Oener (2006). We also map words into O*NET work activities. See Appendix A.4 for more details on the word mappings. For job ad j and task category k, our measure of task intensity is the number of distinct task-specific word mentions per 1,000 ad words.[6] We standardize each task to have mean zero and standard deviation one across all ads.[7]

while dropping exceedingly long ads helps reduce computation time.

[6]We count repeated uses of the same word only once. Hence, repetitiveness of the job description does not inflate the task intensity of the ad. Uses of different task keywords, such as "analyze" and "evaluate," will each be counted and will increase the task intensity measure.

[7]In Atalay, Phongthiengtham, Sotelo, and Tannenbaum (2020), we show robustness to the choice of word mappings, e.g., by including and excluding synonyms of words in the mapping to tasks, and to alternative task units.

Our second approach is novel and uses verb-noun pairs in the job descriptions to define the set of job tasks. This approach avoids using a researcher-defined mapping of words to task categories and leverages the rich database of text using tools from natural language processing. An additional advantage of this approach is that it defines tasks at a highly granular level, allowing us to carefully measure the degree of specialization of jobs that share the same occupational code.

We describe this approach in detail in Appendix B and briefly outline it here. A task is defined as a (verb stem, noun stem) pair, such as "assist customers" or "provide advice." We stem verbs and nouns so that variation in verb and noun forms do not affect the analysis. We extract the task categories as follows: (i) we first identify the section of ad text that refers to job tasks, (ii) within this section of text, we find each verb and the next noun that appears in the sentence, ignoring other parts of speech that may appear in between. For our analysis, we extract the 500 most common verb-noun pairs.[8] Once we have assembled the set of tasks, we represent each job ad as a vector, of which each element corresponds to a distinct task. Verb-noun pairs that appear multiple times in an ad are counted only once, and hence, each element is a zero or one. We exclude 70 verb-noun pairs that in our judgment do not correspond to job tasks, such as "send resume" and "is position," and hence the number of tasks used in the analysis is 430. The ten most common tasks, from most to least frequent, are: "written communication," "perform duties," "working team," "provide customer_service," "be part," "provide service," "work environment," "perform functions," "build relationships," "ensure compliance." While the task extraction process is not perfect, a key strength of our approach is it allows the natural text used by employers, describing the jobs they intend to fill, to define the set of tasks.

2.2 Validation of Data and Task Measures

We demonstrate that information contained in the online ad text captures real information about the labor market in Appendix A.5. We compare the education requirements extracted from the job ads to the education of employed workers in the 2010-2017 ACS, in the same occupation-market. We find these two measures of education are highly correlated, a relationship that holds across large and small markets, within and across occupations. We also validate the task measures extracted from the ads and compare these measures to O*NET. First, we show that occupation-level measures of O*NET Work Activities, which we construct from the text of online ads, are highly correlated with those occupations' measures

[8] We choose 500 tasks to balance the advantage of comprehensively characterizing jobs' tasks against the costs of computational time. We reproduce the key specialization results using the 300 most common tasks, in Appendix C, and show that our results are insensitive to the number of tasks.

in the O*NET database (Figure A.3). Second, in Appendix B.6, we show that our task measures, constructed using either of the two approaches, account for variation in average wages at the occupation level, above and beyond what is captured by occupation fixed effects. These task measures therefore capture occupational characteristics beyond what is available in O*NET, and these characteristics are reflected in market wages.

3 The Geography of Tasks and Technologies

This section presents the main analysis of the geography of job tasks, technology requirements, and worker specialization.

3.1 Job Tasks Across Space

We begin with our first approach to task measurement and study how the five task categories (non-routine interactive, non-routine analytic, non-routine manual, routine cognitive, and routine manual) differ across labor markets of different sizes. For each task k, we regress task intensity $t_{jn}^{(k)}$ of job j in market size decile n on indicators for market size decile. CZs are placed in market size deciles using employment weights so that each decile n has approximately the same number of employed workers. We estimate:

$$t_{jn}^{(k)} = \beta_0 + \sum_{n=2}^{10} D_{jn}\beta_n^{(k)} + \gamma' x_{jn} + \epsilon_{jn}, \tag{1}$$

where D_{jn} are indicators for market size decile n, with the first decile serving as the reference group, and x_{jn} represents a control for ad length and, in some specifications, six-digit SOC fixed effects. The coefficients of interest, $\beta_n^{(k)}$, capture the task intensities relative to the first decile market size. Standard errors are clustered at the commuting zone level.

Figure 1 plots the coefficients on market size decile, $\beta_n^{(k)}$. The primary takeaway is that non-routine interactive and non-routine analytic tasks are increasing in market size, while routine manual tasks are decreasing in market size. The tenth population decile has 0.20 s.d. greater intensity of non-routine interactive tasks and 0.30 s.d. greater intensity of non-routine analytic tasks, while having approximately 0.20 s.d. lower intensity of routine manual tasks. The right panel includes six-digit SOC fixed effects, and shows that the gradient diminishes. This weaker gradient is unsurprising and indeed reassuring, since occupational categories are designed to group jobs by their work activities. Nevertheless, even within occupations, non-routine interactive and analytic tasks are mentioned more frequently (by 0.05 s.d.), and routine manual tasks are mentioned less frequently (by 0.08 s.d.), in the top population decile

CZs relative to the bottom decile CZs. Hence, while much of the variation in job tasks across geography is captured by the composition of occupations, a strong gradient remains even within occupations, which is missed in standard data sources such as O*NET.[9]

Figure C.1 presents the analysis for interactive and analytic tasks separately by the education requirement of the job ad. Jobs requiring a college degree in urban areas are far more intensive in interactive and analytic tasks compared to rural areas, while this gradient is flat for jobs requiring only a high school degree. Figure C.2 shows that jobs that are *jointly* intensive in interactive and analytic tasks represent a greater share in large markets, which mirrors these jobs' increasing importance over time (Deming, 2017). Jobs that are intensive in both analytic and interactive tasks make up 15 percentage points more of jobs in each of the highest three deciles compared to the lowest decile. Jobs that are intensive in only analytic tasks but not interactive tasks make up only about four percentage points more of jobs in the highest three deciles. These qualitative findings also hold within occupations.

Interactive Tasks Inside and Outside the Firm

Having demonstrated the importance of interactive tasks in urban labor markets, we study the nature of interactive tasks and specifically assess the importance of interactions *inside* the firm relative to interactions *outside* the firm. Our findings speak to a recent theoretical and quantitative literature that emphasizes worker interactions and information flows as sources of productivity differences across geographies and worker types.[10]

We use task measures that map to O*NET task categories that separately measure external and internal interactive tasks.[11] We regress each task intensity measure on commuting zone size deciles, with controls for ad length and, where indicated, six-digit SOC fixed effects. Figure 2 plots the coefficients on market size decile, with the first decile as the reference decile. The results show that the gradient of external tasks with market size is stronger than that of internal tasks. Compared to ads in the bottom population decile, ads in the top

[9]We perform a simple decomposition in Appendix C.1 to quantify how much of the variation in job tasks across city size is within versus between occupations. Between the top population quartile and bottom population quartile CZs, 35 percent of the difference in non-routine analytic task intensity occurs within occupations. For the other four task measures, this fraction ranges from -6 percent (for routine cognitive tasks) to 50 percent (for non-routine manual tasks).

[10]For example, Davis and Dingel (2019) present a theory in which the key agglomeration force is the exchange of ideas between firms, rather than within firms. Garicano and Rossi-Hansberg (2015) review the literature on knowledge flows across hierarchies within firm boundaries.

[11]We define *external interactive tasks* as O*NET activities "Selling or Influencing Others" and "Communicating with Persons Outside Organization;" and we define *internal interactive tasks* as O*NET work activities "Guiding, Directing, and Motivating Subordinates," "Developing and Building Teams," "Coaching and Developing Others," "Coordinating the Work and Activities of Others," and "Communicating with Supervisors, Peers, or Subordinates." We list the word mappings in Appendix A.4.

population deciles mention internal interactive tasks (by 0.20 s.d.) and external interactive tasks (by 0.25 s.d.) more frequently. When including six-digit SOC occupation fixed effects, the gradients are substantially smaller, though still economically and statistically significant.

Second, the gradient is driven largely by jobs requiring a college degree, as seen in Figure C.3. College workers in the seventh to tenth decile CZs have about 0.25 s.d. higher intensity of external interactive tasks compared to the first decile, and 0.09 s.d. higher intensity of internal interactive tasks, shown in panel I. This gradient is far more muted for jobs requiring a high school degree, which is shown in panel III, where the corresponding estimates are approximately 0.05 and 0, respectively. Lastly, comparing panel I, which does not have SOC fixed effects, to panel II, which does, shows that while much of the gradient for college degree-requiring jobs is due to occupational composition, even within occupations there is a higher and statistically significant component of external interactive tasks (0.04 to 0.07 s.d.).

A Granular Approach to Measuring Tasks

In our second approach, we refine our task measurements using a more granular set of job tasks, which are verb-noun pairs extracted from the text. We estimate equation (1) separately for each of the tasks, and collect the coefficients $\hat{\beta}_{10}^{(k)}$, which captures the relative difference in task k intensity between tenth decile market size and first decile market size. The coefficients are normalized by dividing by the standard deviation of the task and then sorted by magnitude. Table B.4 presents the largest positive and largest negative estimates across all tasks.

Our results echo, at a much higher resolution, what we found in Figure 1. Placing little guidance on the categorization of tasks, and using the natural language of the job ad descriptions to measure tasks, this exercise reveals that non-routine and abstract tasks have the steepest positive gradient. Examples include "managing projects," "problem-solving skills," and "developing strategies." Communication and group interactions are important, too, as illustrated by the importance of "written communication" and "maintaining relationships." The tasks with the steepest negative gradient reflect more routine activities and emphasize following directions, including "operate cash-register," "greeting customers," and "maintaining inventory."[12]

[12]For robustness, in Table B.5 we reproduce this table with six-digit SOC fixed effects. We also reproduce the table using the verb list from Michaels, Rauch, and Redding (2018); see Table B.6. Both robustness exercises reveal a similar pattern of increased interactiveness and teamwork in urban areas.

3.2 Technology Requirements Across Space

In the previous section, we showed that interactive tasks are more prevalent in urban environments. Past work, hypothesizing that human interaction is a key input in innovative activity, and drawing on patent data or occupational titles, has documented that new technologies are first introduced and adopted in cities (Carlino, Chatterjee, and Hunt, 2007; Lin, 2011). In this section, we systematically explore which technologies are more important in cities, and how this relationship varies with the human capital of workers.

We consider two questions: Are technological requirements more important in urban areas? And how does the technology gradient differ for jobs requiring a college degree compared to the gradient for jobs requiring a high school degree?

We measure the technology requirements of a job by searching for each of O*NET's "Hot Technologies." The list is originally derived from job postings and includes 180 different technologies.[13] As a preliminary exercise, we examine which technologies have the steepest positive gradient with respect to labor market size, and which have the steepest negative gradient. We estimate equation (1), replacing the dependent variable with $tech_{jn}^{(\ell)}$, an indicator for job ad j in market size decile n requiring technology ℓ. We run this regression for each of the 180 technologies, and sort by $\beta_{10}^{(\ell)}$, after normalizing the estimates by dividing by the standard deviation of $tech_{jn}^{(\ell)}$. The results are presented in Appendix B. The technologies with the steepest positive gradient with market size are Microsoft Excel, Python, Microsoft Project, and Linux. Separating the analysis by whether the job requires a college degree or a high school degree only, we find the positive gradient is steeper for jobs requiring a college degree. For college workers, the 15 technologies with steepest gradients have gradients 1.5 to 2 times as large as the 15 technologies with steepest gradients for non-college workers. Moreover, for college workers the technologies with the largest gradient in city size involve computer programming (e.g., Python, Javascript, Linux), while for non-college workers they involve data entry and word processing (e.g., the Microsoft Office suite) and social media (e.g., LinkedIn, Facebook).

Figure 3 presents a job-level regression of an indicator that any of the 180 technologies are a requirement, on CZ size deciles, controlling for log ad length. Panel I is without any occupation controls, and panel II includes six-digit SOC fixed effects. Panel I shows an increase of technological requirements with labor market size. Note that the technology gradient only appears for jobs requiring a college degree. Panel II shows that approximately

[13] We list the technologies in Appendix B.3; the list is also available on the O*NET website: https://www.onetonline.org/search/hot_tech/. The initial list is 182 technologies but we exclude R and C from our main analysis since they are likely to lead to false positives. Appendix B.5 reproduces the main analysis including R and C. The results are unchanged.

15 percent of the gradient remains after including six-digit SOC fixed effects. Once again, the gradient is present only for jobs requiring a college degree.

3.3 Specialization in Tasks Across Space

Economists since Adam Smith have attempted to understand the forces behind the productivity gains that come from a larger marketplace (Young, 1928; Stigler, 1951; Becker and Murphy, 1992; Garicano and Hubbard, 2009). Smith noted that larger markets allow for workers to specialize in narrower sets of activities and, as a result, become more productive. In this section, exploiting our granular task measures, we provide a new measure of worker specialization: the dissimilarity in tasks that workers perform relative to their peers within the same firm-market or occupation-market. We then demonstrate that this measure of specialization increases with market size.

To study specialization, we first need a notion of distance between jobs in task space. We characterize each job j as a vector of tasks, T_j, with each element corresponding to a distinct task. Each element takes a value of one if job ad j's description has the corresponding task, and zero otherwise. We normalize the task vectors to have unit length: $V_j = \frac{T_j}{\sqrt{T_j \cdot T_j}}$. The normalization ensures that our measures of specialization are unaffected by job ad length.[14]

The inner product between two jobs' task vectors is their cosine similarity, which takes a value between zero and one. Intuitively, if two jobs have perfect overlap in tasks, their similarity is one, and if they share no tasks in common, their similarity is zero. We define the task dissimilarity between jobs j and $j^{'}$ as one minus their cosine similarity: $d_{jj'} = 1 - V_j \cdot V_{j'}$.

Our notion of specialization within the firm-market is the average task dissimilarity between job j and other jobs in the firm-market pair. For this analysis, we denote a firm f as a firm name $\times$ six-digit industry NAICS code.[15] Define $d_{jfm} = 1 - V_{jfm} \cdot \overline{V}_{(-j)fm}$, where $\overline{V}_{(-j)fm}$ is the vector of average task content in firm-market fm, averaged over all jobs in the firm-market excluding job j. If the term d_{jfm} is larger, job j has less overlap in task content with other jobs in the firm-market. At the firm level, the degree of specialization is $d_{fm} = \frac{1}{n_{fm}} \sum_{j \in f,m} d_{jfm}$, where n_{fm} is the number of jobs in the firm-market. We emphasize that we cannot construct dissimilarity for all workers in the firm-market but only for vacancies, which captures newly-formed jobs.

[14] In constructing the firm-market sample, we drop ads that contain zero tasks, approximately 15 percent of ads, and ads that are singletons in the firm-market cell, another 4 percent. In constructing the occupation-market sample, the respective numbers are 17 percent and 0.11 percent.

[15] We group by both firm name and industry because the same firm name may, in certain cases, correspond to two separate firms in two different industries. Since these cases are rare, our results are essentially unchanged when grouping by firm name.

Note that we can define task dissimilarity more generally, $d_{jcm} = 1 - V_{jcm} \cdot \overline{V}_{(-j)cm}$, where c may represent job j's firm or its occupation. In our analysis we explore dissimilarity along these two dimensions. We estimate the following regression:

$$d_{cm} = \alpha_0 + \sum_{n=2}^{10} D_{cmn}\alpha_n + x_{cm}^{'}\delta + \epsilon_{cm}, \tag{2}$$

where d_{cm} is the mean task dissimilarity in group c and market m (where c refers to either firm or occupation), D_{cmn} is an indicator that market m is in size decile n, and x_{cm} are our main controls averaged to the group-market cell. In specifications where c refers to occupation, x_{cm} may also include occupation fixed effects.[16]

Figure 4 plots the estimates for α_n. The main result in panels I and II is that task dissimilarity in the firm is increasing in market size, with a steeper gradient for non-tradeable sector firms.[17] This result aligns with the classic theoretical point that the degree of specialization is limited by the extent of the market. Since the market for tradeable sector firms extends beyond their CZs, the gradient of specialization with respect to local market size will be flatter for workers in these sectors. Panels III and IV show that specialization within occupations is also increasing in market size.

In Appendix C.3, we study specialization across firms, in which the unit of analysis is the industry-market, and specialization is defined as dissimilarity in firm-level average tasks between firms in the same industry-market. Relative to firms in rural markets, firms in urban markets are located farther apart in task space.

All of these results together reveal that, as market size grows, there is an increase in both within- and between-firm specialization in tasks.

3.4 Tasks, Technologies, and Wages

In previous sections, we have documented that both technology usage and worker specialization increase with city size. In this section, we show that within-occupation differences in specialization and technology contents help account for the urban wage premium.

[16] In our analysis of specialization within occupations, we use four-digit (rather than six-digit) SOCs as our unit of analysis, to have more job ads in cells with which to calculate task dissimilarity.

[17] As a robustness exercise, we measure the degree of specialization without using the granular task measures and instead examine the distribution of common and rare occupations across space. Occupations that are rare as a share of the entire U.S. labor market make up a greater share of larger markets relative to smaller markets (see Appendix C.3). We also perform the exercise of Tian (2019), showing that in larger markets, there are more distinct job titles per firm, conditional on the number of ads posted by the firm.

We compute the mean task dissimilarity within each occupation-CZ pair,

$$d_{om} = \frac{1}{n_{om}} \sum_{j \in o,m} (1 - V_{jom} \cdot \overline{V}_{(-j)om}),$$

the mean number of technological requirements at the occupation-CZ, $tech_{om}$, and, using our ACS sample, the fraction of employed workers in the occupation-CZ with a BA or above, ba_{om}.

We run the following regression:

$$log(wage)_{om} = \gamma_0 + \gamma_1 d_{om} + \gamma_2 tech_{om} + \gamma_3 ba_{om} + \xi_o + \epsilon_{om}. \quad (3)$$

We include four-digit occupation fixed effects, ξ_o, in some specifications of equation (3) to highlight the role of tasks and technologies in accounting for within-occupation wage differences across markets. The coefficient γ_1 represents the relationship between task dissimilarity and wages. One should be cautious in interpreting γ_1 as causal, since, for example, workers may sort endogenously into occupations by unobservables in local labor markets that may correlate with market size. Nevertheless, it is valuable to assess whether within-occupation differences in tasks account for variation in wages across geography, beyond what is captured by differences in worker skills, and γ_1 is a key parameter for doing so. Similarly, γ_2 is informative about the extent to which technological requirements are important for accounting for within-occupation differences in wages across markets.

Table 1 reports the results. Column 1 shows that a one standard deviation increase in task dissimilarity is associated with an increase in wages by 6.9 percent, while a 10 percentage point increase in the fraction of jobs requiring a technology increases wages by 5.3 percent. Controlling for education in column 2 weakens the relationship somewhat, suggesting that the sorting of educated workers into more technical and specialized jobs explains part of the wage premium, but technologies and tasks remain economically and statistically significant. According to column 3 – with occupation fixed effects now included – variation in technologies and tasks explains a substantial proportion of wage variation.

Columns 4-6 re-estimate equation (3) separately by occupational category. We classify workers into white collar, blue collar, and service workers by two-digit SOC code, as described in the table note. Within occupation-CZ task dissimilarity plays an important role in accounting for the wage premium for white collar occupations, but not for blue collar occupations. Columns 4-6 also show that white collar workers have a large urban premium for technological requirements, while for blue collar workers the relationship is negative and small in magnitude.

Table 1 has several implications. First, across geography, within-occupation task and technological differences are strongly correlated with within-occupation wages, even after controlling for educational differences of workers.[18] Second, the correlation between task dissimilarity and wages is stronger for white collar occupations, while it is reduced for service occupations and disappears for blue collar occupations. These empirical findings suggest that the returns to employment in large cities differ for workers in blue and white collar occupations. They also suggest that the task content within occupations varies from rural to urban areas, which may limit the mobility of occupation-specific human capital.

4 Conclusion

By bringing in new data and incorporating tools from natural language processing, we examine in detail the differential task and technology content of urban and rural areas – one that captures heterogeneity within occupations. We also introduce an approach to define job tasks at a granular level, and use these to characterize the relation between market size and specialization, which is a key economic mechanism for productivity gains that has eluded direct measurement. We view our findings as critical for understanding questions such as what drives the urban wage premium, what are the limits to human capital mobility across regions, and how can policies enhance labor market fluidity.

[18] Appendix C.4 provides additional evidence that within-occupation task and technology differences, between rural and urban areas, account for variation in the urban wage premium. To briefly summarize, job ads' task and technology content explain 21 percent of the variation within occupations across markets of different population deciles. The relationship between job tasks and technologies and the urban wage premium is larger for white collar occupations.

References

Atalay, E., P. Phongthiengtham, S. Sotelo, and D. Tannenbaum (2018): "New Technologies and the Labor Market," *Journal of Monetary Economics*, 97, 48–67.

——— (2020): "The Evolution of Work in the United States," *American Economic Journal: Applied Economics*, 12(2), 1–34.

Autor, D., D. Dorn, and G. Hanson (2019): "When Work Disappears: Manufacturing Decline and the Falling Marriage-Market Value of Young Men," *American Economic Review: Insights*, 1(2), 161–178.

Autor, D. H. (2013): "The Task Approach to Labor Markets: An Overview," *Journal for Labour Market Research*, 46(3), 185–199.

Autor, D. H. (2019): "Work of the Past, Work of the Future," *AEA Papers and Proceedings*, 109, 1–32.

Bacolod, M., B. S. Blum, and W. C. Strange (2009): "Urban Interactions: Soft Skills Versus Specialization," *Journal of Economic Geography*, 9(2), 227–262.

Baum-Snow, N., M. Freedman, and R. Pavan (2018): "Why Has Urban Inequality Increased?," *American Economic Journal: Applied Economics*, 10(4), 1–42.

Becker, G. S., and K. M. Murphy (1992): "The division of labor, coordination costs, and knowledge," *Quarterly Journal of Economics*, 107(4), 1137–1160.

Bloom, N., T. Hassan, A. Kalyani, J. Lerner, and A. Tahoun (2020): "The Geography of New Technologies," Discussion paper, Institute for New Economic Thinking.

Carlino, G. A., S. Chatterjee, and R. M. Hunt (2007): "Urban density and the rate of invention," *Journal of Urban Economics*, 61(3), 389 – 419.

Davis, D., and J. Dingel (2019): "A Spatial Knowledge Economy," *American Economic Review*, 109(1), 153–70.

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. (2017): "The Growing Importance of Social Skills in the Labor Market," *Quarterly Journal of Economics*, 132(4), 1593–1640.

Duranton, G., and H. Jayet (2011): "Is the Division of Labour Limited by the Extent of the Market? Evidence from French Cities," *Journal of Urban Economics*, 69(1), 56–71.

Eckert, F., S. Ganapati, and C. Walsh (2019): "Skilled Scalable Services: The New Urban Bias in Recent Economic Growth," Discussion paper.

Frank, M. R., L. Sun, M. Cebrian, H. Youn, and I. Rahwan (2018): "Small Cities Face Greater Impact from Automation," *Journal of the Royal Society Interface*, 15(139).

Garicano, L., and T. N. Hubbard (2009): "Specialization, Firms, and Markets: The Division of Labor Within and Between Law Firms," *Journal of Law, Economics, and Organization*, 25(2), 339–371.

Garicano, L., and E. Rossi-Hansberg (2015): "Knowledge-Based Hierarchies: Using Organizations to Understand the Economy," *Annual Review of Economics*, 7(1), 1–30.

Giannone, E. (2019): "Skill-Biased Technical Change and Regional Convergence," Discussion paper.

Hemelt, S., B. Hershbein, H. Le, S. Martin, and K. Stange (2020): "The Skill Content of College Majors: Evidence from the Universe of Online Job Ads," Discussion paper.

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

Hoberg, G., and G. Phillips (2016): "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 124(5), 1423–1465.

Kim, S. (1989): "Labor Specialization and the Extent of the Market," *Journal of Political Economy*, 97(3), 692–705.

Lin, J. (2011): "Technological Adaptation, Cities, and New Work," *Review of Economics and Statistics*, 93(2), 554–574.

Michaels, G., F. Rauch, and S. J. Redding (2018): "Task Specialization in U.S. Cities from 1880 to 2000," *Journal of the European Economic Association*, 17(3), 754–798.

Modestino, A. S., J. Ballance, and D. Shoag (2020): "Upskilling: Do Employers Demand Greater Skill When Workers are Plentiful?," *Review of Economics and Statistics*, 102(4), 793–805.

Moretti, E. (2013): *The New Geography of Jobs*. Mariner Books/Houghton Mifflin Harcourt.

Ruggles, S., S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, and M. Sobek (2020): IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS.

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.

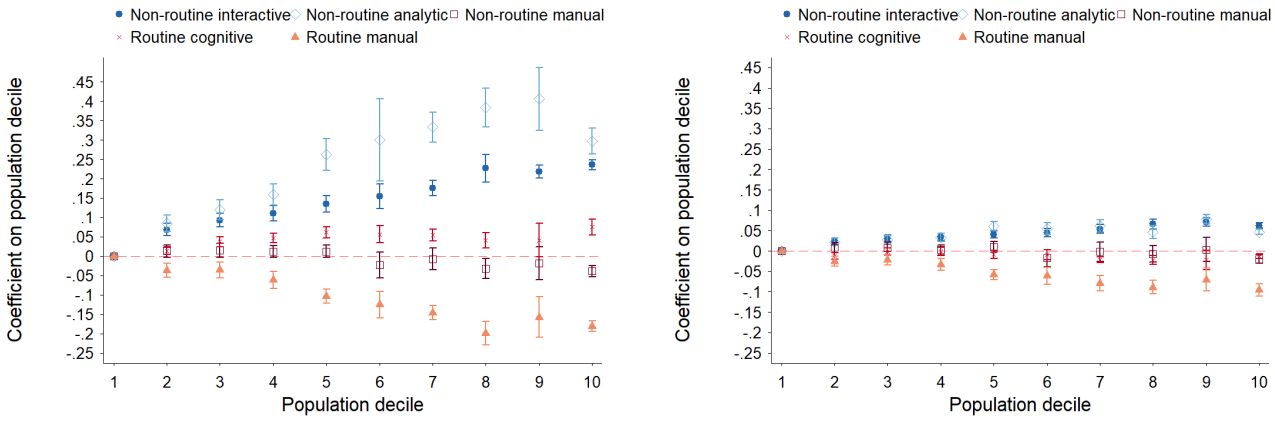
Stigler, G. J. (1951): "The Division of Labor is Limited by the Extent of the Market," *Journal of Political Economy*, 59(3), 185–193.

Tian, L. (2019): "Division of Labor and Productivity Advantage of Cities: Theory and Evidence from Brazil," Discussion paper.

Young, A. A. (1928): "Increasing Returns and Economic Progress," *Economic Journal*, 38(152), 527–542.

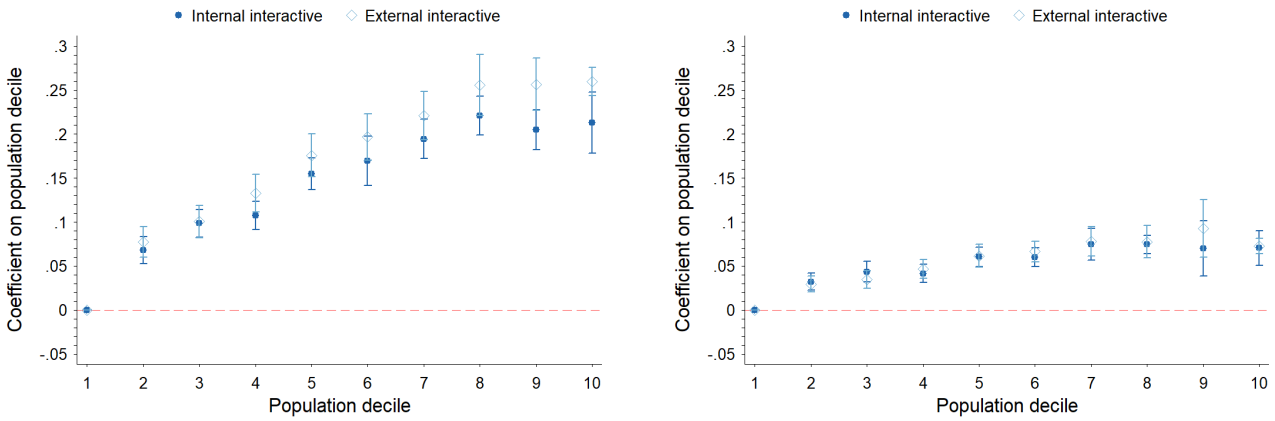
Figures and Tables

Figure 1: Tasks and Market Size



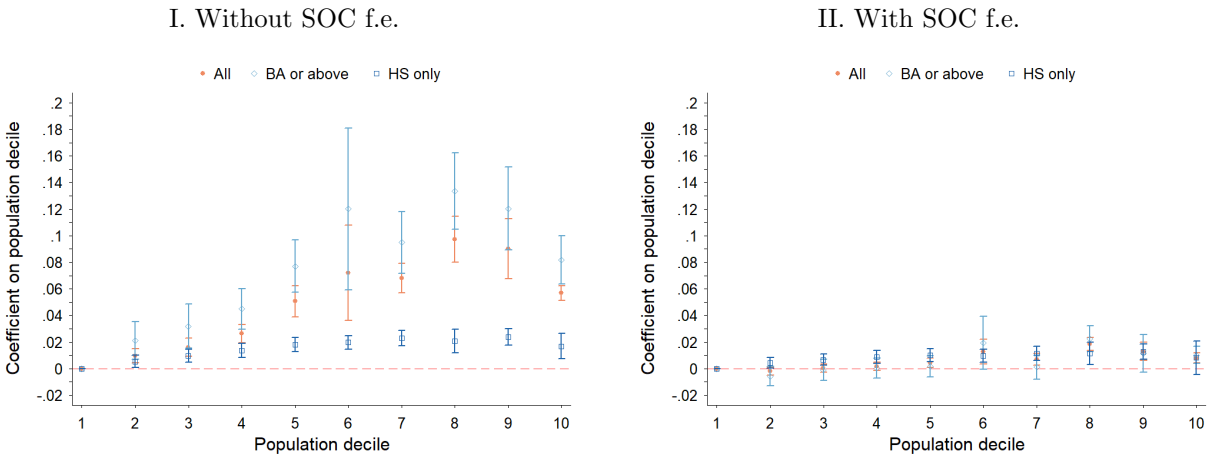
This figure presents estimates of equation (1), which depict the task gradient with market size. We control for log total ad words, and, in the right panel, six-digit SOC fixed effects. The dependent variable is task intensity.

Figure 2: O*NET Interactive Tasks Gradient



This figure presents the estimates from a regression at the job vacancy level of equation (1). We control for log total ad words, and, in the right panel, six-digit SOC fixed effects. The dependent variable is task intensity.

Figure 3: The Technology Gradient

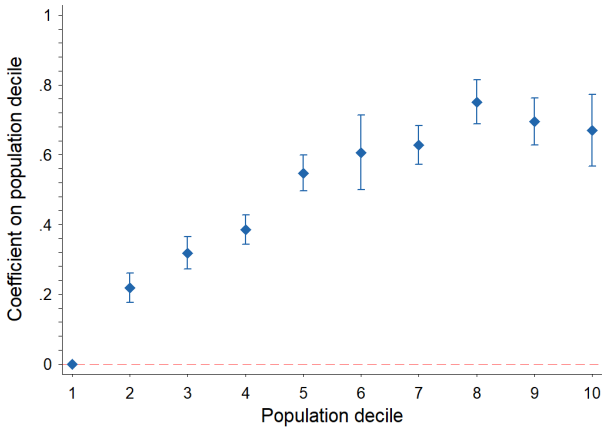


The dependent variable is an indicator for a job mentioning any of O*NET's "Hot Technologies," which is regressed on a vector of deciles for CZ. For reference, the first decile mean is 0.07 across all job ads, 0.14 for BA or above, and 0.07 for HS only. We control for log total ad words in the ad. Panel II includes six-digit SOC fixed effects. Standard errors are robust and clustered at the CZ level.

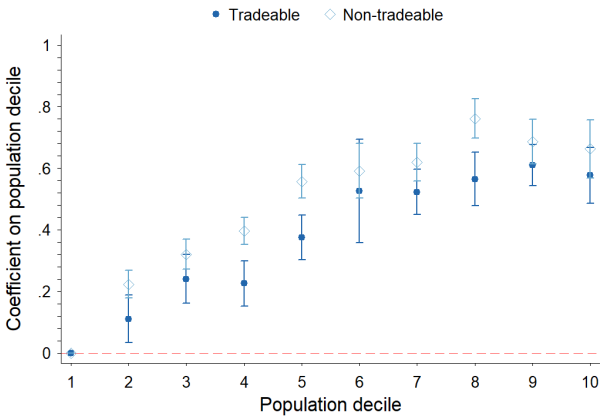
Figure 4: Specialization Gradient: Task Dissimilarity Within Firms and Occupations

A. Firms

I. All

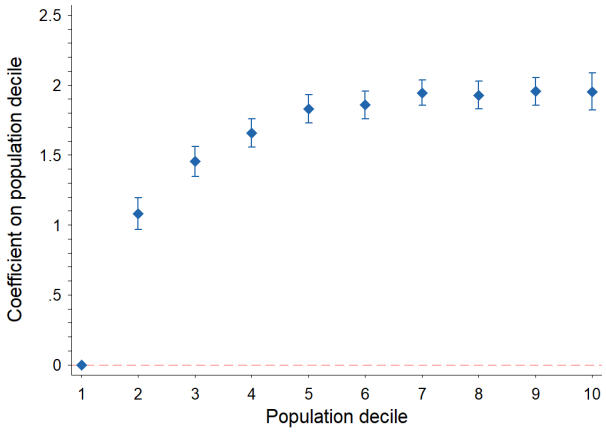


II. Tradeable v. Non-tradeable

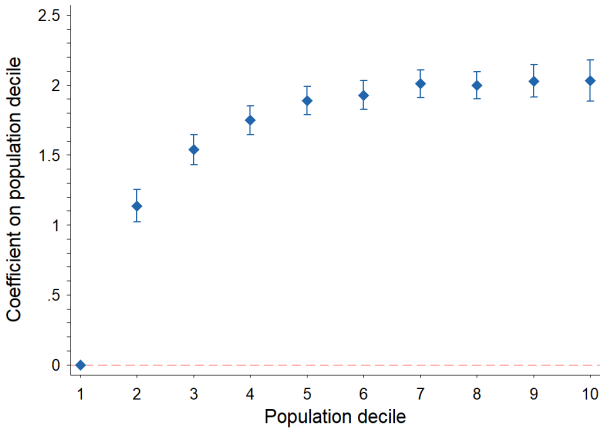


B. Occupations

III. Without SOC f.e.



IV. With SOC f.e.



The figures above present estimates of equation (2) and study how task dissimilarity within the firm (panel A) and within the occupation (panel B) vary with market size. Panel A uses the firm-market sample and the dependent variable is the mean task-dissimilarity in the firm-market, while panel B uses the occupation-market sample and the dependent variable is mean task dissimilarity in the occupation-market. We control for log total ad words, which is averaged to the cell level. Firm-market regressions are weighted by number of ads in the cell; occupation-market regressions are weighted by ACS employment in the cell. Standard errors are clustered at the CZ level. For reference, the first CZ decile mean for the top left panel is -0.51, and, for the top right panel, is -0.54 for the non-tradeable sample and -0.03 for the tradeable sample. The first CZ decile mean for the bottom two panels is -1.92. We define tradeable by two-digit NAICS code: agriculture, forestry, fishing and hunting (11), mining, quarrying, and oil and gas extraction (21), and manufacturing (31-33).

Table 1: Task Dissimilarity, Technologies, and Wages

	All			White collar	Blue collar	Service
	(1)	(2)	(3)	(4)	(5)	(6)
Task dissimilarity	0.069***	0.057***	0.030***	0.049***	0.006**	0.038***
	(0.004)	(0.003)	(0.002)	(0.004)	(0.003)	(0.004)
Technology requirements	0.527***	0.189***	0.165***	0.219***	-0.055**	0.064
	(0.012)	(0.008)	(0.025)	(0.028)	(0.027)	(0.046)
BA or above		1.067***	0.872***	0.941***	0.479***	0.639***
		(0.027)	(0.070)	(0.077)	(0.059)	(0.061)
SOC f.e.	No	No	Yes	Yes	Yes	Yes
Number of observations	45,156	45,156	45,156	24,447	11,338	9,130
R^2	0.140	0.536	0.839	0.819	0.580	0.670
Mean of dependent var.	10.769	10.769	10.769	10.968	10.561	10.224
Mean task dissimilarity	0.000	0.000	0.000	0.158	-0.169	-0.408
Mean technology requirements	0.247	0.247	0.247	0.322	0.129	0.106
Mean BA or above	0.363	0.363	0.363	0.518	0.076	0.132

The unit of observation is the occupation-market. The dependent variable is log wages, regressed on occupation-market task dissimilarity (normalized to have mean zero, standard deviation one across jobs), mean number of technologies, and fraction of workers with a BA or above, and, where indicated, four-digit SOC fixed effects. Regressions are weighted by employment. Standard errors are clustered at the CZ level. Occupations are classified into blue collar, white collar, and service occupations by two-digit SOC codes as follows. Blue collar: farming, fishing and forestry (45); construction and extraction (47); installation, maintenance and repair (49); production (51); and transportation and material moving (53). White-collar: management, business and finance (11–13); professional (15–29); sales (41); and office and administrative support (43). Service: healthcare support (31); protective service (33); food preparation and serving (35); building and grounds cleaning and maintenance (37); and personal care and service (39). Military (55) is excluded from columns 4-6.

A Validating the Online Job Ads Data

This section presents supplementary information and validation of the job ads data. Appendix A.1 provides summary statistics on the CZ deciles. Appendix A.2 provides details on the construction and cleaning of the sample used in the paper. Appendix A.3 discusses the representativeness of online vacancies relative to total vacancies as measured in JOLTS. In Appendix A.4, we show that when we create occupation-level task measures from the job ad text that correspond to O*NET task categories, these measures are highly correlated with O*NET importance scales. In Appendix A.5, we show that the education requirements in the job ads data correlate strongly with the education of employed workers in the ACS in the same occupation-market, and that this relationship holds across large and small markets, and within and between occupations. In Appendix A.6, we show that while there are trends in job ad length across space – larger markets have longer job ads – once we control for ad length the gradient of job description keywords with respect to market size becomes economically insignificant.

A.1 CZ Decile Summary Statistics

Table A.1 presents summary statistics by CZ decile, including the total number of job ads in the decile, the median CZ population, and the name(s) of the median population CZ(s) within the decile.

Table A.1: CZ Decile Summary Statistics

Decile	Total ads	Median CZ pop.	Median CZ name(s)
1	506.8	54.9	Norfolk & Madison Counties, NE
2	575.0	304.7	Jackson & Hillsdale & Lenawee Counties, MI; Bloomington, IN
3	595.2	609.6	Wichita, KS
4	599.8	1,033.4	Tulsa, OK; Naples-Marco Island, FL
5	732.3	1,639.0	Nashville-Davidson-Murfreesboro, TN
6	692.3	2,441.2	St. Louis, MO
7	705.1	3,453.2	Minneapolis-St. Paul, MN; Hartford-Bridgeport-Stamford-Norwalk, CT
8	858.4	5,056.6	Atlanta, GA
9	685.3	6,159.5	Newark-Trenton-White Plains NJ-NY; Houston, TX
10	385.4	15,273.6	New York, NY; Los Angeles, CA

The table above presents summary statistics by CZ decile, including the total number of job ads in the decile (expressed in 1,000s), the median CZ population in the decile (in 1,000s), and the name(s) of the median population CZ(s) within the decile. In cases where the median CZ population is the average of two CZs, we provide both names separated by a semicolon.

A.2 Details on Sample Construction

We use a five percent sample of the online job ads data that we purchased from EMSI. The sample of our dataset covers January 2012 to March 2017. We exclude ads with fewer than the first percentile number of words and greater than the 95th percentile number of words. These restrictions ensure that the ads have enough content to measure tasks and also are not so long as to considerably slow down processing time. This step limits the sample to ads with length between 11 and 841 words and reduces the sample to 7.0 million ads. We exclude Hawaii and Alaska from the analysis, which drops another 35,529 ads. We also exclude ads that do not contain a county FIPS code, and therefore cannot be mapped to a CZ. This step drops another 503,051 ads. Finally, we drop ads that have no SOC code, another 102,154 ads. This leaves 6.3 million ads for our occupational analysis. Table A.2 presents the number of ads by year in the sample.

For the firm-level analysis sample, we make a few additional restrictions. We drop ads placed by staffing or placement agencies, since they act as intermediaries between the worker and firm hiring the worker. This step drops 596,578 ads.[19] We drop ads without a firm name, which is another 107,317 ads. Finally, we drop firms with a missing NAICS code, another 3,771 ads. These restrictions yield approximately 5.6 million ads for the sample used for the firm-level analysis.

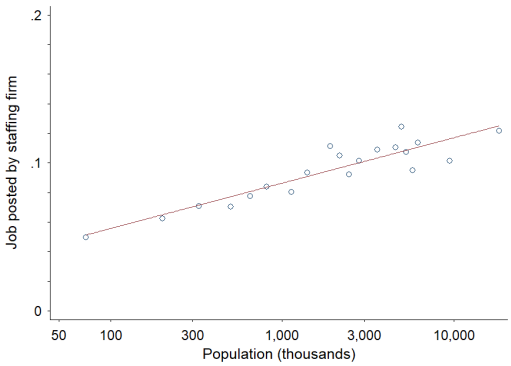
Table A.2: Job Vacancy Counts by Year

Occupation-level dataset		Firm-level dataset	
Year	Count	Year	Count
2012	591,682	2012	504,618
2013	860,961	2013	751,387
2014	1,021,805	2014	904,882
2015	1,465,475	2015	1,327,579
2016	1,905,368	2016	1,709,801
2017	490,287	2017	429,645
Total	6,335,578	Total	5,627,912

The table above presents the number of job ads by year after making the sample restrictions discussed in Appendix A.2.

[19] Figure A.1 presents a binscatter of an indicator for the job ad being posted by a staffing firm, against CZ population. The figure shows a slight positive gradient with market size.

Figure A.1: Job Posted by a Staffing Firm



This figure presents a binned scatterplot of an indicator for the job ad being posted by a staffing firm on log population at the CZ-level.

A.3 Representativeness of Online Vacancies

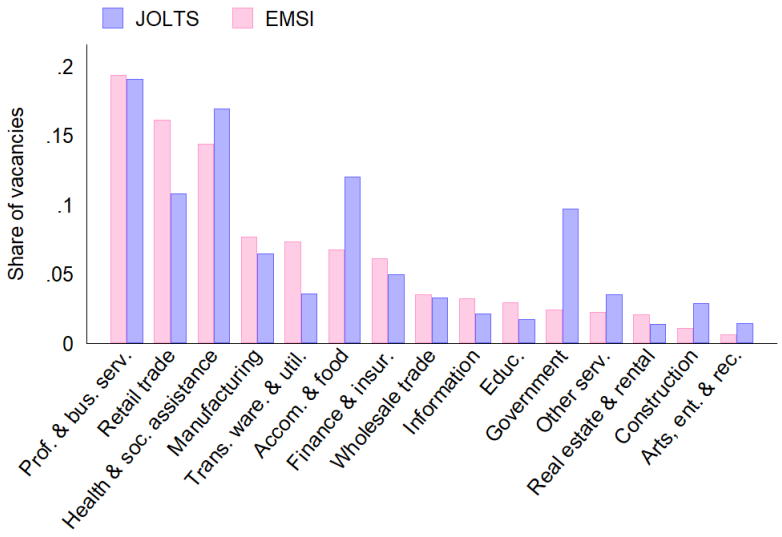
The standard resource for measuring job vacancies in the U.S. is the Job Openings and Labor Turnover Survey (JOLTS), conducted by the Bureau of Labor Statistics of the U.S. Department of Labor. The dataset consists of monthly job openings at the national level by major industry category.[20] JOLTS is based a survey of a random subset of establishments covered by state or federal unemployment insurance laws.[21]

Figure A.2 plots the distribution of job ads by sector for JOLTS and EMSI. Certain industries, such as Manufacturing, Finance and Insurance, and Education have higher representation in EMSI compared to JOLTS, while others, such as Health and Social Assistance, Government, and Accommodation and Food, have higher representation in JOLTS.

[20] The JOLTS dataset also has vacancies at the census region level, but not at the region-by-industry level. JOLTS has no finer geographic unit than census region.

[21] JOLTS defines job openings as "positions that are open (not filled) on the last business day of the month. A job is 'open' only if it meets all three of the following conditions: (1) A specific position exists and there is work available for that position. The position can be full-time or part-time, and it can be permanent, short-term, or seasonal; (2) The job could start within 30 days, whether or not the establishment finds a suitable candidate during that time; (3) There is active recruiting for workers from outside the establishment location that has the opening."

Figure A.2: Distribution of EMSI Job Ads v. JOLTS



This figure plots the distribution of EMSI job ads and JOLTS job openings across major industries, from 2012-2017. The industries are sorted on the x-axis by their share of job ads in EMSI.

A.4 Measuring Occupational Tasks

This section provides additional details on how we measure jobs' task content. These measures correspond to those used in past research: Spitz-Oener (2006) and the O*NET database. We then compare occupations' task content -- according to these measures -- using the EMSI dataset with measures directly observed in the O*NET database. These two sets of measures align, validating our use of the EMSI dataset.

Mapping Words to Tasks

We map job description words to the five Spitz-Oener (2006) task categories: non-routine analytic, non-routine interactive, non-routine manual, routine cognitive, and routine manual. We use the word-to-task mappings we develop in Atalay, Phongthiengtham, Sotelo, and Tannenbaum (2020). These mappings are available on our project website: https://occupationdata.github.io/. We use the continuous bag of words model list of word mappings, which is described in detail in the data documentation on the website.

Comparing Tasks from Job Ads to O*NET

A key limitation of O*NET is that it only measures tasks at the occupation level. Hence, O*NET is unable to speak to geographic variation in tasks aside from those arising from different employment shares across regions. Nevertheless, O*NET is valuable for testing the validity of our job ads for extracting occupation-level tasks. We construct occupation-level task content using the EMSI ads data and plot the correlation with O*NET's Work Activities.

The specific tasks we compare are O*NET's "Selling or Influencing Others," "Communicating with Persons Outside Organization," "Guiding, Directing, and Motivating Subordinates," "Developing and Building Teams," "Coaching and Developing Others," "Coordinating the Work and Activities of Others," and "Communicating with Supervisors, Peers, or Subordinates." We adopt the mapping of words to O*NET Work Activities listed below.[22] Note that this mapping is necessarily somewhat ad hoc. We count, for each ad, the total number of occurrences of any of the corresponding words. We then normalize the count so that it is expressed per 1,000 job ad words. The first two bullet points refer to interactive tasks that are external to the firm; the remaining five refer to internal interactive tasks.

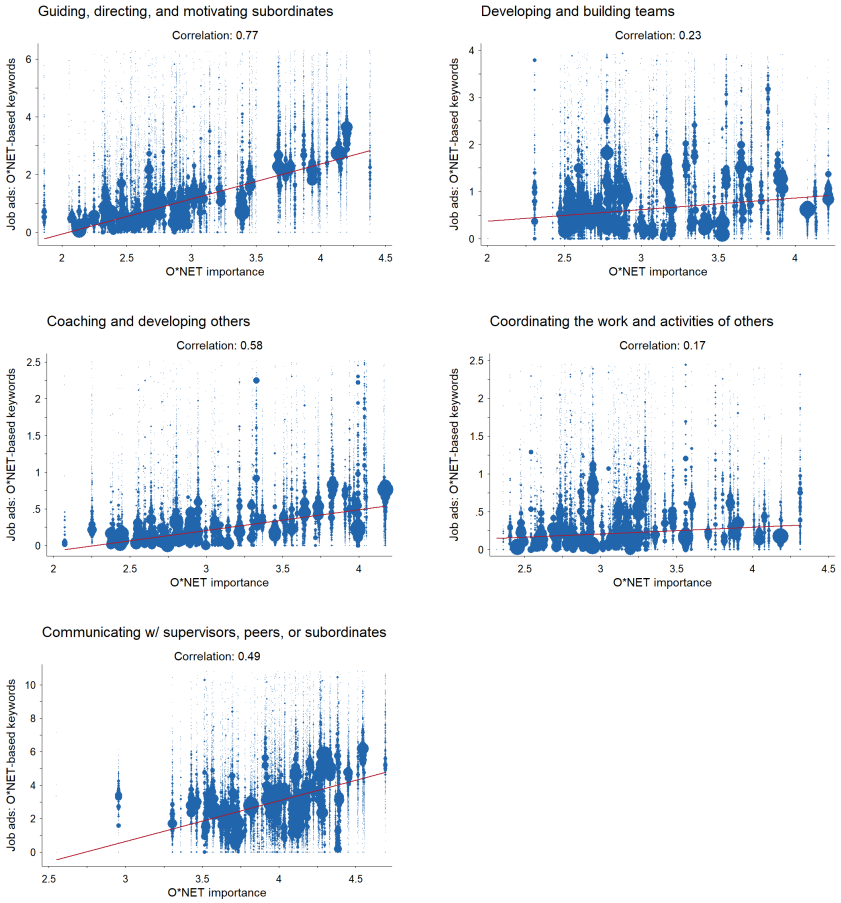
- *Selling or Influencing Others*: sales marketing advertising advertise merchandising promoting telemarketing market plan
- *Communicating with Persons Outside Organization*: clients client vendor vendors public interface communicate communication communicating coordinating conferring public relation
- *Guiding, Directing, and Motivating Subordinates*: directing direction guidance leadership motivate motivating motivational subordinate supervise supervising
- *Developing and Building Teams*: team-building "team build" project leader
- *Coaching and Developing Others*: mentor mentoring coaching
- *Coordinating the Work and Activities of Others*: coordinate coordination coordinator
- *Communicating with Supervisors, Peers, or Subordinates*: peer subordinate subordinates supervisor supervisors manager managers interface communicate communication communicating coordinating conferring

[22] We count instances of each word separately, for example, "public" and "relations" are searched for separately rather than as the bigram "public relations." We make one exception for "team build" because in our judgment "build" on its own is likely to return false positives. In Atalay, Phongthiengtham, Sotelo, and Tannenbaum (2020) and in the word mappings on our project website, some task-related words are bigrams.

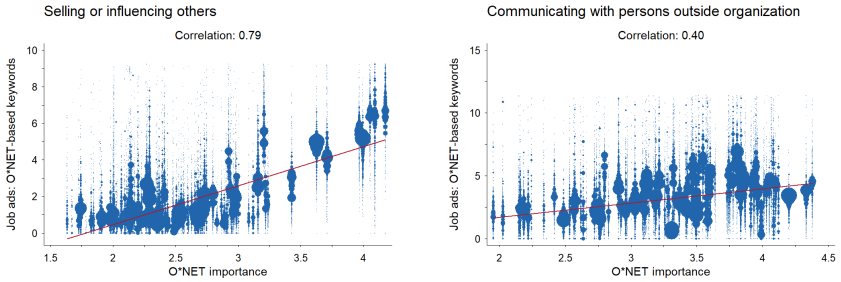
Figure A.3 demonstrates that our job ad-based task data have, for the most part, a high degree of correlation with O*NET tasks. We should not expect a perfect correlation, as O*NET itself has well-known limitations of small sample sizes, status quo bias, and subjective scales (Autor, 2013). But these correlations indicate that the job description text provides meaningful information about the task content of occupations.

Figure A.3: Comparing Tasks from Job Ads to O*NET

I. Internal Interactive Tasks



II. External Interactive Tasks



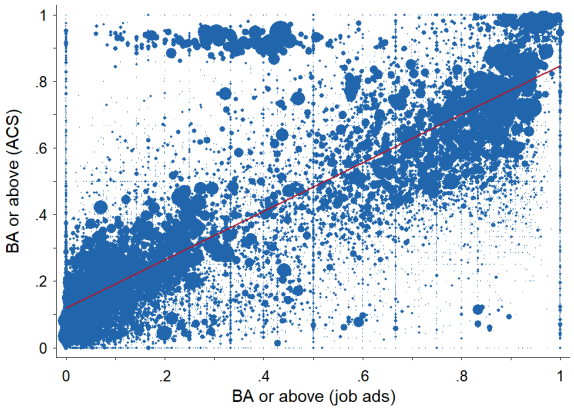
The figures above plot the correlations between occupation-level tasks extracted from the job ads to those based on from O*NET. Each dot represents a four-digit SOC × CZ. The correlations are weighted by ACS employment. (The figures exclude task intensities over the 99th percentile in both the reported correlations and the scatterplots.)

A.5 Education Requirements: Job Ads v. ACS Employment

In this section, again with the aim of validating the EMSI dataset, we compare education levels across occupations and commuting zones. For each six-digit SOC × CZ, we compute the fraction of job ads requiring a BA degree or above (among ads mentioning an educational requirement), and the fraction of employed workers, measured in the ACS, with a BA degree or higher. Figure A.4 correlates these two measures, with weights for employment in the cell. There is a strong correlation, suggesting that job ads contain valuable information about the educational requirements of the occupation. The share of ads with a given educational requirement is somewhat greater than the corresponding share of workers with that level of educational attainment. This result is perhaps unsurprising, given that job vacancies represent the frontier of occupational change, and the supply of educated workers has increased over time. Figure A.5 plots the same regression by CZ population quartile, showing a strong correlation for both large and small labor markets.

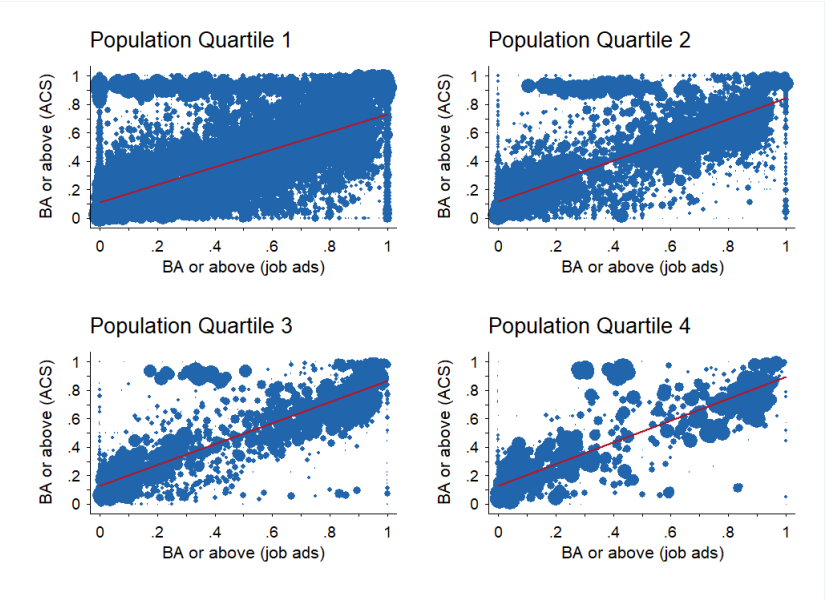
Using the same data, Figure A.6 depicts the gradient of educational requirements across CZ population deciles for the job vacancy data, and, next to it, the gradient of educational attainment of employed workers in the ACS. The gradient looks remarkably similar, both within and across occupations, suggesting again that the job vacancy data are picking up meaningful variation in the educational requirements of jobs across geography.

Figure A.4: Education Requirements in ACS v. Job Ads



Each dot in the figure above corresponds to a four-digit SOC × market. The cells are weighted by employment. The y-axis corresponds to the fraction of workers in the ACS with at least a college degree. The x-axis corresponds to the fraction of job ads that require a BA degree or higher (among ads that mention any education requirement).

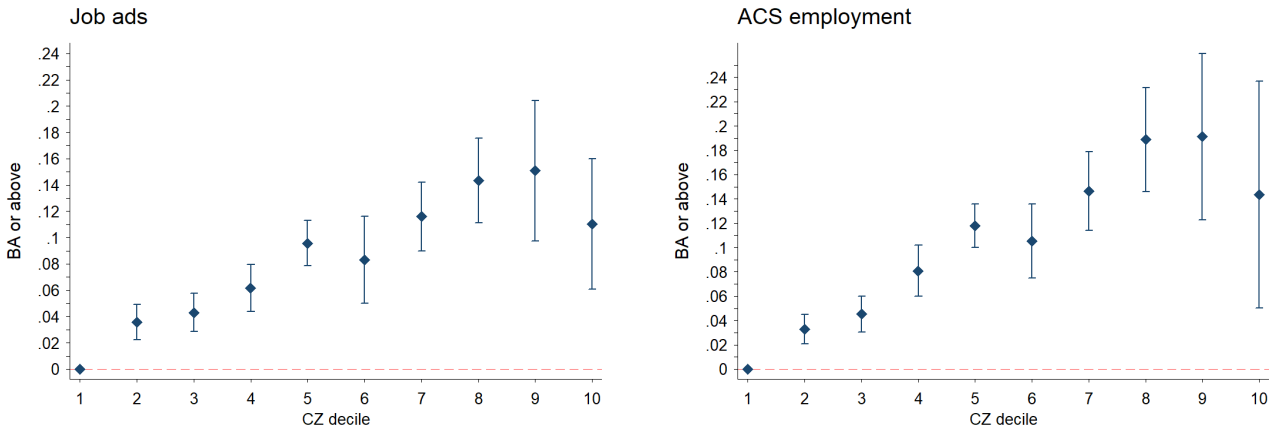
Figure A.5: Education Requirements in ACS v. Job Ads



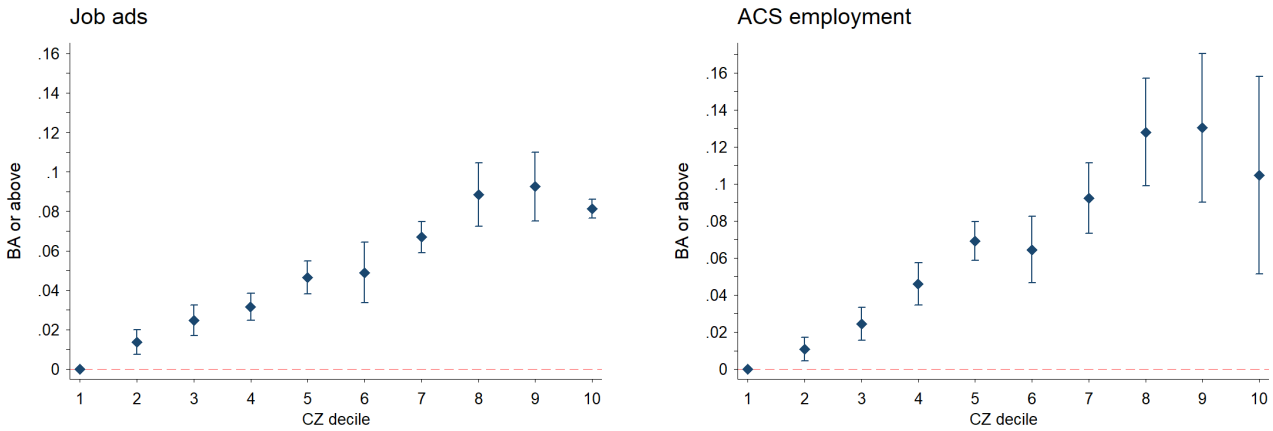
The figure above replicates Figure A.4 separately by CZ population quartile.

Figure A.6: Education Gradient with Market Size: ACS v. Job Ads

I. Without SOC f.e.



II. With SOC f.e.



The top left panel plots the coefficients in a regression of the fraction of job ads having an education requirement of a BA or above (conditional on having an educational requirement) on dummies for CZ decile, in an occupation-market cell. The cells are weighted by employment, and standard errors are clustered at the CZ level. The top right panel plots the same regression except where the dependent variable is the fraction of employed workers with a BA or above. The bottom two panels reproduce the top two panels with four-digit fixed effects.

A.6 Job Ad Length and Description Keywords Across Space

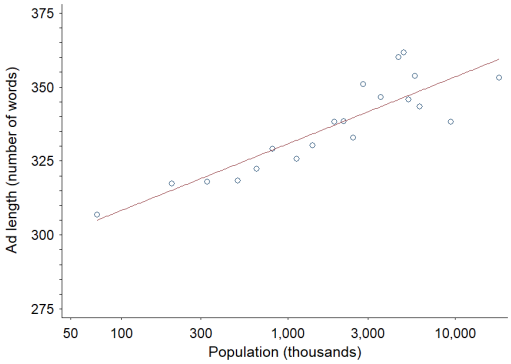
We next consider the content of the job ads and how it differs across geography. First, we plot a binned scatterplot of job ad length (i.e., the number of words) against the log CZ population (Figure A.7). This exercise shows that larger markets have longer job ads on

average. Motivated by this pattern, we control for job ad length throughout our analysis and standardize our task measures to be per 1,000 ad words, and normalize our granular task measures so that each task vector has unit length.

In Appendix B.1, we describe the approach to extracting job tasks from the text. The first step is to first identify the part of the text corresponding to the job description. We use a set of keywords to identify this portion of the ad. Figure A.8 examines the gradient of the job ad containing one of these keywords with market size, after controlling for ad length. The left panel shows a negligible relationship between market size and the presence of a keyword.

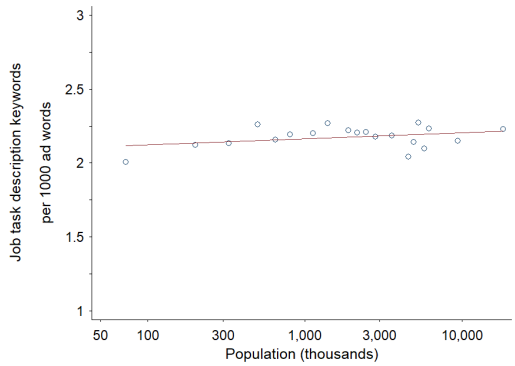
Lastly, we show that our novel task-extraction methodology – using job descriptions and parts of speech to let the text define the job tasks – passes a simple validation check. We calculate the cosine similarity between each job and the occupation-market average, and then take the average. This exercise reveals that similarity is higher for more narrowly defined occupational categories. Specifically, the cosine similarity is 0.052 for two-digit SOCs, 0.072 for four-digit SOCs, 0.104 for six-digit SOCs, and 0.166 for job titles. Thus, the text-based tasks of occupations are more similar within more narrowly defined occupational categories. It is perhaps unsurprising that narrower occupational categories share more job ad words, but this finding is reassuring and suggests the text contains valuable information about occupational characteristics that is reflected in standard occupational classifications.

Figure A.7: Job Ad Text Across Geography



The figure above presents a binned scatterplot of job ad length (number of words) on log population at the CZ-level. Cells are weighted by the number of job ads in the cell.

Figure A.8: Job Description Keywords Across Geography



The left panel above presents a binned scatterplot of an indicator of the job ad having a keyword in our task-extraction algorithm – "responsibilities," "duties," "summary," "tasks" – normalized per 1,000 ad words, against log CZ population.

B Task Extraction and Validation

This section outlines our approach to measuring job tasks. We first describe and illustrate the procedure for extracting job tasks from the text (Appendix B.1); present the most common tasks and technologies (Appendices B.2 and B.3); evaluate the relationships among tasks, technologies, and market size (Appendices B.4 and B.5); and then show that these tasks account for variation in wages across geography, above and beyond what is captured by occupational codes (Appendix B.6).

B.1 Extracting Job Tasks from the Raw Text

We first use the job ad text to generate a list of job tasks, which we call the *vocabulary of tasks*. Once we have the task vocabulary, we represent each job ad as a vector, of which each element corresponds to a distinct task.

We define a task as a verb-noun pair. There are two main steps to extracting verb-noun pairs from the text:

1. We first isolate the section of the text that pertains to job tasks. To do this we search for keywords in the text that suggest a list of tasks will follow. The keywords we use are "duties," "summary," "description," "tasks." We isolate the section of text that begins

at one of these keywords and ends at the next period.[23]

2. Using the section of text extracted from step 1, we find all (verb stem, noun stem) pairs, which will make up our task vocabulary. Examples of pairs include "assist customers" and "provide advice." Since verbs and nouns are stemmed, "writing memo" and "writes memos" are recorded as the same task. This step works as follows: we extract each verb and the noun that appears next in the sentence. Hence if the job ad says "writing lucid memos to prepare for depositions" or "writes legal memos for court hearings," these will both be recorded by our algorithm as "writes memos."[24] If multiple verbs correspond to the same noun (for instance, "prepares and revises memos" our algorithm extracts two tasks: "prepares memos" and "revises memos."[25] Since there are many tasks that appear only once in the data, we require the task to appear in at least ten ads, to limit the set of tasks in the vocabulary.

Once we have the vocabulary of tasks, according to steps 1 and 2 above, we vectorize all job ads according to the task vocabulary created in step 2. Hence, we are not limiting our analysis to ads with keywords described in step 1. We represent each ad as a vector, in which each element of the vector corresponds to a particular task in the vocabulary and takes a value of one if the job ad has that particular task and zero otherwise.

B.2 Task List

Below we list the 430 tasks that we extract from the job ad text as verb-noun pairs along with the fraction of ads with each task ($\times$ 100).

Tasks Extracted from Verb-Noun Pairs

written communication	13.0257	working store	0.9374	bagging merchandise	0.3460
perform duties	9.4926	meet business	0.9364	handling cash	0.3437
working team	7.4251	providing product	0.9180	procedures cash	0.3257
provide customer_service	6.6934	using equipment	0.9115	using eye	0.3249

[23]The purpose of this step is to eliminate portions of the job ad that refer to worker skills or firm characteristics. This step significantly improves the precision of the task extraction. Note that not all ads will have these keywords, and hence an important check is whether the presence of these words varies systematically from rural to urban labor markets. Figure A.8 investigates this relationship and finds little evidence for a systematic pattern. In step 2, when we vectorize all job ads based on the task vocabulary created in this step, we do not restrict the data to jobs that include these keywords. Also, in step 2, we perform the vectorization on all ad text, not just the portion of text that follows a keyword.

[24]We do not group synonyms, so "write memo" and "draft memo" are recorded as distinct tasks.

[25]We do not do the analogous procedure when a verb is followed by a list of nouns (for instance, "writes memos, opinions, and letters"); in this situation, our algorithm extracts one task – the verb and the first noun ("writes memos").

be part	5.6583	protect company	0.8972	taking vehicle	0.3210
provide service	5.3395	carry pounds	0.8943	maintained times	0.3133
lifting pounds	4.6136	ensuring merchandising	0.8941	damaged merchandise	0.3108
providing support	4.4229	following policies	0.8890	move trays	0.3104
work environment	4.2684	ensure operation	0.8781	needed customer_satisfaction	0.3092
perform functions	4.2277	responding customer	0.8579	increase customer_satisfaction	0.3044
build relationships	3.8635	ensure service	0.8539	following pogs	0.3041
ensure compliance	3.5870	including cash	0.8443	responsibilities duties	0.3031
performing tasks	3.5125	developed sales	0.8352	document counts	0.3024
assisting customers	3.2288	communicate information	0.8348	assigned skills	0.3022
provide customer	3.1077	closes store	0.8229	may store	0.2908
maintaining relationships	3.0468	developing strategies	0.8218	leads customers	0.2905
problem_solving skills	2.9784	working sales	0.8212	maintaining program	0.2901
making decisions	2.9349	writing skills	0.8198	are reporting	0.2875
required knowledge	2.9310	answering phones	0.8154	executes store	0.2866
ensure customer	2.8990	increase sales	0.8052	according needs	0.2864
lift lbs	2.8608	be lbs	0.8021	supporting activities	0.2829
provides quality	2.8342	maintaining environments	0.8014	lead store	0.2827
provides leadership	2.5047	handle tasks	0.7909	serving quality	0.2689
develop relationship	2.5011	support business	0.7870	include staff	0.2668
perform job	2.4971	are manages	0.7822	maintain pharmacy	0.2627
have experience	2.4283	ensure adherence	0.7739	remove items	0.2540
playing role	2.3877	require walking	0.7711	requiring security	0.2536
leading team	2.3856	ensure employees	0.7655	required paperwork	0.2522
are position	2.3332	working variety	0.7644	include hand	0.2513
achieve goals	2.2844	assume responsibilities	0.7592	seek customer	0.2444
working relationships	2.2757	ensure completion	0.7577	lifting merchandise	0.2430
continuing education	2.1940	maintain productivity	0.7455	promote shopping	0.2401
serving customers	2.1819	are duties	0.7342	merchandising product	0.2349
have years	2.1553	identifies problems	0.7329	scheduling activities	0.2295
following company	2.1392	asking questions	0.7320	set displays	0.2265
required qualifications	2.1347	include service	0.7303	has client	0.2240
providing care	2.0627	providing environment	0.7301	stored areas	0.2206
make recommendations	2.0457	writing reports	0.7265	maintain card	0.2199
meet requirements	2.0141	managing operations	0.7249	training sessions	0.2183
meet deadlines	1.9775	including training	0.7245	conducting employee	0.2130

provides training	1.9577	providing expertise	0.7104	evaluates employees	0.2116
provided information	1.8973	ensure client	0.7027	include shelves	0.2112
will customers	1.8947	assigned store	0.6921	permitted law	0.2062
resolve issue	1.8601	maintain communication	0.6920	using phone	0.2054
work flexible_schedule	1.8575	assist development	0.6902	vacuum face	0.2037
demonstrate knowledge	1.8571	generate sales	0.6839	assigns directs	0.2007
taking actions	1.8503	working departments	0.6815	using greet	0.1836
provide feedback	1.8131	using knowledge	0.6813	discontinued items	0.1835
provide assistance	1.8073	include development	0.6663	using orders	0.1808
providing solutions	1.8068	answering telephone	0.6570	outdated merchandise	0.1800
driving sales	1.7791	develop productivity	0.6569	prepare returns	0.1797
ensure quality	1.7532	developing implement	0.6548	greeting card	0.1794
helping customer	1.7479	established guidelines	0.6539	work stock	0.1765
works custom	1.7189	maintain work_environment	0.6482	securing company	0.1763
communicate customer	1.6945	preparing foods	0.6481	crews customer_service	0.1761
follow instructions	1.6791	existing clients	0.6366	recalled merchandise	0.1759
managing projects	1.6743	ensure guests	0.6231	crew directing	0.1758
maintain store	1.6554	including work	0.6221	change bulbs	0.1738
is service	1.6483	maximizes profitability	0.6159	labeling prescriptions	0.1735
greeting customers	1.6384	required driver	0.6138	maximizing customer_satisfaction	0.1723
work shift	1.6339	provide client	0.6136	needed in_store	0.1708
will teams	1.6264	meet clients	0.6114	reset departments	0.1703
answer questions	1.6252	set goals	0.6112	return system	0.1703
ensure product	1.6196	including business	0.6068	signing maintain	0.1701
provide guidance	1.6020	are compliance	0.6046	preventing trafficking	0.1699
detail ability	1.5925	move store	0.6043	windows ceilings	0.1698
includes ability	1.5900	provide technical_support	0.6015	windows removal	0.1690
maintaining inventory	1.5885	provide recommendations	0.5896	sweeping stock	0.1688
include sales	1.5879	opens store	0.5815	signing shelves	0.1688
written skills	1.5729	obtain information	0.5811	dump baskets	0.1688
preferred knowledge	1.5285	ensuring team	0.5669	photofinishing orders	0.1688
work schedule	1.5256	is walks	0.5581	regarding cash_register	0.1688
achieving sales	1.5248	assigned supervisor	0.5577	bags counter_tops	0.1687
resolve problems	1.5085	requires merchandise	0.5567	measuring drugs	0.1684
stand periods	1.4931	managing sales	0.5564	putting drug	0.1682
committed diverse	1.4668	include design	0.5528	seal trays	0.1682

maintaining standards	1.4602	hiring training	0.5491	capping vials	0.1679
assist store	1.4362	ensure projects	0.5474	closing duties	0.1672
meets customer	1.4272	will career	0.5421	make offer	0.1641
requires travel	1.4230	conducting research	0.5416	ensures quality_assurance	0.1606
work others	1.4230	assisting clients	0.5355	following reports	0.1567
work week_ends	1.4150	assisted sales	0.5328	communicating field	0.1554
written instructions	1.3752	maintain awareness	0.5270	execute cash	0.1530
operating cash_register	1.3735	include knowledge	0.5175	returned check	0.1492
resolving customer	1.3628	reaching pulling	0.5157	following vendor	0.1492
develop business	1.3594	traveling store	0.5122	execute display	0.1459
maintain working	1.3569	unloading trucks	0.5120	request help	0.1459
maintain knowledge	1.3533	move merchandise	0.5054	including translation	0.1426
providing direction	1.3523	develop test	0.5026	appropriate use	0.1422
are sales	1.3488	including performance	0.4901	perform register	0.1418
establish relationships	1.3468	including maintenance	0.4849	opening duties	0.1410
perform variety	1.3458	supervising store	0.4845	executing set	0.1401
ensure safety	1.3232	guided values	0.4785	sustained work	0.1397
handling customer	1.3140	ensuring food	0.4728	pay policy	0.1393
interact customers	1.3129	handle merchandise	0.4725	securing door	0.1390
achieve results	1.3078	build customer	0.4707	execute completion	0.1379
exceed sales	1.3000	make adjustments	0.4695	pay vendors	0.1377
ensure stores	1.2915	include merchandising	0.4597	checking employee	0.1375
developing team	1.2807	manages business	0.4588	check_in merchandise	0.1374
develop solutions	1.2723	taking orders	0.4545	check acceptance	0.1371
preferred ability	1.2457	ensuring communications	0.4525	skating carhop	0.1368
using computer	1.2323	including systems	0.4524	maintain prescription	0.1365
maintain appearance	1.2284	meets standards	0.4505	sustained periods	0.1365
identify opportunities	1.2281	manage relationships	0.4499	pulls deposits	0.1360
weighing pounds	1.2267	including preparation	0.4490	apprehend company	0.1358
growing business	1.2217	ensure policies	0.4467	document cash	0.1356
make changes	1.2214	comply state	0.4383	adapting store	0.1355
maintain custom	1.2155	include program	0.4380	secure change	0.1352
completing tasks	1.2050	ensure restaurant	0.4377	identify shoplifters	0.1350
existing customers	1.1991	may merchandise	0.4361	react program	0.1350
on_going training	1.1942	may floor	0.4279	in_store repairs	0.1350
performing work	1.1759	put customer	0.4249	resolve rejections	0.1350

including nights	1.1743	scheduling appointments	0.4193	organized pharmacy	0.1348
work projects	1.1730	assisting team	0.4184	signing crew	0.1348
develop planning	1.1620	providing coaching	0.4137	react shoplifters	0.1347
stand walk	1.1526	have merchandise	0.4125	using enhancements	0.1346
maximize sale	1.1489	including support	0.4115	execute walk_through	0.1346
sells products	1.1478	causing discomfort	0.4102	intern communication	0.1344
written oral_communication	1.1286	provides performance	0.4035	according hipaa	0.1344
ensure customer_satisfaction	1.1274	processing transactions	0.4030	locking setting	0.1340
operate equipment	1.1250	offer products	0.3978	sweep room	0.1339
meet goals	1.1221	include client	0.3976	adjust facings	0.1335
use hands	1.1209	containing materials	0.3974	trash rest	0.1335
analyzing data	1.1207	may slippery	0.3958	bulletins action	0.1335
meet sales	1.1067	maintain area	0.3946	dcr photofinishing	0.1335
prepare reports	1.1062	receives service	0.3945	maintain pull	0.1335
assigned management	1.1047	transforming delivery	0.3921	comply cvs	0.1332
knowledge skills	1.1011	maintain files	0.3918	pharmacist communicate	0.1331
according company	1.0815	become slippery	0.3917	needed inventory_management	0.1330
including management	1.0743	causing walking	0.3916	according cvs	0.1330
engage customers	1.0722	causing drafts	0.3916	cvs workflow	0.1330
provides input	1.0682	appear floor	0.3915	greeting operations	0.1274
perform maintenance	1.0614	floors work	0.3912	sorting merchandise	0.1226
prioritize tasks	1.0197	passing emit	0.3910	delegated photo	0.1214
managing teams	1.0034	include customer_service	0.3894	merchandising directives	0.1102
ensure accuracy	1.0017	focus team_work	0.3883	preventing terrorists	0.1075
working business	1.0001	as_needed assist	0.3864	supervisor team	0.0957
improving quality	1.0000	retrieving information	0.3735	driving culture	0.0908
team members	0.9907	assist staff	0.3715	drive_in employees	0.0902
establish policies	0.9903	maintaining business	0.3691	identifying conditions	0.0699
assisting management	0.9799	include order	0.3660	assigned reading	0.0413
maintain records	0.9741	generating business	0.3639	customer_service culture	0.0241
desired skills	0.9651	staffing needs	0.3632		
ensure delivery	0.9489	establish priorities	0.3496		

As described in the text, we exclude 70 tasks from the original list of 500 most common verb-noun pairs, using our judgment to select pairs that do not correspond to tasks. These excluded verb-noun pairs describe worker skills (e.g., "high school diploma," "ged years,"

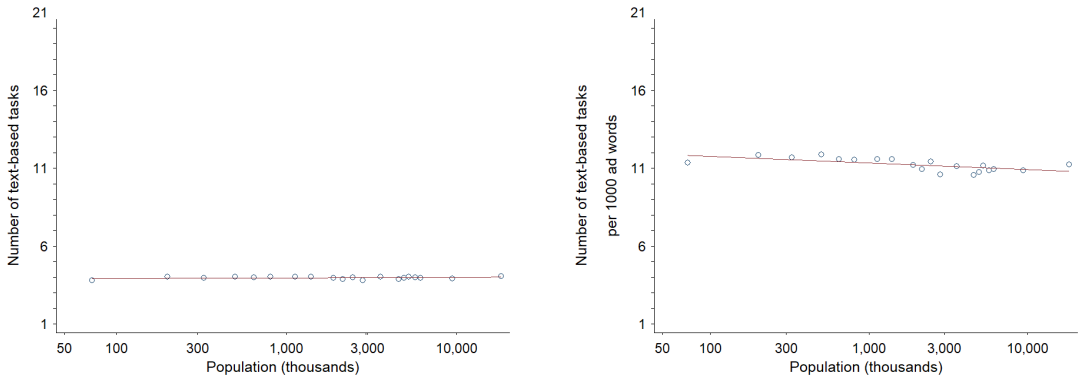
"required bachelor"), firm attributes (e.g., "is company," "is equal_opportunity"), aspects of the job search process ("pass drug"), or are simply uninformative ("meet needs," "be duties"). The excluded verb-noun pairs are:

Verb-Noun Pair Drop List

be years	is position	is delivery
is equal_opportunity	work part	are company
arc bach	are time	ged years
must years	ensure execution	include duties
high_school diploma	bach bets	required position
demonstrated ability	be team	be duties
required employee	travel travel	pass drug
bachelor degree	is experience	required bachelor
meet needs	may materials	are accordance
required ability	are drafts	sporting goods
required years	be doors	have ability
required skills	can doors	based business
according state	are business	ensuring aspects
include customers	requested react	assigned job
work hours	are store	be ability
are customers	including evenings	may duties
be customer	is law	are fast_growing
preferred years	is customer	requires state
required experience	earned degree	must_have driver
s degree	is ability	will business
arc setter	send resume	s level
end caps	s journal	is job
preferred experience	eas program	be company
including products		

Figure B.1 presents the frequency of text-extracted job tasks per ad. The left panel is a binscatter of number of tasks at the ad-level on CZ size, while the right panel presents the same figure but first normalizes the number of tasks per 1,000 ad words. There are about four tasks per ad on average (out of 430 total tasks), and when we normalize by ad length, as in the right panel, the number of tasks decreases with market size.

Figure B.1: Number of Tasks and Market Size



The left panel above presents a binned scatterplot of number of tasks against log CZ population. The right panel presents the same figure, except the dependent variable is normalized per 1,000 ad words.

B.3 Technology List

The table below lists the O*NET "Hot Technologies" that we identify in the job ads text along with the fraction of ads with each technology (× 100). To be counted as a technology appearance, all words in the technology name must appear in the vacancy text, although we do not require the words to appear in order.

Technologies Extracted from Job Vacancy Data (with Frequency per 100)

microsoft excel	2.0566	geographic information system gis software	0.0134
facebook	1.6065	microsoft dynamics gp	0.0133
sap	1.4853	transact-sql	0.0132
linux	1.4065	unified modeling language uml	0.0125
microsoft project	1.3218	apache cassandra	0.0119
microsoft word	1.1720	apache pig	0.0097
javascript	1.1669	extensible markup language xml	0.0077
unix	1.0452	cascading style sheets css	0.0077
microsoft office	1.0363	oracle business intelligence enterprise edition	0.0076
linkedin	1.0249	apache kafka	0.0071
microsoft access	0.8903	spring boot	0.0071
microsoft windows	0.8149	integrated development environment ide software	0.0068
react	0.7996	delphi technology	0.0065

microsoft outlook	0.7230	apache groovy	0.0060
python	0.7208	adobe systems adobe creative cloud	0.0057
microsoft powerpoint	0.6548	enterprise resource planning erp software	0.0054
microsoft sql server	0.5013	atlassian bamboo	0.0053
oracle java	0.4844	virtual private networking vpn software	0.0046
chef	0.4732	node.js	0.0045
sas	0.4551	ibm spss statistics	0.0045
ruby	0.4071	google angularjs	0.0037
youtube	0.4027	hypertext markup language html	0.0036
tax software	0.3962	job control language jcl	0.0030
ajax	0.3503	apache subversion svn	0.0019
mysql	0.3412	oracle hyperion	0.0015
git	0.2910	backbone.js	0.0014
swift	0.2735	customer information control system cics	0.0013
microsoft sharepoint	0.2653	oracle primavera enterprise project portfolio management	0.0013
citrix	0.1815	adobe systems adobe aftereffects	0.0009
microsoft visio	0.1793	practical extraction and reporting language perl	0.0007
nosql	0.1579	microsoft asp.net	0.0007
tableau	0.1526	ca erwin data modeler	0.0006
bash	0.1416	microsoft active server pages asp	0.0002
microsoft visual studio	0.1412	common business oriented language cobol	0.0001
microsoft dynamics	0.1411	salesforce software	0.0001
relational database management software	0.1397	google analytics	0.0001
microsoft exchange server	0.1342	computer aided design cad software	0.0001
google drive	0.1230	ibm websphere	0.0000
epic systems	0.1166	qlik tech qlikview	0.0000
objective c	0.1140	oracle peoplesoft	0.0000
microsoft sql server reporting services	0.1110	junit	0.0000
selenium	0.1097	oracle taleo	0.0000
puppet	0.1069	yardi	0.0000
spring framework	0.1022	national instruments labview	0.0000
apache tomcat	0.1010	microsoft .net framework	0.0000
data entry software	0.0952	microsoft asp.net core mvc	0.0000
microsoft visual basic	0.0860	autodesk autocad civil d	0.0000
symantec	0.0858	supervisory control and data acquisition scada software	0.0000
mongodb	0.0846	apache solr	0.0000

red hat enterprise linux	0.0769	microstrategy	0.0000
ruby on rails	0.0690	minitab	0.0000
postgresql	0.0617	enterprise javabeans	0.0000
microsoft azure	0.0549	c++	0.0000
shell script	0.0532	microsoft visual basic scripting edition vbscript	0.0000
scala	0.0508	oracle jd edwards enterpriseone	0.0000
teradata database	0.0492	medical procedure coding software	0.0000
drupal	0.0486	blackbaud the raiser's edge	0.0000
nagios	0.0476	php: hypertext preprocessor	0.0000
confluence	0.0466	apple macos	0.0000
verilog	0.0458	smugmug flickr	0.0000
adobe systems adobe acrobat	0.0457	oracle weblogic server	0.0000
mcafee	0.0448	github	0.0000
docker	0.0442	splunk enterprise	0.0000
oracle jdbc	0.0439	red hat wildfly	0.0000
adobe systems adobe photoshop	0.0438	oracle pl/sql	0.0000
intuit quickbooks	0.0433	aws redshift	0.0000
eclipse ide	0.0408	medical condition coding software	0.0000
fund accounting software	0.0348	bentley microstation	0.0000
apache hadoop	0.0337	dassault systemes solidworks	0.0000
adobe systems adobe illustrator	0.0325	advanced business application programming abap	0.0000
oracle fusion applications	0.0322	autodesk revit	0.0000
google docs	0.0314	dassault systemes catia	0.0000
ubuntu	0.0307	oracle solaris	0.0000
apache maven	0.0298	adobe systems adobe dreamweaver	0.0000
django	0.0282	adobe systems adobe indesign	0.0000
structured query language sql	0.0282	oracle javaserver pages jsp	0.0000
apache http server	0.0250	javascript object notation json	0.0000
hibernate orm	0.0245	the mathworks matlab	0.0000
meditech software	0.0237	netsuite erp	0.0000
apache ant	0.0231	ibm infosphere datastage	0.0000
ansible software	0.0229	trimble sketchup pro	0.0000
autodesk autocad	0.0219	healthcare common procedure coding system hcpcs	0.0000
ibm notes	0.0186	marketo marketing automation	0.0000
atlassian jira	0.0182	ibm cognos impromptu	0.0000
adp workforce now	0.0178	wireshark	0.0000

apache struts	0.0156	microsoft powershell	0.0000
sap crystal reports	0.0148	handheld computer device software	0.0000
esri arcgis software	0.0146	google adwords	0.0000
jquery	0.0140	elasticsearch	0.0000
apache hive	0.0135	c#	0.0000

B.4 Tasks and Market Size

Table B.4 presents the tasks with the steepest positive and negative gradients with market size, estimating equation (1). The results broadly support the finding that more interactive tasks and tasks emphasizing teamwork are more common in urban areas. Table B.5 presents the same results except we include six-digit SOC fixed effects as controls. Table B.6 re-runs equation (1) and instead of using our task list that is extracted from the text itself, we use a pre-determined list of verbs from Michaels, Rauch, and Redding (2018). The takeaway is quite similar. Using only the verb list, more abstract or non-routine verbs, such as "design," "project," "research," and "manage" have the steepest positive gradient, while more routine verbs, such as "store," "clean," and "count," and manual verbs, such as "fuel" and "rotate," have the steepest negative gradient.

Table B.4: Tasks with the Steepest Gradient: Extracting Tasks Directly from Ads

Positive gradient		Negative gradient	
Task	$\hat{\beta}_{10}$	Task	$\hat{\beta}_{10}$
written communication	0.1596	maintain store	-0.1763
managing projects	0.1157	maximizes profitability	-0.1692
committed diverse	0.1100	operating cash_register	-0.1653
meet deadlines	0.1075	protect company	-0.1641
providing support	0.0956	make changes	-0.1431
maintaining relationships	0.0943	provide customer_service	-0.1394
written skills	0.0922	preventing trafficking	-0.1373
work environment	0.0910	greeting customers	-0.1343
problem_solving skills	0.0881	skating carhop	-0.1334
required knowledge	0.0844	procedures cash	-0.1264
working relationships	0.0844	maintaining inventory	-0.1234
develop business	0.0833	assist store	-0.1221
developing strategies	0.0754	unloading trucks	-0.1191
identify opportunities	0.0751	ensure employees	-0.1143
prioritize tasks	0.0739	drive_in employees	-0.1104

We estimate equation (1) separately for each task, excluding controls. We normalize the estimates by dividing by the standard deviation of the task. The table above presents the tasks with the steepest positive and negative gradients with respect to market size, as captured by $\hat{\beta}_{10}$, which reflects the difference between tenth and first decile market size. All coefficients are statistically significant at the one percent level.

Table B.5: Tasks with the Steepest Gradient: Extracting Tasks Directly from Ads (with SOC f.e.)

Positive gradient		Negative gradient	
Task	$\hat{\beta}_{10}$	Task	$\hat{\beta}_{10}$
committed diverse	0.0915	maximizes profitability	-0.1597
achieving sales	0.0701	protect company	-0.1501
ensure safety	0.0686	maintain store	-0.1339
written skills	0.0580	operating cash_register	-0.1256
stand walk	0.0573	make changes	-0.1249
driving sales	0.0572	greeting customers	-0.1094
exceed sales	0.0556	procedures cash	-0.1080
work environment	0.0541	skating carhop	-0.1064
providing environment	0.0523	ensure employees	-0.1041
providing coaching	0.0510	unloading trucks	-0.1005
according company	0.0500	drive_in employees	-0.0981
prioritize tasks	0.0500	maintaining inventory	-0.0948
working relationships	0.0488	assigned store	-0.0873
handle tasks	0.0487	working store	-0.0852
using eye	0.0461	provide customer_service	-0.0848

The table above reproduces Table B.4 with six-digit SOC f.e. as controls. All estimates are statistically significant at the one percent level.

Table B.6: Verbs with the Steepest Gradient

Positive gradient		Negative gradient	
Task	$\hat{\beta}_{10}$	Task	$\hat{\beta}_{10}$
design	0.0812	pay	-0.0625
project	0.0797	truck	-0.0623
experience	0.0660	store	-0.0559
research	0.0632	earn	-0.0513
develop	0.0616	clean	-0.0506
manage	0.0581	license	-0.0452
web	0.0560	fuel	-0.0448
finance	0.0499	get	-0.0421
analyze	0.0492	rotate	-0.0396
process	0.0483	authorize	-0.0392
create	0.0461	count	-0.0362
content	0.0437	trash	-0.0321
lead	0.0432	average	-0.0320
market	0.0431	retail	-0.0307
track	0.0426	sign	-0.0301

The table above reproduces Table B.4, except uses the list of verbs from Michaels, Rauch, and Redding (2018) as tasks instead of the verb-noun pairs extracted from job descriptions. This exercise is conducted on a one percent sample of all job ads, rather than five percent, for computational speed, since the verb list includes 1,665 verbs. All estimates are statistically significant at the one percent level.

B.5 Technology Requirements and Market Size

Table B.7 re-estimates equation (1) where the dependent variable is a specific technology requirement. We estimate this regression separately for each O*NET technology and report the technologies with the steepest positive gradient with respect to market size. We estimate equation (1) using the entire sample of job ads, using the subsample of those requiring a high school degree only, and using the subsample requiring a college degree or above.

Table B.7 has several implications. First, the magnitude of the technology gradient is stronger for technologies requiring a college degree than a high school degree. Second, the technologies required of higher-skilled workers are more advanced and include computer programming (e.g., Python, Linux, Javascript, Unix) and not simply word processing (e.g., the Microsoft Office suite) or social media (e.g., Facebook, LinkedIn).[26]

[26]Table B.7 omits technologies with the steepest negative gradient because the estimates are small in magnitude and only two are statistically significant at the five percent level. First, pooling all ads, the coefficient estimate for Swift is -0.0593, and is significantly different from 0. It is likely that for many job ads "swift" is simply an adverb and not a reference to a technological requirement. For jobs requiring a high school degree, no technologies have a negative gradient that are statistically significant. For jobs requiring a

Lastly, we check the sensitivity of our result on the market size gradient of technologies with respect to our decision to exclude R and C from the technology list. Figure B.2 reproduces Figure 3 but includes the technologies R and C, which are potentially susceptible to false positives in processing the job vacancy text. Our main result is largely unaffected.

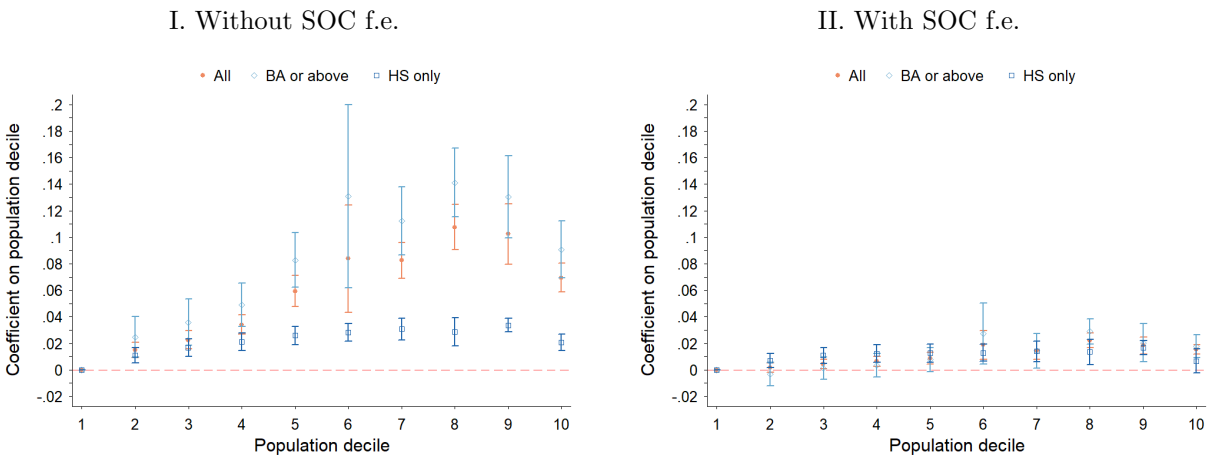
Table B.7: Technologies with the Steepest Gradient

All		**College**		**High School**	
Technology	$\hat{\beta}_{10}$	Technology	$\hat{\beta}_{10}$	Technology	$\hat{\beta}_{10}$
Microsoft Excel	0.1131	Python	0.0980	Microsoft Excel	0.0721
Python	0.0843	Microsoft Excel	0.0889	LinkedIn	0.0689
Microsoft Project	0.0789	Javascript	0.0844	Microsoft Outlook	0.0527
Linux	0.0785	SAS	0.0710	Facebook	0.0469
Microsoft Office	0.0720	Linux	0.0708	Microsoft Word	0.0453
SAP	0.0686	Microsoft Project	0.0706	Microsoft Office	0.0412
Microsoft Access	0.0685	Microsoft Access	0.0650	Youtube	0.0334
Microsoft Powerpoint	0.0680	Git	0.0644	React	0.0277
Microsoft Outlook	0.0630	MySQL	0.0591	Microsoft Access	0.0250
LinkedIn	0.0621	Facebook	0.0558	Microsoft Powerpoint	0.0239
Unix	0.0589	Tax Software	0.0553	Objective C	0.0216
Facebook	0.0589	Microsoft Office	0.0550	Python	0.0179
SAS	0.0584	Unix	0.0549	Ajax	0.0170
Geographic Information System (GIS)	0.0579	Ajax	0.0542	Yardi	0.0167
Git	0.0563	Microsoft SQL Server	0.0531	Citrix	0.0164

We estimate equation (1) where the dependent variable is a specific technology requirement, excluding controls. We estimate this regression separately for each O*NET technology. All coefficients are normalized by dividing by the standard deviation of the technology. We report the technologies with the steepest positive gradient with respect to market size, $\hat{\beta}_{10}$, which reflects the tenth decile technology intensity relative to the first decile. All estimates are statistically significant at the five percent level, with the following exceptions in the High School column: React ($p = 0.48$) and Ajax ($p = 0.09$).

college degree or above, only Apache Pig has a statistically significant negative gradient (-0.0134).

Figure B.2: The Technology Gradient (including R and C)



The figure above reproduces Figure 3 but includes the technologies R and C.

B.6 Wages and Tasks Across Space

This section demonstrates that tasks extracted from job vacancy ads account for variation in wages across geography, wage variation above and beyond what is captured by occupational codes.

For this analysis, we construct occupation-education-market average tasks from the job ads data. We then merge mean wages at the occupation-education-market level from the IPUMS-ACS. We then regress log wages on tasks, with different sets of controls. All regressions are weighted by employment in the cell.

Note that these regressions probably understate the explanatory power of job tasks in accounting for wage variation since we do not observe job-level wages and these are regressions of mean wages on mean tasks, using variation across geography-education cells. Second, while it is tempting to interpret these estimates as hedonic regressions that are delivering "task prices," we should avoid this interpretation because tasks are endogenous to unobserved worker sorting or job characteristics.

Table B.8 first shows that task variation across geography accounts for variation in wages above and beyond what is captured by occupation fixed effects. This result can be seen by the statistically significant coefficients on tasks in columns 3-6. Note that the slight increase in R^2 between columns 2 and 3 indicates that the five task categories only captures 0.1 percent of wage variation beyond occupation categories. But the granular task measures account for an additional 1.9 percent of wage variation, as seen by comparing R^2 between columns 3 and 4. Thus, the granular tasks extracted from job descriptions capture meaningful information

about job tasks that are reflected in wages. Note that for jobs requiring a college degree, non-routine analytic tasks have a stronger relationship with wages than for jobs requiring a high school degree only.

Table B.9 presents regressions of log wages on log population, tasks, and tasks interacted with population. In the coefficient on log-population, we confirm the finding in the literature that the relationship between population and wages is stronger for higher educated workers. We also see that the interaction terms between population and tasks appears important. For example, column 2 shows that an increase in interactive tasks in larger labor markets accounts for higher wages of jobs requiring a college degree, while an increase in interactive tasks for jobs requiring a high school degree has a weaker correlation with wages. Note that this table uses *within-occupation* variation in tasks across geography in accounting for higher wages. Overall, Tables B.8 and B.9 show that task variation across space accounts for variation in wages above and beyond occupation codes.

Table B.8: Wages and Tasks

	Baseline				HS only	BA or above
	(1)	(2)	(3)	(4)	(5)	(6)
Non-routine analytic	0.229***		0.050***	0.042***	0.020**	0.060***
	(0.013)		(0.010)	(0.012)	(0.008)	(0.016)
Non-routine interactive	0.085***		-0.003	-0.009	0.013	-0.005
	(0.012)		(0.006)	(0.006)	(0.010)	(0.009)
Routine cognitive	-0.008**		-0.021***	-0.002	-0.025***	-0.014
	(0.004)		(0.004)	(0.004)	(0.005)	(0.011)
Routine manual	0.059***		-0.018***	-0.007	-0.020***	-0.056***
	(0.005)		(0.006)	(0.005)	(0.006)	(0.011)
Non-routine manual	0.040***		0.010*	0.002	0.005	-0.057***
	(0.008)		(0.005)	(0.005)	(0.005)	(0.013)
SOC f.e.	No	Yes	Yes	Yes	Yes	Yes
Text-based tasks	No	No	No	Yes	No	No
Number of observations	58,494	58,494	58,494	58,494	33,859	24,635
R^2	0.489	0.784	0.785	0.804	0.552	0.694
Adjusted R^2		0.784	0.785	0.803	0.551	0.693
Mean of dep. var.	10.65	10.65	10.65	10.65	10.44	10.94

The unit of observation is the occupation-education-market. The dependent variable is log wages, regressed on Spitz-Oener (2006) task-related keywords per 1,000 ad words, which are standardized to have mean zero and standard deviation one across ads before averaging to the cell. Column 4 includes the verb-noun tasks averaged to the cell. The only controls are education category dummies, and four-digit SOC f.e., which are included in columns 2-5. Regressions are weighted by employment. Standard errors are clustered at the CZ level.

Table B.9: Wages and Task-Population Gradient

	HS only	BA or above
	(1)	(2)
Log pop.	0.043***	0.013***
× non-routine analytic	(0.006)	(0.004)
Log pop.	0.015**	0.029***
× non-routine interactive	(0.006)	(0.006)
Log pop.	0.002	0.009
× routine cognitive	(0.002)	(0.007)
Log pop.	-0.018***	-0.012***
× routine manual	(0.003)	(0.004)
Log pop.	0.002	-0.014
× non-routine manual	(0.003)	(0.009)
Log population	0.076***	0.081***
	(0.007)	(0.008)
SOC f.e.	Yes	Yes
Number of observations	33,859	24,635
R^2	0.594	0.766
Mean of dep. var.	10.44	10.94

The unit of observation is the occupation-education-market. The dependent variable is log wages, which is regressed on tasks, log population, and log population interacted with tasks. The only controls are education category dummies and four-digit SOC f.e. where indicated. Tasks are standardized to have mean zero, standard deviation one across ads before averaging to the cell. Regressions are weighted by employment. Task coefficients are not reported above. Standard errors are clustered at the market level. Tasks correspond to the classification in Spitz-Oener (2006).

C Analysis Appendix

This section presents tables and figures to supplement the main analysis.

C.1 Task Differences Across Geography: Within and Between Occupations

To evaluate whether the variation in occupational tasks across geography is due to within versus between occupation variation in task content, we perform a simple decomposition. Denote the average task k content in market size quartile q as, $t_{kq} = \sum_{o \in \mathcal{O}} t_{koq} s_{koq}$, which is expressed as the average task content of occupation o in quartile q, t_{koq}, multiplied by occupation o's share of quartile q's employment, s_{koq}. We express the difference in task

content between two quartiles, q and $\tilde{q}$ as

$$t_{kq} - t_{k\tilde{q}} = \sum_{o \in \mathcal{O}} (t_{koq} - t_{ko\tilde{q}}) \bar{s}_{ko} + \sum_{o \in \mathcal{O}} \bar{t}_{ko} (s_{koq} - s_{ko\tilde{q}}), \tag{4}$$

where $\bar{s}_{ko} = (s_{koq} + s_{ko\tilde{q}})/2$ and $\bar{t}_{ko} = (t_{koq} + t_{ko\tilde{q}})/2$. The first term on the right hand side of equation (4) represents the within component, and the second term represents the between component. Dividing both sides by $(t_{kq} - t_{k\tilde{q}})$ yields the within and between shares.

Table C.1 presents the results of this decomposition. For non-routine analytic tasks, 23 percent of the variation between first quartile and fourth quartile CZs is within occupation. For non-routine interactive tasks, the corresponding figure is 35 percent. This result implies that standard data sources fail to capture much of the variation in tasks between rural and urban markets.

Table C.1: Summary Statistics: Task Decomposition Across Markets

	NR-Analytic		NR-Interactive		NR-Manual		R-Cognitive		R-Manual	
Q1	4.39		4.73		0.84		0.64		2.98	
Q2	4.76		5.18		0.77		0.67		2.86	
Q3	5.13		5.63		0.79		0.70		2.70	
Q4	7.07		6.41		0.78		0.77		2.31	
	Between and Within Occupational Decomposition									
	Between	Within	Between	Within	Between	Within	Between	Within	Between	Within
Q2-Q1	0.61	0.39	0.43	0.57	0.54	0.46	0.86	0.14	0.64	0.36
Q3-Q2	0.73	0.27	0.65	0.35	0.39	0.61	1.68	-0.68	0.40	0.60
Q4-Q3	0.81	0.19	0.78	0.22	-0.16	1.16	1.22	-0.22	0.57	0.43
Q4-Q1	0.77	0.23	0.65	0.35	0.50	0.50	1.06	-0.06	0.56	0.44

The top panel plots the average task content in each of four market size quartiles. Tasks are expressed as number of task-word mentions per 1,000 ad words. The bottom panel presents a decomposition of the within and between shares of the total difference between population quartiles.

C.2 Appendix to Section 3.1

In this appendix, we present additional figures on the relationships among job tasks and population.

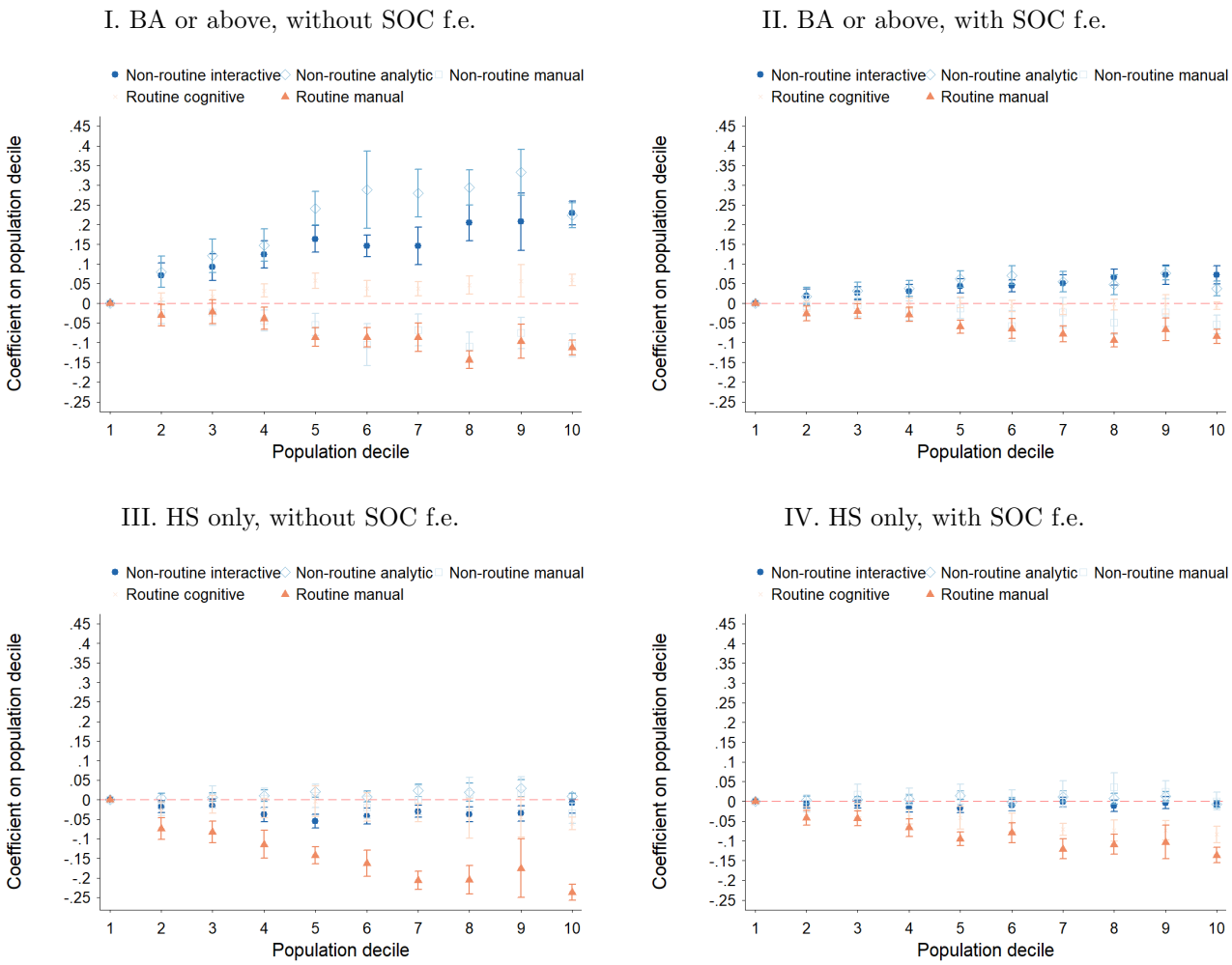
In Figure 1 of the paper, we demonstrate that non-routine interactive and non-routine analytic tasks are mentioned more frequently in larger markets, and routine manual and routine cognitive tasks less frequently. In Figure C.1, we demonstrate that these relationships are primarily due to differences among ads that require a college degree.

Figure C.2 considers whether there is evidence for jobs being jointly intensive in interactive and analytic tasks in large markets, as Deming (2017) found them to be increasingly

important over time. We place each job into one of four groups, based on whether it is above or below the median non-routine interactive task content, and above or below the median non-routine analytic task content. We then plot, for each decile, the difference between the proportion of jobs in each of the four groups relative to the proportion of jobs in the same group in the first CZ decile. This plot is presented as the left panel of Figure C.2. We find that jobs that are intensive in *both* analytic and interactive tasks make up 15 percentage points more of jobs in each of the highest three deciles compared to the lowest decile. Jobs that are intensive in only analytic tasks but not interactive tasks make up only about four percentage points more of jobs in the highest three deciles, while jobs that are only interactive but not analytical make up a smaller share of total jobs in the highest decile markets, relative to smallest decile markets. This finding holds even after removing the mean task content at the six-digit SOC level before categorizing into the four groups, as seen in the right panel of Figure C.2.

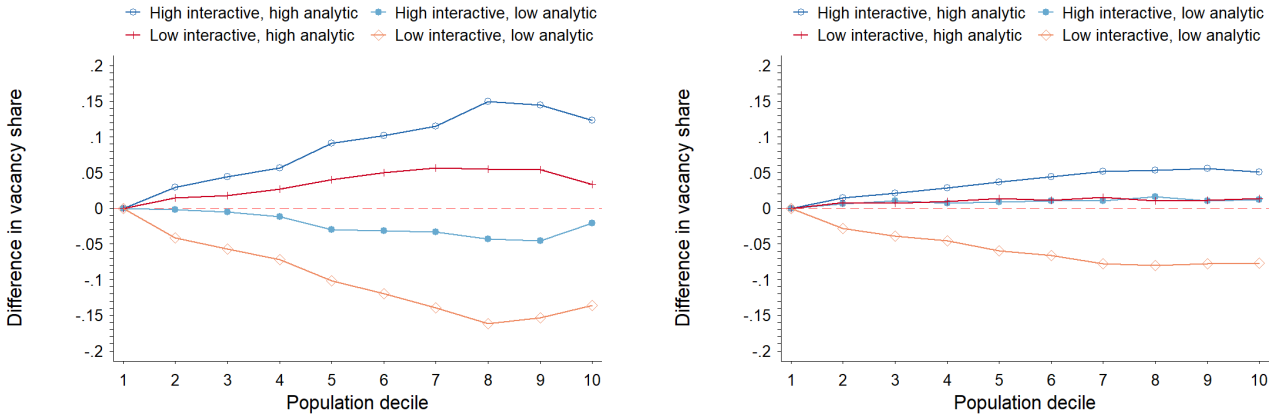
In Figure C.3, we explore whether the gradients presented in Figure 2 differ according to the jobs' educational requirements. For the most part, gradients are steeper for jobs requiring a college degree. However, in specifications with six-digit SOC occupation fixed effects, the difference between these gradients is minor.

Figure C.1: Task Gradient by Educational Requirements



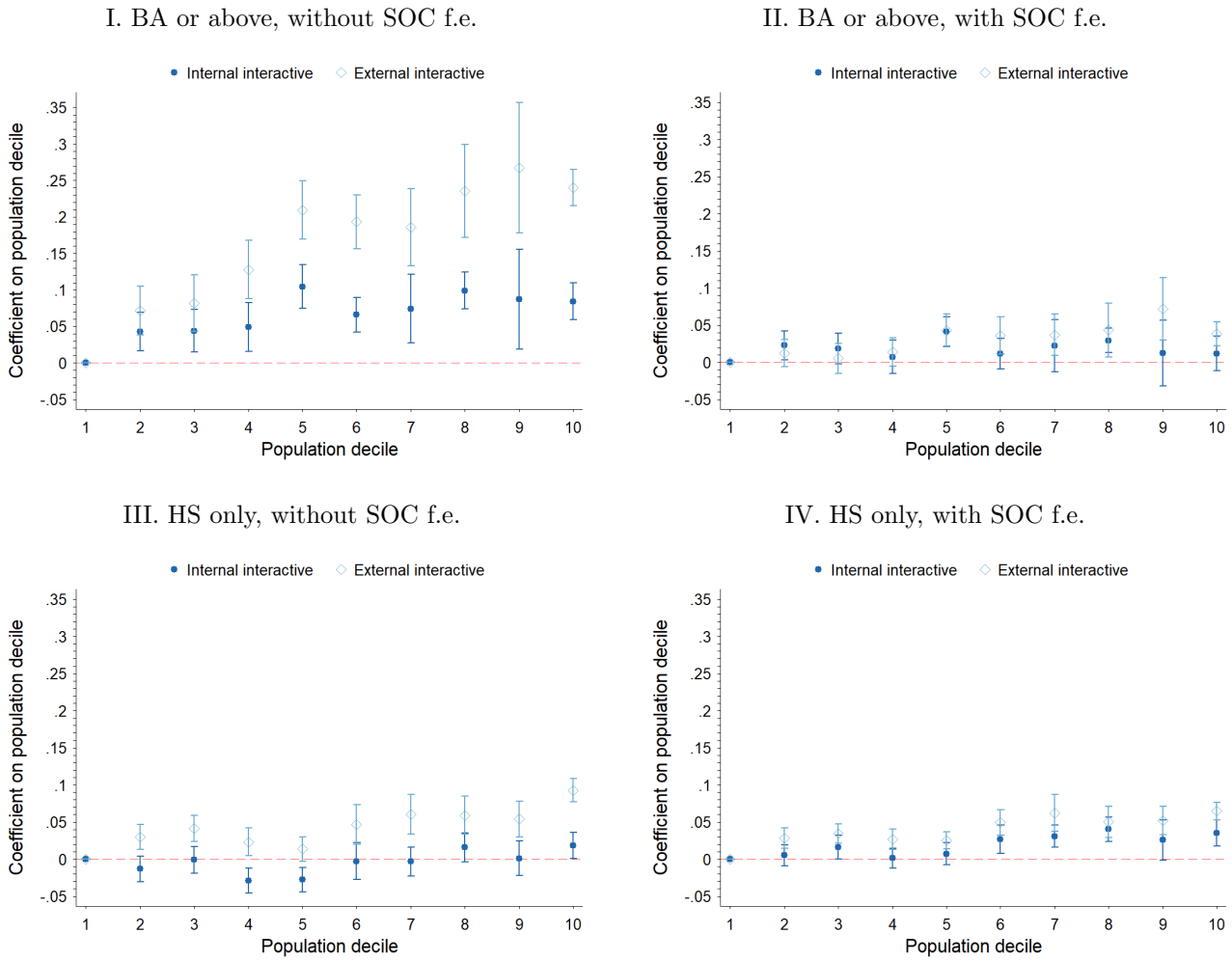
This figure reproduces Figure 1, separately by the educational requirement of the job. Panels I and II restrict the sample to ads requiring a BA or above, while panels III and IV restrict the sample to ads requiring high school only.

Figure C.2: Interactive and Analytic Tasks and Market Size



The panels above depict the distribution of jobs across space. To construct the left panel, we first place job ads into one of four mutually exclusive groups, based on whether they are above or below the median non-routine interactive task content and non-routine analytic task content. We then plot the difference between the proportion of jobs in each of the four categories (high or low, analytic or interactive) relative to the proportion of jobs in the same category in the first CZ decile. The right panel is constructed in the same way except we first subtract the SOC mean task content from each job before placing jobs into groups, and hence the right panel reflects within-occupation changes in task content across space.

Figure C.3: O*NET Interactive Tasks Gradient



This figure reproduces Figure 2, separately by the educational requirement of the job. Panels I and II restrict the sample to ads requiring a BA or above, while panels III and IV restrict the sample to ads requiring high school only.

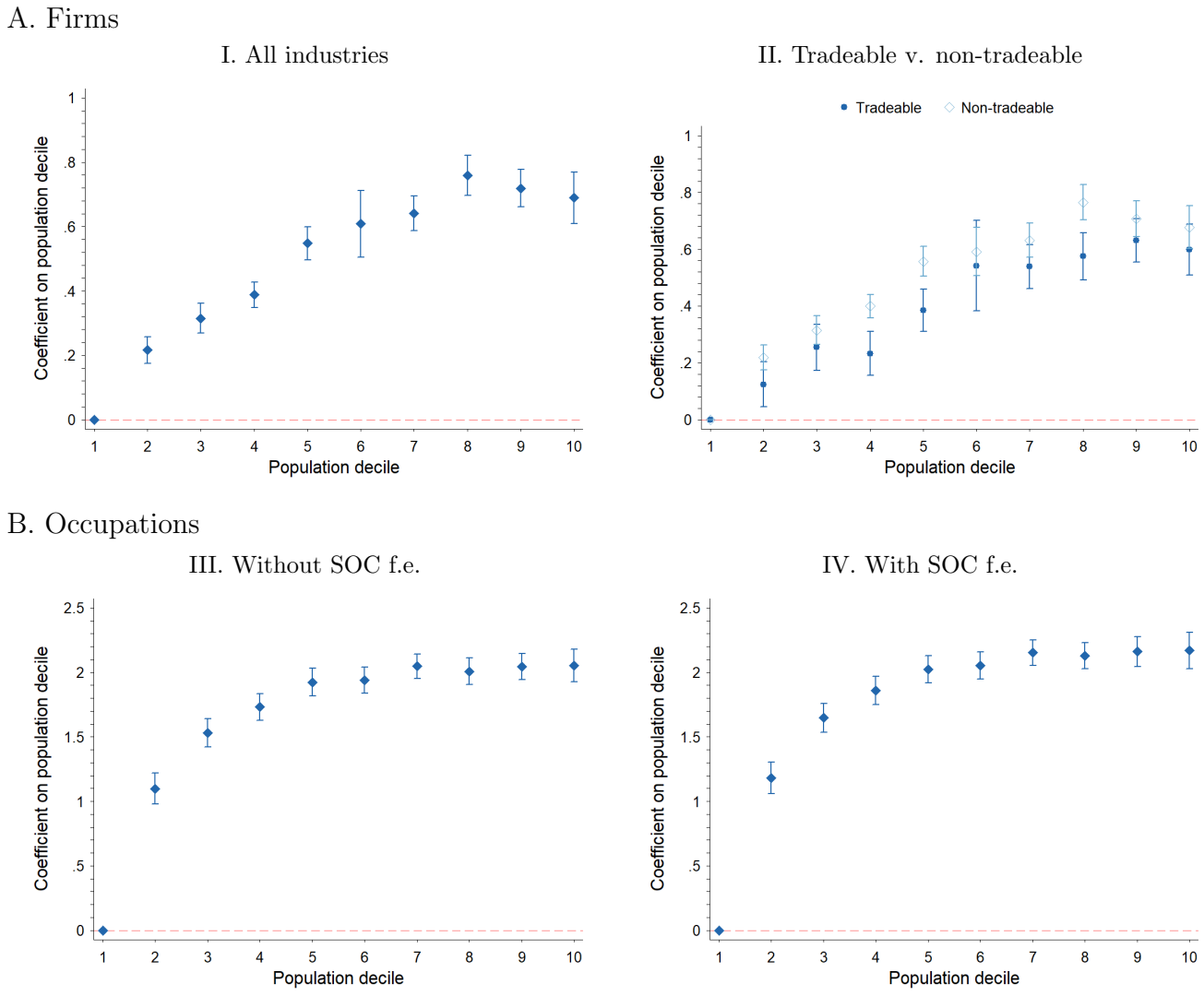
C.3 Specialization and Market Size

This section provides supplemental evidence on the relationship between specialization within and between firms and market size.

Robustness to the Number of Tasks

Our measurement approach requires setting a threshold for the number of tasks (verb-noun pairs) we use to study specialization. In the paper, we use a task list of 500 verb-noun pairs, which we winnow to 430 by excluding those that, according to our judgment, do not reflect job tasks. In this section, we reduce the number of tasks to 300 – i.e., keeping the most common 300 of the 430 remaining tasks – and reproduce Figure 4, the main figure that uses these granular task measures. Figure C.4 shows that the results are not sensitive to the choice of number of tasks.

Figure C.4: Specialization Gradient: Task Dissimilarity Within Firms and Occupations (300 Tasks)

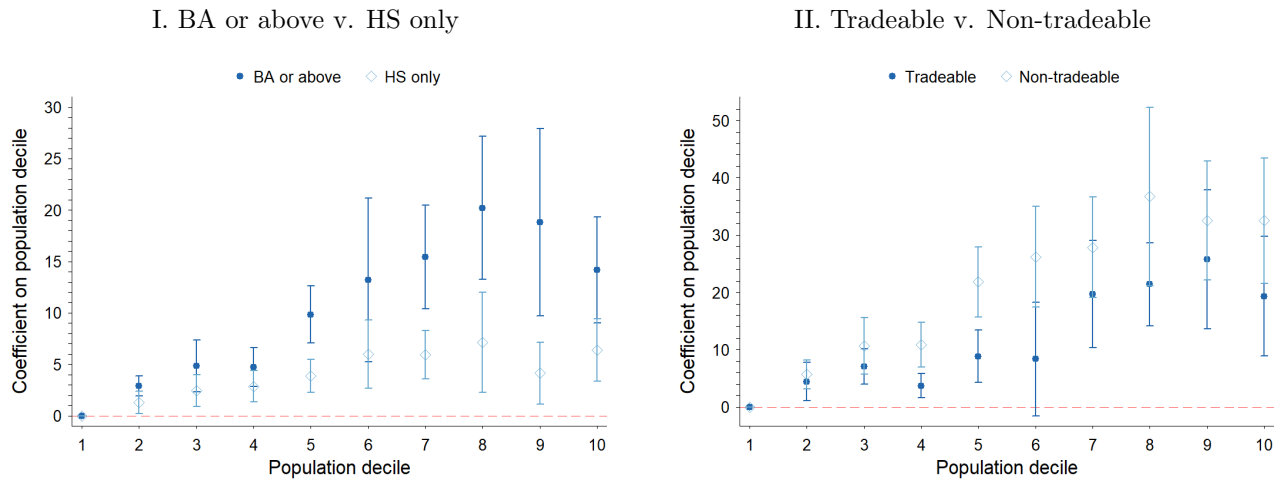


This figure reproduces Figure 4 using a task list of 300 verb-noun pairs.

Number of Job Titles

Prior research – notably, Tian (2019) – explores evidence for specialization by counting the number of distinct occupation codes in a firm-market. The idea behind this exercise is that a greater number of distinct occupations implies greater specialization in production. We examine this relationship in Figure C.5, using our job vacancy data to count distinct job titles within a firm name × six-digit industry NAICS × CZ. We produce these market size gradients separately for high and low education level job titles, and for tradeable and non-tradeable sector firms. The key takeaway is that we do see a positive relationship between market size and the degree of worker specialization, and this relationship is stronger for workers with a BA degree or above and for non-tradeable sector firms.

Figure C.5: Specialization Gradient: Number of Job Titles



The unit of observation is the firm-market (CZ). We regress the number of distinct job titles on market size deciles, controlling for total number of ads placed by the firm in the CZ, two-digit NAICS code, and the average log ad length. The left panel depicts two regressions. In the first, the dependent variable is the number of job titles requiring a high school degree, and in the second, the dependent variable is the number of distinct job titles requiring a college degree. In the right panel, the dependent variable is the number of distinct job titles, and the regression is estimated separately on tradeable and non-tradeable sector firms. All regressions are weighted by number of ads in the firm-market. Standard errors robust and clustered at the CZ level. The figure plots the coefficients on the CZ size deciles. For reference, in the left panel, the first decile CZ mean for BA or above is 2.58 and for HS only is 3.11. In the right panel, the first decile CZ mean for tradeable is 9.96 and for non-tradeable is 10.68.

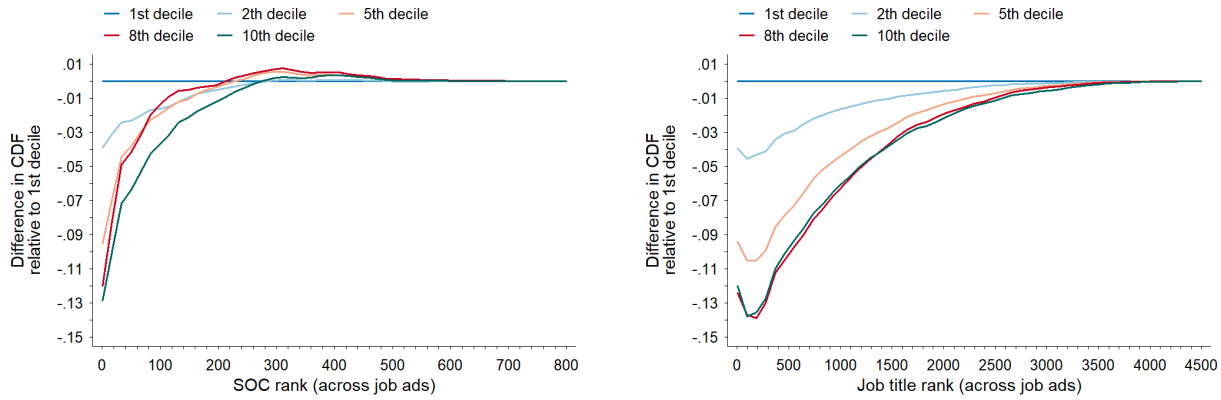
The Distribution of Common and Rare Occupations

As another robustness exercise, we measure the degree of specialization by examining the distribution of common and rare occupations across space.

We rank six-digit SOCs based on their share of all ads in the full sample. The x-axis presents SOCs in descending order based on their overall rank in the sample. We then compute the share of each SOC in each market size decile and plot the difference relative to the share in the first decile CZ. The left panel of Figure C.6 shows that the most common occupations are over-represented in small markets, while more rare occupations are over-represented in large markets. For example, among the ten most common occupations economy-wide, the tenth decile market has 11-13 percentage points lower share of these occupations, compared to the first CZ decile. For the 300-400 most common occupations, the tenth decile market has about a 0.3 percentage point greater share relative to the first decile.

This finding – that rare jobs represent a larger share of total jobs in larger markets – is even more pronounced when we do the analysis at the job title level. Note that the job title is not observed in standard datasets such as the ACS or the Current Population Survey (CPS), and hence represents an additional virtue of the job ads data used here. The right panel presents the analysis at the job title level, showing even more dramatically that common jobs are over-represented in smaller markets (as a share of total jobs).

Figure C.6: Common and Rare Occupations and Job Titles



The left panel is constructed as follows. We first generate the empirical cdf of occupational shares for each CZ decile. On the x-axis the six-digit SOCs are ranked in order of their shares of all job ads in the sample, from highest to lowest. The left panel presents the difference between each CZ decile cdf and the first decile CZ's cdf. The right panel is constructed analogously except the unit of analysis is the job title rather than the six-digit SOC. A local polynomial smoother is applied to both panels.

Specialization Across Firms

In Section 3.3, we considered the relationship between market size and within-firm specialization. In this appendix, we consider specialization across firms. The unit of analysis in this section is the industry-market, and the degree of specialization is computed across firms in the industry-market.

Define the dissimilarity between firm f in industry i and market m and other firms in the industry-market as $d_{fim} = 1 - \overline{V}_{fim} \cdot \overline{V}_{(-f)im}$. In this equation, $\overline{V}_{fim}$ is the vector of average tasks for the firm-industry-market, and $\overline{V}_{(-f)im}$ is the vector of average tasks for all firms other than f in the industry-market. The dependent variable in this analysis of across-firm specialization is $d_{im} = \frac{1}{n_{im}} \sum_{f,m} d_{fim}$, where n_{im} is the number of firms in the industry-market cell.

Table C.2 presents regressions at the industry-market level. There are three takeaways: in larger markets, (1) there are more firms, (2) firms have more workers, (3) firms are farther apart in task space. These effects are larger for the non-tradeable sector.

Taking these results together with those of Section 3.3, both within- and between-firm specialization increase with market size.

Table C.2: Specialization Across Firms

	Mean	OLS			
		(1)	(2)	(3)	(4)
Firm size	3.544	0.892***	1.093***	0.693***	1.158***
	(4.409)	(0.047)	(0.057)	(0.053)	(0.061)
		103,501	103,501	20,560	82,273
Number of firms	18.237	8.968***	9.374***	1.500***	8.090***
	(32.266)	(0.728)	(0.675)	(0.131)	(0.591)
		103,501	103,501	20,560	82,273
Task dissimilarity between firms	0.895	0.023***	0.027***	0.018***	0.029***
	(0.158)	(0.001)	(0.002)	(0.002)	(0.002)
		46,331	46,331	5,528	40,144
NAICS f.e.			Yes	Yes	Yes
Tradeable sector only				Yes	
Non-tradeable sector only					Yes

The unit of analysis is the industry-market. Each cell in columns 1-4 is a separate regression of the dependent variable indicated in the row on log CZ population. The estimate for log CZ population is presented along with the standard error (in parentheses) and the number of observations. Task dissimilarity between firms is the dissimilarity between the firm-level average and the leave-out firm average in the industry-market. Firm size is measured as the total number of job ads in the market, and the number of firms is measured as the number of firms with a posting in the industry-market cell. The only controls included are average log total ad words in the cell, and, where indicated, six-digit NAICS code. Regressions are weighted by the number of firms. Standard errors are clustered at the CZ level.

C.4 Tasks, Technologies, and the Urban Wage Premium

This section supplements Section 3.4 to show that job tasks and technology requirements have implications for the urban wage premium.

We construct task dissimilarity, technology requirements, and education in each CZ size decile n relative to the first decile. Hence, we measure the mean task content of each task in each four-digit SOC $\times$ CZ decile bin, and then compute the task dissimilarity between each occupation-CZ decile and the same occupation in the first CZ decile: $\Delta d_{on} = 1 - \overline{V}_{on} \cdot \overline{V}_{o1}$. We also measure the technology gap, $\Delta tech_{on} = tech_{on} - tech_{o1}$, the average number of technological requirements per job in the occupation-CZ decile, relative to the same occupation in the first CZ decile. And we measure the education gap, $\triangle ba_{on} = ba_{on} - ba_{o1}$, the difference in the fraction of workers in the occupation-CZ decile with a BA degree or higher relative to the same occupation in the first CZ decile, using the ACS.

We run the following regression:

$$\Delta log(wage)_{on} = \gamma_0 + \gamma_1 \Delta d_{on} + \gamma_2 \Delta tech_{on} + \gamma_3 \Delta ba_{on} + \epsilon_{on}, \quad (5)$$

and exclude the first decile from the regression sample. The coefficient γ_1 represents the effect of task dissimilarity on the urban wage premium; specifically, γ_1 captures the extent to which *within*-occupation differences in tasks across geography account for variation in the urban wage premium, beyond what is captured by differences in worker skills. Similarly, γ_2 is informative about the extent the technological requirements of jobs are important for accounting for the urban wage premium.

Table C.3 shows that occupational tasks play an important role in the urban wage premium. The mean dissimilarity between tenth and first decile CZs is 0.075 and hence, column 1 shows that this rural-to-urban increase in task dissimilarity is associated with an increase in wages of about 3.8 percent (0.075×0.501). This represents about 13 percent of the mean wage gap between the tenth and first deciles. Controlling for educational differences (i.e., including Δba_{on} as a control) does not weaken the relationship between tasks and wages, suggesting that the sorting of educated workers does not explain away the importance of job tasks.

Columns 3-5 re-estimate equation (5) separately by occupational category. We classify workers into white collar, blue collar, and service workers by two-digit SOC code, as described in the table note. Descriptively, the within-occupation urban wage premium is largest for white collar occupations, as can be seen in the mean of the dependent variable (0.34), and smallest for blue collar occupations (0.008), and the education gap is largest for white collar occupations (0.16) and smallest for blue collar occupations (0.04). A key finding from columns 3-5 is that within-occupation task dissimilarity plays an important role in accounting for variation in the urban wage premium for white collar occupations and much less for service and blue collar occupations, echoing our result in Section 3.4.

Columns 1 and 2 show that the technology gap between urban and rural areas is predictive of the urban-rural wage gap. Much of the technology premium is eliminated once we condition on the education gap in column 2, suggesting that the education requirements and technology requirements of jobs increase jointly in cities. Columns 3-5 show that service workers have the strongest urban wage premium for technology requirements.

Table C.3 has several takeaways. First, occupations differ between rural and urban areas in the U.S. in both their task content and technological requirements, which are unobserved using standard data sources. Second, within-occupation task differences are strongly correlated with the within-occupation urban wage premium, even after controlling for educational differences of workers. Third, the correlation between task dissimilarity and the urban wage premium is strong for white collar occupations but weak for blue collar and service occupa-

tions. Similarly, the technology premium is highest for white collar and service occupations.

These empirical findings suggest that the returns to urban migration differ for workers in blue versus white collar occupations. The findings also suggest that the task content within occupations varies from rural to urban areas, which may affect the ease of migrating occupation-specific human capital from rural to urban labor markets.

Table C.3: Explaining the Urban Wage Premium

	All		White collar	Blue collar	Service
	(1)	(2)	(3)	(4)	(5)
Task dissimilarity	0.501**	0.745***	0.802***	-0.117*	0.020
	(0.191)	(0.107)	(0.109)	(0.058)	(0.052)
Technology requirements gap	0.128***	0.020	0.020*	-0.130	0.150***
	(0.016)	(0.011)	(0.010)	(0.073)	(0.037)
Education gap		1.760***	1.510***	0.920***	1.366***
		(0.029)	(0.029)	(0.085)	(0.144)
Number of observations	1,080	1,080	510	320	220
R^2	0.213	0.696	0.609	0.196	0.395
Mean of dependent var.	0.291	0.291	0.335	0.008	0.157
Mean task dissimilarity	0.075	0.075	0.074	0.067	0.124
Mean technology gap	0.424	0.424	0.495	0.026	0.074
Mean education gap	0.141	0.141	0.159	0.039	0.041

In the regressions reported above, the unit of analysis is occupation-CZ decile. The dependent variable is the log wage gap between the occupation-CZ decile and the same occupation in the first CZ decile. The right hand side includes mean task dissimilarity between the occupation-CZ decile and the same occupation in the first CZ decile, and, in columns 2-5, the difference in the fraction of workers in the occupation-CZ decile with a BA degree or higher, relative to the first CZ decile. (The sample excludes the first CZ decile.) Standard errors are robust and clustered at the CZ decile and observations are weighted by employment.

C.5 Robustness to Data Source

In this appendix, we reproduce some of our main empirical exercises using a sample of ads from Burning Glass. Our EMSI dataset has its own advantages. In particular, it contains the ads' raw text, allowing us to isolate the tasks that employers mention. In contrast, Burning Glass co-mingles jobs' skills, technologies, and tasks. Nevertheless, since Burning Glass has been so commonly used in recent analyses of the labor market, we check the robustness of our results to this alternate data source.

We draw a random sample of 1.2 million ads from January 2012 to December 2017. For this sample, so that we may replicate Figure 2, we compute measures of internal-to-

the-firm interactive tasks[27] and external-to-the-firm interactive tasks.[28] As in Section 3.1, we compute the number of task mentions per 1000 ad words. Second, as in Section 3.2, for each ad we compute whether the ad mentions individual O*NET "Hot Technologies." So that we may compute specialization, as in Section 3.3, for each job ad j we define a 400-dimensional vector, T_j, with each element characterizing whether ad j mentions the individual Burning Glass element. As in Section 3.3, define the normalized task vectors $V_j = \frac{T_j}{\sqrt{T_j \cdot T_j}}$, and the distance between job j and other jobs in the occupation- (or firm-) and market as $d_{jcm} = 1 - V_{jcm} \cdot \overline{V}_{(-j)cm}$.

First, Figure C.7 replicates Figure 2. As in Section 3.1, internal and external tasks each increase in city size, both within and between six-digit SOC occupations.

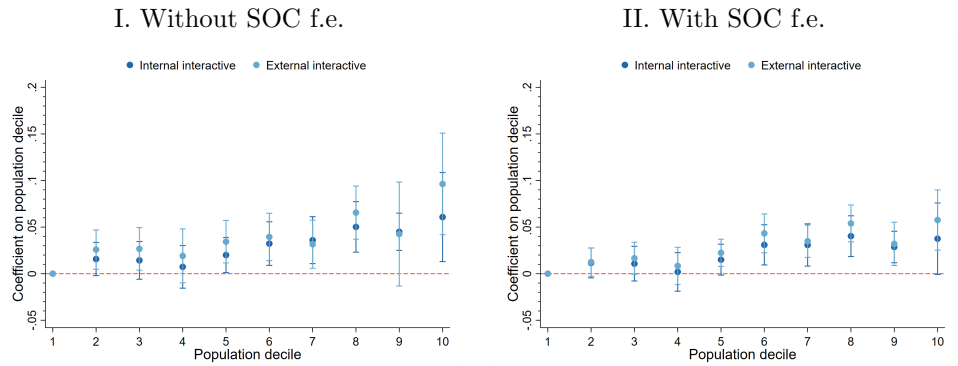
Second, we reproduce Figure 4. As in Figure 4, Figure C.8 indicates that within-occupation and within-firm specialization is greater in more populous commuting zones, with a steeper gradient for firms in non-tradeable industries than for firms in tradeable industries (panel II).

Finally, we reproduce Table 1. As in Table 1, Table C.4 indicates that wages are higher in markets with greater specialization, with greater technology usage, and with a greater share of workers with a college degree. The relationships between wages and within-occupation $\times$ market specialization is stronger in white collar than in blue collar occupations. In contrast to Table 1, however, the association between wages and technology mentions is stronger in blue collar occupations.

[27] We map the following Burning Glass elements to internal interactive tasks: "Agile coaching," "Communication Skills," "Employee Coaching," "Executive Coaching," "Leadership," "Leadership Development," "Leadership Training," "Mentoring," "Oral Communication," "Peer Review," "Personal Coaching," "Supervisory Skills," "Team Building," "Verbal / Oral Communication," and "Written Communication."

[28] We map the following Burning Glass elements to external interactive tasks: "Advertising," "Client Base Retention," "Client Care," "Client Needs Assessment," "Client Relationship Building and Management," "Communication Skills," "Digital Marketing," "Market Planning," "Marketing," "Marketing Communications," "Marketing Programs," "Marketing Sales," "Marketing Strategy Development," "Merchandising," "Oral Communication," "Print Advertising," "Product Marketing," "Professional Services Marketing," "Prospective Clients," "Public Relations," "Public Relations Campaigns," "Public Relations Industry Knowledge," "Public Relations Strategy," "Sales," "Telemarketing," "Vendor Interaction," "Vendor Performance Monitoring," "Vendor Relations," "Verbal / Oral Communication," and "Written Communication."

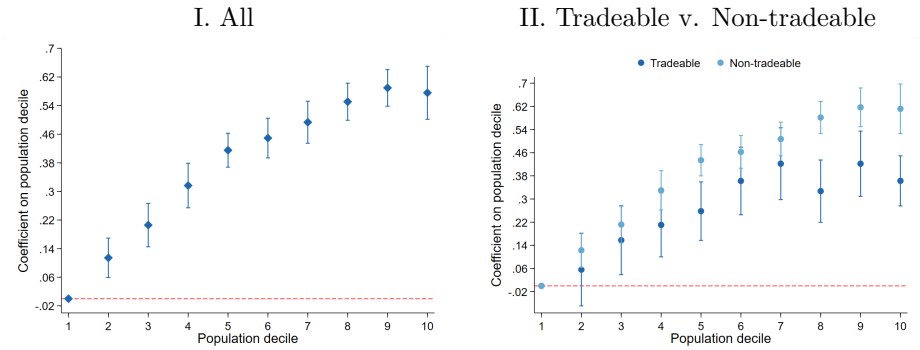
Figure C.7: O*NET Interactive Tasks Gradient



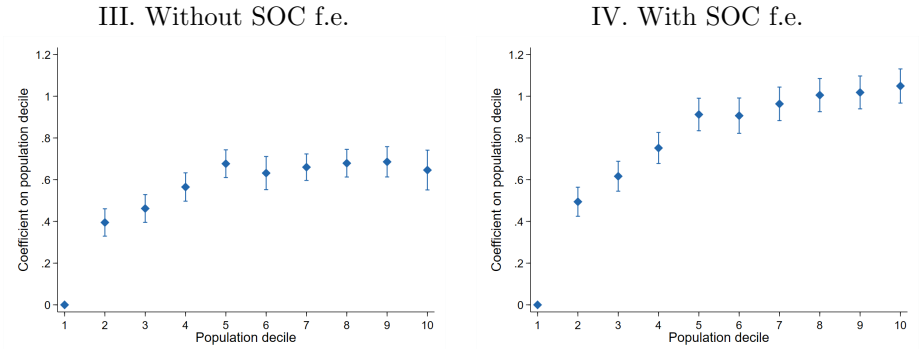
See the caption for Figure 2. In contrast, our task measures here come from Burning Glass.

Figure C.8: Specialization Gradient: Task Dissimilarity Within Firms and Occupations

A. Firms



B. Occupations



See the caption for Figure 4. In contrast, the task dissimilarity and technology measures here come from Burning Glass.

Table C.4: Task Dissimilarity, Technologies, and Wages

	All			White collar	Blue collar	Service
	(1)	(2)	(3)	(4)	(5)	(6)
Task dissimilarity	0.062***	0.046***	0.042***	0.091***	-0.019*	0.011**
	(0.004)	(0.003)	(0.003)	(0.016)	(0.010)	(0.005)
Technology requirements	0.306***	0.027***	0.015***	-0.021	0.114***	0.338***
	(0.008)	(0.006)	(0.005)	(0.020)	(0.023)	(0.072)
Education		0.978***	0.964***	0.947***	0.562***	0.657***
		(0.023)	(0.023)	(0.047)	(0.061)	(0.077)
SOC f.e.						
Number of observations	29,211	29,211	29,211	18,519	6,039	4,511
R^2	0.184	0.573	0.587	0.539	0.170	0.199
Mean of dependent var.	10.788	10.788	10.788	10.974	10.568	10.238
Mean task dissimilarity	0.000	0.000	0.000	0.043	-0.004	-0.198
Mean technology requirements	0.580	0.580	0.580	0.772	0.251	0.170
Mean BA or above	0.389	0.389	0.389	0.555	0.072	0.087

See the caption for Table 1. In contrast, the task dissimilarity and technology measures here come from Burning Glass.