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 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.5, 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. Furthermore, drawing on a survey conducted as part of the Princeton Data Improvement Initiative, we validate our ad-based task measures using within-occupation variation. 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

There are 722 CZs in our analysis sample. 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. CZs are assigned to market size deciles using employment weights so that each decile n has approximately the same number of employed workers. Note that Table A.1 shows that the number of job ads in each decile differs somewhat due to the discreteness of assigning each CZ to one decile.

Table A.1: CZ Decile Summary Statistics

Decile	Total ads	Pct. urban	Density	Median CZ pop.	Median CZ name(s)
1	506.8	42.2	16.9	54.9	Norfolk & Madison Counties, NE
2	575.0	67.3	72.9	304.7	Jackson & Hillsdale & Lenawee Counties, MI; Bloomington, IN
3	595.2	76.6	132.7	609.6	Wichita, KS
4	599.8	83.9	202.3	1,033.4	Tulsa, OK; Naples-Marco Island, FL
5	732.3	88.7	426.8	1,639.0	Nashville-Davidson-Murfreesboro, TN
6	692.3	92.1	440.8	2,441.2	St. Louis, MO
7	705.1	94.9	461.1	3,453.2	Minneapolis-St. Paul, MN; Hartford-Bridgeport-Stamford-Norwalk, CT
8	858.4	96.0	666.5	5,056.6	Atlanta, GA
9	685.3	96.6	1,103.4	6,159.5	Newark-Trenton-White Plains NJ-NY; Houston, TX
10	385.4	98.5	920.7	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 mean fraction of the population that is urban, the mean population density (persons per square kilometer), 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 in which the median CZ population is the average of two CZs, we provide both names. Area and percent urban are provided by the U.S. Census's 2010 Percent Urban and Rural by County report, which we link to CZ and then report mean CZ statistics in the decile.

A.2 Details on Sample Construction

We use a 5 percent sample of the online job ads data we purchased from EMSI. The sample of our dataset covers January 2012 to March 2017. We exclude ads with fewer than the 1st 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 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 impose a few additional restrictions. We drop ads placed by staffing or placement agencies, since they act as intermediaries between the worker and the firm hiring the worker. These ads are identified with a flag in the EMSI data. This step drops 596,578 ads. (We discuss ads placed by staffing agencies in the next paragraph.) We drop ads without a firm name, which is another 107,317 ads. Finally, we drop firms with no NAICS code—another 3,771 ads. These restrictions yield approximately 5.6 million ads for the sample used for the firm-level analysis.

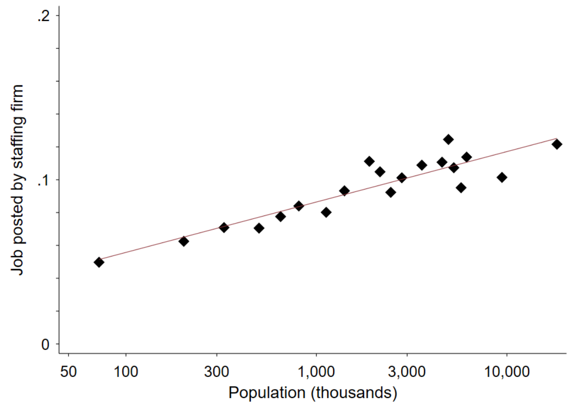
Jobs are more likely to be posted by a staffing agency in larger markets. This gradient can be seen in Figure A.1, which presents a binscatter of an indicator for the job ad's being posted by a staffing firm against the CZ population. To better understand these types of ads, we estimate a job ad-level regression of an indicator for a BA-requirement (or a HS-only requirement) on an indicator for the job being posted by a staffing agency, and estimate this regression for the sample of jobs with a non-missing education requirement. These results are reported in Table A.3. The main takeaway is a job ad posted by a staffing firm, on average, has an 19 percentage point higher likelihood of requiring a BA, which is mainly driven by occupational composition. Controlling for six-digit occupation fixed effects, the estimate is 4.4 percentage points. In our analysis, we find that higher-skilled jobs are more specialized; hence, the greater composition of staffing firm-posted vacancies in larger CZs is likely to lead us to understate the specialization gradient. We check the sensitivity of our specialization results by reproducing Figure 4, panel A and including the staffing firms; these results are in Figure A.2. The main takeaway that within-firm specialization is increasing in market size is unchanged.

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 applying the sample restrictions described in Appendix A.2.

Figure A.1: Job Posted by a Staffing Firm



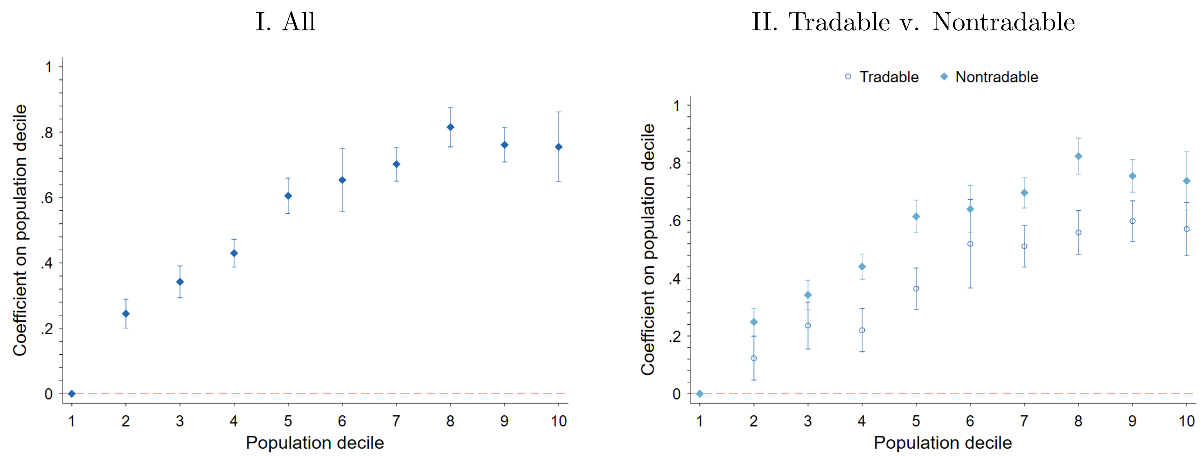
This figure presents a binned scatterplot of an indicator for the job ad's being posted by a staffing firm on log population at the CZ level.

Table A.3: Education Requirements and Staffing Firm-Posted Ads

	BA or above		HS only	
	(1)	(2)	(3)	(4)
Staffing firm	0.199***	0.044***	-0.175***	-0.032***
	(0.007)	(0.002)	(0.006)	(0.002)
SOC f.e.	No	Yes	No	Yes
Number of observations	3,209,292	3,209,292	3,209,292	3,209,292
R^2	0.012	0.555	0.010	0.542
Mean of dep. var.	0.51	0.51	0.35	0.35

The unit of analysis is the job ad and the regression sample includes all job ads with a posted education requirement. The dependent variable is an indicator for the job requiring a college degree (column 1-2), and an indicator for the job requiring a high school degree only (column 3-4). The right-hand side includes an indicator for the job being posted by a staffing firm and, in columns 2 and 4, six-digit SOC f.e. Standard errors are clustered at the CZ level.

Figure A.2: Specialization Gradient (Including Staffing Firms)



This figure reproduces panel A of Figure 4 but includes staffing firms in the sample.

A.3 Evaluating Online Vacancies: Representativeness and Selection

This section examines the coverage of job ads across SOC-CZ cells. It then compares the proportion of online job ads across industries to the proportion as measured in the Job Openings and Labor Turnover Survey (JOLTS).

We first characterize the coverage of job ads across SOC-CZ cells. We use the BLS 2010 version of the SOC classification, which are available in the IPUMS ACS. The IPUMS ACS has 111 four-digit SOCs that are non-aggregated and there are 722 CZs.[35] Of the 80,142 potential cells, about 9 percent are missing from the ACS, because rural commuting zones do not have employment in every occupation. Of the populated ACS four-digit SOC-CZ cells, the job ads data cover 71.0 percent of cells, which represent 98.3 percent of employment. The most common missing cells in the job ads data are those in the Military and those in Farm, Fishing, and Forestry occupations. Turning to two-digit SOCs, there are 722 CZs and 23 two-digit SOCs. Of the 16,606 possible cells, 0.86 percent are missing in the ACS. Our job ads data cover 90.7 percent of the remaining cells, which represent 99.8 percent of employment.

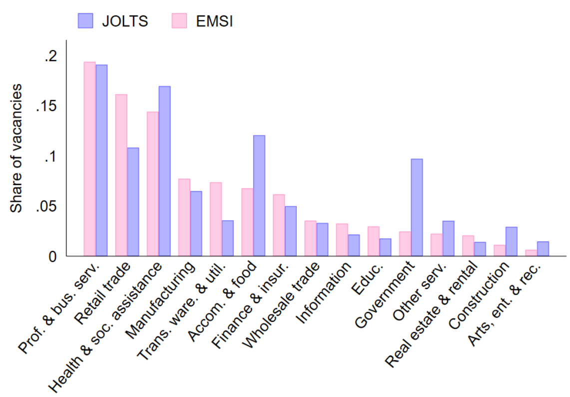
The standard resource for measuring job vacancies in the U.S. is JOLTS, conducted by the Bureau of Labor Statistics of the U.S. Department of Labor. The dataset consists of

[35]Some IPUMS occupation codes are aggregated because they lack an exact match to a Census occupation code or to preserve confidentiality because there are fewer than 10,000 individuals in the cell nationwide.

monthly job openings at the national level by major industry category.[36] JOLTS is based on a survey of a random subset of establishments covered by state or federal unemployment insurance laws.[37]

Figure A.3 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 than in JOLTS, while others, such as Health and Social Assistance, Government, and Accommodation and Food, have higher representation in JOLTS. Overall, however, there is a high correspondence in industries' vacancy shares in the two datasets.

Figure A.3: 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. Industries are sorted by their share of job ads in EMSI.

A.4 Additional Validation of Job Ads Data: Education Requirements in Job Ads v. ACS Employment, and Worker Search Behavior

In this section, again with the aim of validating the EMSI dataset, we perform two exercises. First, we compare education levels in job ads vs. ACS employed workers, across occupations and commuting zones. Second, we study the propensity of workers to search for jobs online,

[36] 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.

[37] 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." See https://www.bls.gov/help/def/jl.htm. Accessed February 23, 2021.

across large and small markets, for both college and non-college workers.

For each four-digit SOC $\times$ CZ, we compute the fraction of job ads requiring a BA degree or above (in 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.

A final potential concern is that firm recruiting strategies may differ between large and small markets, due to larger pool of applicants in large markets, which may create selection into the types of jobs posted. We indirectly test this concern by studying the search behavior of workers, using the CPS Computer and Internet Use Supplement for 2011-2017. We regress an indicator for using the internet to search for jobs on CZ size decile indicators and present the results in Figure A.7. The two panels—with and without occupation fixed effects—show no evidence that using the internet for job search varies with CZ size. We perform this analysis separately for workers with a BA or above and for workers with a high school degree only. None of these analyses reveal worker search behavior differing with market size. These results provide at least suggestive evidence that selection of job postings online is not a major concern.

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

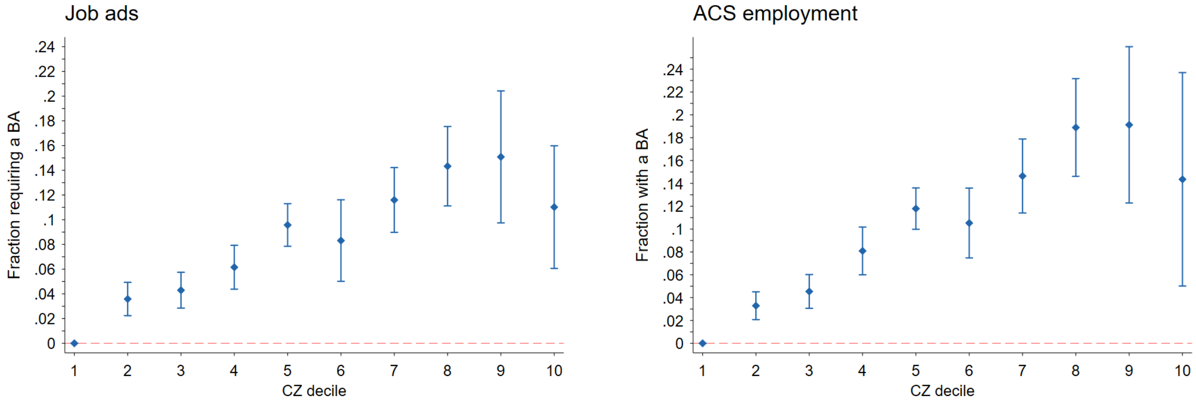
Each dot in the figure above corresponds to a four-digit SOC $\times$ market. The cells are weighted by employment. The y-axis corresponds to the fraction of workers in the ACS with at least a college degree. The x-axis corresponds to the fraction of job ads that require a BA degree or higher (among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

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

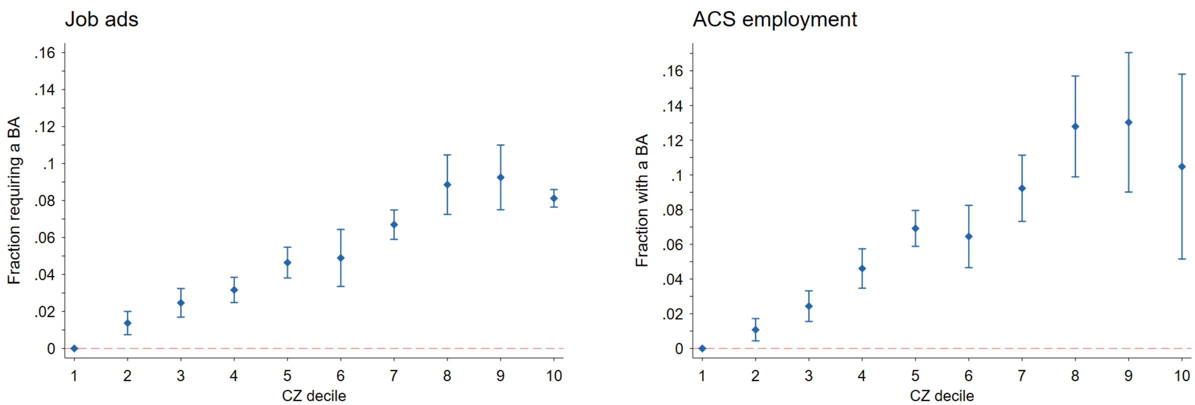
The figure above replicates Figure A.4 separately by CZ population quartile.

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

I. Without SOC f.e.

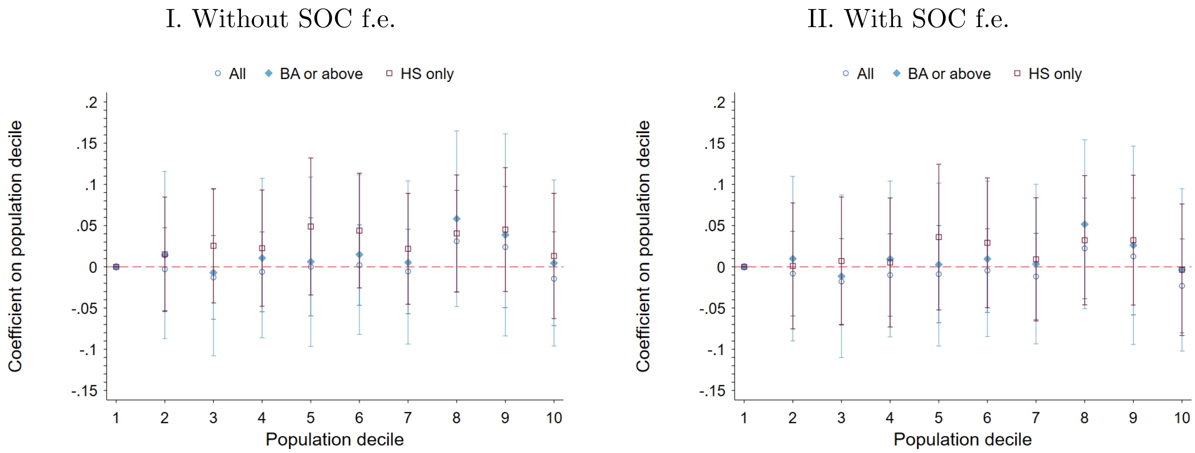


II. With SOC f.e.



Observations are four-digit SOC-CZ pairs. 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. 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 using the ACS. The bottom two panels reproduce the top two panels with four-digit fixed effects.

Figure A.7: Worker Job Search Using the Internet



The table above uses the CPS Computer and Internet Use Supplement for 2011-2017. The dependent variable is an indicator for the worker using the internet to search for jobs, which is regressed on a vector of deciles for CZ. Panel II includes fixed effects for CPS OCC2010 codes. Standard errors are clustered at the CZ level.

A.5 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. We also compare our data to within-occupation measures available from data collected by the Princeton Data Improvement Initiative (PDII) and find supportive evidence that our measures align with the PDII's within-occupation measures.

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 et al. (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 measures tasks only 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.[38] 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

[38] 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 et al. (2020) and in the word mappings on our project website, some task-related words are bigrams.

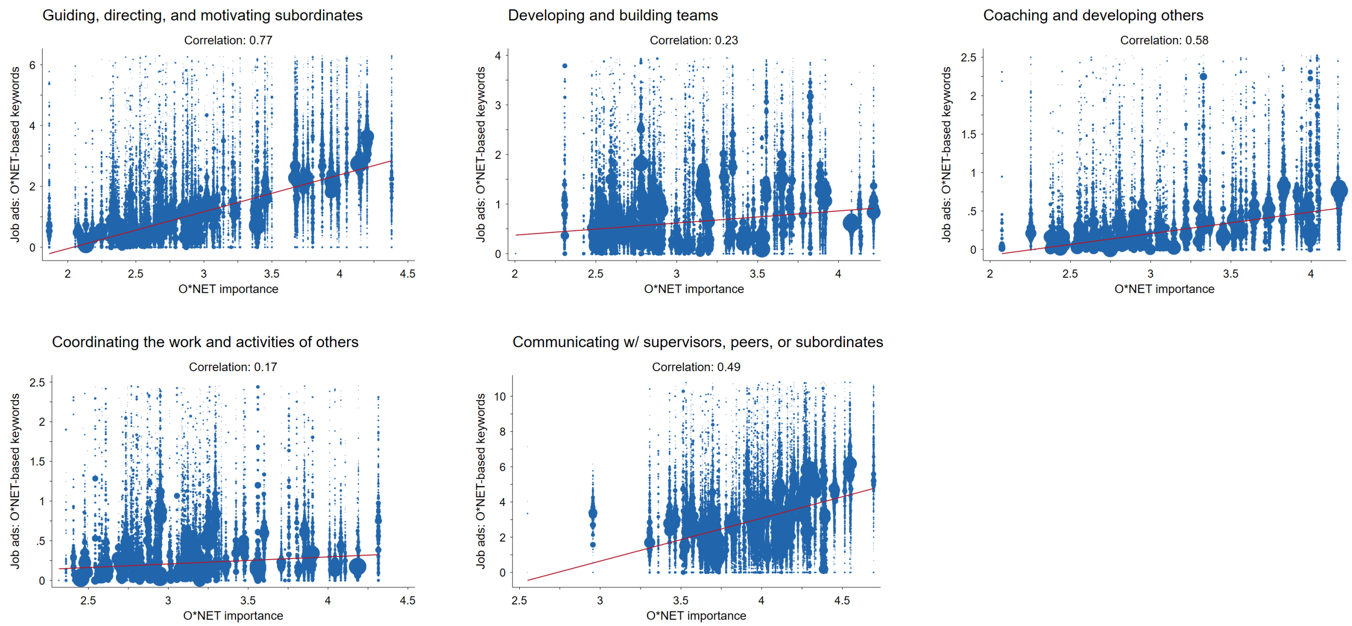
Figure A.8 demonstrates that our job ad-based task data have, for the most part, a high degree of correlation with O*NET tasks. We should not expect a perfect correlation, as O*NET itself has well-known limitations of small sample sizes, status quo bias, and subjective scales (Autor, 2013). But these correlations indicate that the job description text provides meaningful information about the task content of occupations.

Comparing Occupation-Level Market Size Gradients: O*NET v. Job Ads

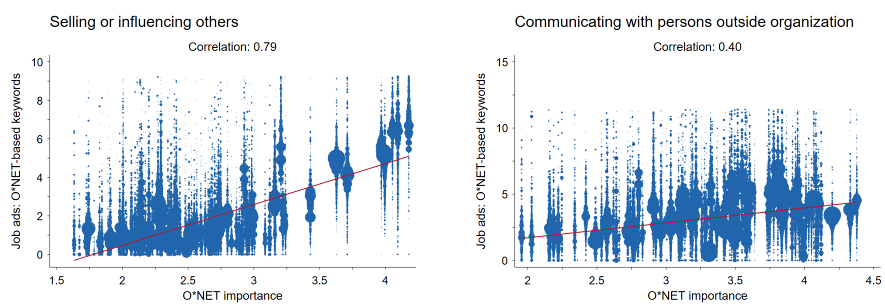
As additional evidence of the usefulness of job ads for studying job tasks in the labor market, we use occupation-level tasks extracted from job ads and compare these to widely-used occupation-level task measures from O*NET. We ask whether we would draw similar conclusions about the task gradients with market size using job ads as we would using O*NET, using a purely occupation-level analysis. To measure O*NET-based tasks, we adopt the O*NET items and categorization of Acemoglu and Autor (2011). We regress these tasks on CZ deciles, using ACS employment weights, and plot them in Figure A.9, panel I. We next construct our five task measures using job ads and applying the word mappings from Spitz-Oener. For this exercise, we constrain tasks to be fixed at the occupation level. We regress these tasks on CZ deciles using employment weights, and plot them in Figure A.9, panel II. The task gradients are strikingly similar across data sources, particularly for the non-routine analytic, non-routine interactive, routine cognitive and routine manual task categories, lending support to job ads data being useful for measuring tasks.

Figure A.8: Comparing Tasks from Job Ads with O*NET

I. Internal Interactive Tasks

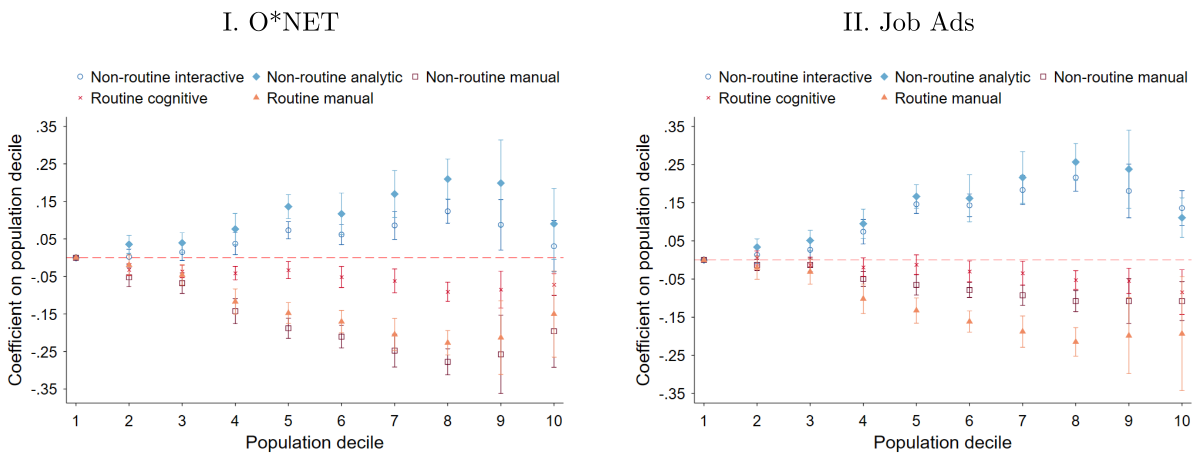


II. External Interactive Tasks



The figures above plot the correlations between occupation-level tasks extracted from the job ads to those based on O*NET. Each dot represents a four-digit SOC $\times$ 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.)

Figure A.9: Comparing Job Ads-based Tasks to O*NET-based Tasks



The left panel adopts the O*NET-based task measures and categories of Autor and Acemoglu (2011). O*NET items are averaged at the four-digit occupation level before being standardized to have mean zero and standard deviation 1 in the labor market. The job ads-based tasks are from the Spitz-Oener task categories that are first averaged at the four-digit occupation-level and then standardized to have mean zero and standard deviation 1 in the labor market. ACS employment in four-digit occupation-CZ cells are merged to the occupation-level task measures. Each task intensity measure is regressed on CZ deciles, with ACS employment weights, and with standard errors clustered at the CZ level.

Comparing Task Measures from EMSI to those in the Princeton Data Improvement Initiative

While useful, the preceding validation of our measurement of job task content relied solely on between-occupation task variation. In this section, we employ data from the Princeton Data Improvement Initiative (PDII)[39] to examine whether our measures align — looking within occupations as well — with those derived from existing datasets.[40]

[39]These data are described in Hallock (2013) and are used by Autor and Handel (2013) and Blau and Kahn (2013), among others.

[40]For our purposes, there are at least three advantages of EMSI over the PDII when measuring differences in job content between small versus large commuting zones. First, the PDII is drawn from a much smaller sample, with correspondingly large sampling error. Second, the geographic information available in the PDII include only the state of the survey respondent and whether the respondent was located in a metropolitan statistical area or not. Thus, with the PDII we cannot estimate CZ-size task gradients. Finally, the EMSI data permit measurement of individual tasks and technologies at a granular level across a wide variety of tasks and technologies, something that the PDII does not seek to measure. Nevertheless, the PDII provide the opportunity to validate our measurement of task categories using an existing, well-known dataset.

As part of the Princeton Data Improvement Initiative, 2,513 adults were asked a wide array of questions about the types of skills required and tasks performed in their jobs. From these questions, we construct indices of non-routine analytic tasks, non-routine interactive tasks, routine cognitive tasks, and manual tasks.[41] To compare the indices based on the PDII data to those based on EMSI job ads, we take the average values from each of the two datasets by four-digit SOC code and geography. The finest level of geography available in both datasets is the interaction of state and whether the individual resides in a metropolitan statistical area (MSA).

First, in Table A.4, we regress PDII-task measures on an indicator of whether the observation (a four-digit occupation-by-state-by-metropolitan status) is in a metropolitan area. The first four columns of this table present coefficient estimates from regressions without four-digit SOC fixed effects; the final four columns present results from specifications where these fixed effects are included. Overall, we find that individuals report spending more time on non-routine analytic tasks—and less time on routine cognitive tasks and manual tasks—in metro areas. These results align, for the most part, with those in Figure 1. There, we also report greater mentions of non-routine analytic tasks in larger CZs and fewer mentions of routine manual and non-routine manual tasks in larger CZs. The one (partial) discrepancy relates to routine cognitive tasks. In Figure 1, in specifications without fixed effects, there is a greater propensity for firms to mention routine cognitive tasks in larger CZs; this relationship disappears in specifications with SOC fixed effects. By contrast, in the PDII data, there is a lower intensity of routine cognitive tasks in metro areas.

[41] We consider (i) the frequency with which the individual takes 30 minutes to solve a problem, (ii) the frequency with which the individual uses math to solve problems, and (iii) the longest document typically read for a job as measures of non-routine analytic tasks; (i) the frequency of managing/supervising, and (ii) how much face-to-face contact with others as measures of non-routine interactive tasks; the frequency of short/repetitive tasks as our measure of routine cognitive skills; and the frequency of physical tasks as our measure of manual tasks. We could not find separate measures of routine manual and non-routine manual tasks. Before constructing each of these four indices, we standardize questions from each individual survey question. We then take the mean of these standardize values, and, finally, standardize the resulting indices.

Table A.4: Relationship Between PDII Task Measures and Metropolitan Status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Metro	0.197***	-0.008	-0.215***	-0.327***	0.080	0.005	-0.096**	-0.158***
	(0.057)	(0.061)	(0.047)	(0.057)	(0.052)	(0.058)	(0.048)	(0.044)
Dependent variable	NR-Analytic	NR-Interactive	R-Cognitive	Manual	NR-Analytic	NR-Interactive	R-Cognitive	Manual
R^2	0.009	0.000	0.011	0.024	0.371	0.307	0.334	0.535
Number of observations	1602	1609	1598	1607	1602	1609	1598	1607
SOC f.e.	No	No	No	No	Yes	Yes	Yes	Yes

An observation is a occupation-state-metro status triple. Observations are weighted equally. The dependent variable in each regression is the standardized index of task measures, using questions from the PDII. Standard errors are clustered at the four-digit SOC level.

In Tables A.5 and A.6, we regress measures PDII measures of task intensity against corresponding measures from our EMSI dataset. In Table A.5, occupation-by-state-by-MSA status observations are weighted equally, while in Table A.6 we weight observations according to the number of EMSI job ads in the occupation-by-state-by-MSA status triple. In unweighted specifications, we find that the PDII and EMSI data are correlated with other, but that these correlations disappear once conditioning on four-digit SOC. In weighted specifications, the two datasets' measures align not only between but also within occupations.

Overall, we conclude that patterns identified from our job ad data align reasonably well with those constructed using the PDII.

Table A.5: Relationship Between PDII and EMSI Task Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Non-routine	0.837***				-0.006			
Analytic	(0.089)				(0.109)			
Non-routine		0.285***				0.166		
Interactive		(0.091)				(0.127)		
Routine			0.145*				-0.021	
Cognitive			(0.075)				(0.111)	
Routine				0.432***				-0.007
Manual				(0.078)				(0.058)
Non-routine				0.221***				-0.036
Manual				(0.074)				(0.073)
Dependent variable	NR-Analytic	NR-Interactive	R-Cognitive	Manual	NR-Analytic	NR-Interactive	R-Cognitive	Manual
R^2	0.159	0.021	0.005	0.123	0.370	0.308	0.332	0.530
Number of observations	1602	1609	1598	1607	1602	1609	1598	1607
SOC f.e.	No	No	No	No	Yes	Yes	Yes	Yes

An observation is a occupation-state-metro status triple. Observations are weighted equally. The dependent variable in each regression is the standardized index of task measures, using questions from the PDII. Standard errors are clustered at the four-digit SOC level.

Table A.6: Relationship Between PDII and EMSI Task Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Non-routine Analytic	0.977***				0.799***			
	(0.098)				(0.220)			
Non-routine Interactive		0.359***				0.464**		
		(0.118)				(0.231)		
Routine Cognitive			0.343***				0.612**	
			(0.095)				(0.282)	
Routine Manual				0.506***				0.062
				(0.102)				(0.145)
Non-routine Manual				0.230**				-0.081
				(0.093)				(0.175)
Dependent variable	NR-Analytic	NR-Interactive	R-Cognitive	Manual	NR-Analytic	NR-Interactive	R-Cognitive	Manual
R^2	0.223	0.035	0.024	0.151	0.467	0.402	0.429	0.571
Number of observations	1602	1609	1598	1607	1602	1609	1598	1607
SOC f.e.	No	No	No	No	Yes	Yes	Yes	Yes

An observation is a occupation-state-metro status triple. Observations are weighted according to the number of ads in the EMSI data in the occupation-state-metro status triple. The dependent variable in each regression is the standardized index of task measures, using questions from the PDII. Standard errors are clustered at the four-digit SOC level.

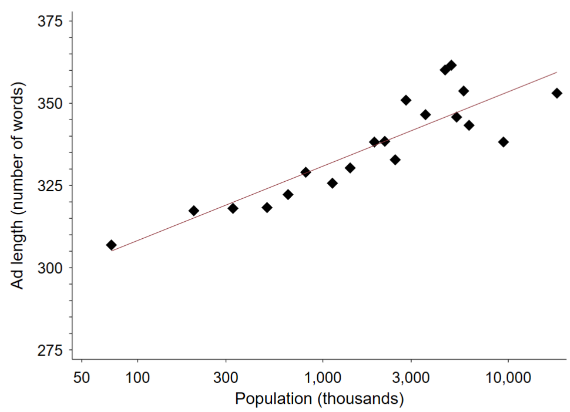
A.6 Job Ad Length and Description Keywords Across Space

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.10). 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.

As described in Section 2, the first step of our approach to extracting job tasks from the text to identify the part of the text corresponding to the job description. We use a set of keywords to identify this portion of the ad: "duties," "summary," "description," and "tasks." Figure A.11 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.

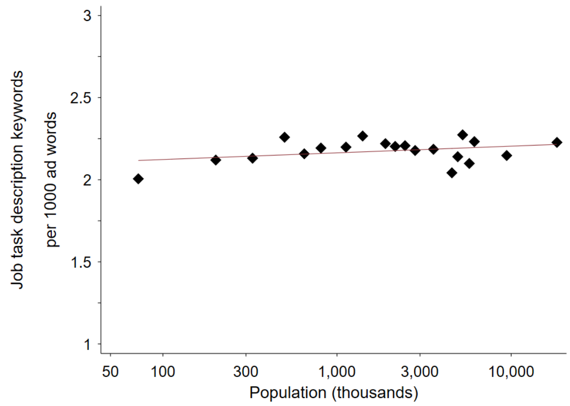
Lastly, we show that our novel task-extraction methodology—using job descriptions and parts of speech to let the text define the job tasks—passes a simple validation check. We calculate the cosine similarity between each job and the occupation-market average, and take the average. This exercise reveals 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 that the text contains valuable information about occupational characteristics that is reflected in standard occupational classifications.

Figure A.10: 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.11: Job Description Keywords Across Geography



The figure above presents a binned scatterplot of an indicator of the job ad's 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 illustrate the algorithm and present the most common tasks, the list of excluded tasks, and a scatterplot of the number of granular tasks and market size (Appendix B.1); present a validation exercise using multi-establishment firms (B.2); present the most common technologies (Appendix B.3); evaluate the relationships among tasks, technologies, and market size (Appendices B.4 and B.5); and 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 Task List

We first present two sample ads and the granular tasks extracted by the algorithm in Table B.1.

Table B.1: Illustrating the Algorithm to Extract Verb-Noun Tasks

Job Title	Job Ad Text	Tasks Extracted
Electrician	licensed electrician electronic control systems is seeking a full_time licensed electrician to **perform** commercial , residential , and industrial electrical **maintenance** and repair . candidates would be **assisting clients** in dade , bro ward and palm beach counties . candidate must be organized and motivated as we are looking for a person with skills and good working habits . specific responsibilities include , but are not limited to : assembling , installing , testing and maintaining electrical or electronic wiring , equipment , appliances , apparatus and fixtures **using hand** tools and power tools . diagnosing malfunctioning systems and components connecting wires to circuit breakers , transformers or other components . inspecting electrical systems , equipment and components to identify hazards , defects and the need for adjustment or repair , and to **ensure compliance** with codes . maintaining current electrician 's license or identification card to meet governmental regulations . . licensed electrician active journeyman electrician must be licensed 5 years of experience minimum (residential , commercial & industrial) proficient knowledge of local codes and safety regulations must speak fluent english work in dade , bro ward and palm beach counties must_have valid drivers_license and dependable transportation	perform maintenance, assisting clients, use hands, ensure compliance
Assistant Store Manager	general_summary : as a family dollar assistant store manager you will responsible for **providing** exceptional **service** to our customers . a key priority includes assisting the store manager in the daily operation of the store . under the direction of the store manager , you will also be responsible for **maintaining inventories** , **store** appearance and completing daily paperwork . principal duties & responsibilities : greets and **assists customers** in a positive , approachable manner . answers questions and resolves customer inquiries and concerns . maintains a presence in the store by **providing** excellent **customer_service** . **ensures** a clean , well_stocked **store** for customers . at the direction of the store manager , supervises , trains , and develops store team members on family dollar operating practices and procedures . assists in unloading all merchandise from delivery truck , organizes merchandise , and transfers merchandise from stockroom to store . **assists store** manager in ordering merchandise and record_keeping to include payroll , scheduling and cash_register deposits and receipts . supports store manager in loss_prevention efforts . assumes certain management responsibilities in absence of store manager . **follows** all **company** policies and procedures . bach f6f5fe bets arc setter maintaining store store .	provide service, maintaining inventory, maintain store, assisting customers, provide customer_service, ensure stores, assist store, following company

The table above presents the full text of two sample job ads and highlights in bold the verb-noun tasks extracted by our algorithm. Note that not all verb-noun pairs in the job ad text are highlighted as tasks because we define the set of tasks as the 500 most common verb-noun pairs.

Next, we list the 399 tasks we extract from the job ad text as verb-noun pairs along with the fraction of ads with each task ($\times$ 100). For readability, instead of listing the word stems we present the verb-noun pairs as they appear in their first occurrence, which leads to variations in noun forms and verb conjugations in the table (e.g., "provide service" vs. "providing support").

Table B.2: Tasks Extracted from Verb-Noun Pairs

written communication	13.0257	developed sales	0.8352	damaged merchandise	0.3108
working team	7.4251	communicate information	0.8348	move trays	0.3104
provide customer_service	6.6934	closes store	0.8229	needed customer_satisfaction	0.3092
provide service	5.3395	developing strategies	0.8218	increase customer_satisfaction	0.3044
lifting pounds	4.6136	working sales	0.8212	following pogs	0.3041
providing support	4.4229	writing skills	0.8198	responsibilities duties	0.3031
build relationships	3.8635	answering phones	0.8154	document counts	0.3024
ensure compliance	3.5870	increase sales	0.8052	assigned skills	0.3022
assisting customers	3.2288	maintaining environments	0.8014	may store	0.2908
provide customer	3.1077	handle tasks	0.7909	leads customers	0.2905
maintaining relationships	3.0468	support business	0.7870	maintaining program	0.2901
problem_solving skills	2.9784	ensure adherence	0.7739	executes store	0.2866
making decisions	2.9349	require walking	0.7711	supporting activities	0.2829
ensure customer	2.8990	ensure employees	0.7655	lead store	0.2827
lift lbs	2.8608	working variety	0.7644	serving quality	0.2689
provides quality	2.8342	assume responsibilities	0.7592	include staff	0.2668
provides leadership	2.5047	ensure completion	0.7577	maintain pharmacy	0.2627
develop relationship	2.5011	maintain productivity	0.7455	remove items	0.2540
perform job	2.4971	identifies problems	0.7329	requiring security	0.2536
leading team	2.3856	asking questions	0.7320	required paperwork	0.2522
achieve goals	2.2844	include service	0.7303	include hand	0.2513
working relationships	2.2757	providing environment	0.7301	seek customer	0.2444
continuing education	2.1940	writing reports	0.7265	lifting merchandise	0.2430
serving customers	2.1819	managing operations	0.7249	promote shopping	0.2401
following company	2.1392	including training	0.7245	merchandising product	0.2349
providing care	2.0627	providing expertise	0.7104	scheduling activities	0.2295
make recommendations	2.0457	ensure client	0.7027	set displays	0.2265
meet requirements	2.0141	assigned store	0.6921	has client	0.2240
meet deadlines	1.9775	maintain communication	0.6920	stored areas	0.2206
provides training	1.9577	assist development	0.6902	maintain card	0.2199

provided information	1.8973	generate sales	0.6839	training sessions	0.2183
will customers	1.8947	working departments	0.6815	conducting employee	0.2130
resolve issue	1.8601	using knowledge	0.6813	evaluates employees	0.2116
work flexible_schedule	1.8575	include development	0.6663	include shelves	0.2112
demonstrate knowledge	1.8571	answering telephone	0.6570	using phone	0.2054
taking actions	1.8503	develop productivity	0.6569	vacuum face	0.2037
provide feedback	1.8131	developing implement	0.6548	assigns directs	0.2007
provide assistance	1.8073	established guidelines	0.6539	using greet	0.1836
providing solutions	1.8068	maintain work_environment	0.6482	discontinued items	0.1835
driving sales	1.7791	preparing foods	0.6481	using orders	0.1808
ensure quality	1.7532	existing clients	0.6366	outdated merchandise	0.1800
helping customer	1.7479	ensure guests	0.6231	prepare returns	0.1797
works custom	1.7189	including work	0.6221	greeting card	0.1794
communicate customer	1.6945	maximizes profitability	0.6159	work stock	0.1765
follow instructions	1.6791	required driver	0.6138	securing company	0.1763
managing projects	1.6743	provide client	0.6136	crews customer_service	0.1761
maintain store	1.6554	meet clients	0.6114	recalled merchandise	0.1759
greeting customers	1.6384	set goals	0.6112	crew directing	0.1758
work shift	1.6339	including business	0.6068	change bulbs	0.1738
will teams	1.6264	are compliance	0.6046	labeling prescriptions	0.1735
answer questions	1.6252	move store	0.6043	maximizing customer_satisfaction	0.1723
ensure product	1.6196	provide technical_support	0.6015	needed in_store	0.1708
provide guidance	1.6020	provide recommendations	0.5896	reset departments	0.1703
detail ability	1.5925	opens store	0.5815	return system	0.1703
maintaining inventory	1.5885	obtain information	0.5811	signing maintain	0.1701
include sales	1.5879	ensuring team	0.5669	preventing trafficking	0.1699
written skills	1.5729	assigned supervisor	0.5577	windows ceilings	0.1698
work schedule	1.5256	requires merchandise	0.5567	windows removal	0.1690
achieving sales	1.5248	managing sales	0.5564	sweeping stock	0.1688
resolve problems	1.5085	include design	0.5528	signing shelves	0.1688
stand periods	1.4931	hiring training	0.5491	dump baskets	0.1688
maintaining standards	1.4602	ensure projects	0.5474	photofinishing orders	0.1688
assist store	1.4362	conducting research	0.5416	regarding cash_register	0.1688
meets customer	1.4272	assisting clients	0.5355	bags counter_tops	0.1687
work others	1.4230	assisted sales	0.5328	measuring drugs	0.1684
requires travel	1.4230	maintain awareness	0.5270	putting drug	0.1682

work week_ends	1.4150	include knowledge	0.5175	seal trays	0.1682
written instructions	1.3752	reaching pulling	0.5157	capping vials	0.1679
operating cash_register	1.3735	traveling store	0.5122	closing duties	0.1672
resolving customer	1.3628	unloading trucks	0.5120	make offer	0.1641
develop business	1.3594	move merchandise	0.5054	ensures quality_assurance	0.1606
maintain working	1.3569	develop test	0.5026	following reports	0.1567
maintain knowledge	1.3533	including performance	0.4901	communicating field	0.1554
providing direction	1.3523	including maintenance	0.4849	execute cash	0.1530
establish relationships	1.3468	supervising store	0.4845	returned check	0.1492
perform variety	1.3458	guided values	0.4785	following vendor	0.1492
ensure safety	1.3232	ensuring food	0.4728	execute display	0.1459
handling customer	1.3140	handle merchandise	0.4725	request help	0.1459
interact customers	1.3129	build customer	0.4707	including translation	0.1426
exceed sales	1.3000	make adjustments	0.4695	appropriate use	0.1422
ensure stores	1.2915	include merchandising	0.4597	perform register	0.1418
developing team	1.2807	manages business	0.4588	opening duties	0.1410
develop solutions	1.2723	taking orders	0.4545	executing set	0.1401
preferred ability	1.2457	ensuring communications	0.4525	sustained work	0.1397
using computer	1.2323	including systems	0.4524	pay policy	0.1393
maintain appearance	1.2284	meets standards	0.4505	securing door	0.1390
identify opportunities	1.2281	manage relationships	0.4499	execute completion	0.1379
weighing pounds	1.2267	including preparation	0.4490	pay vendors	0.1377
growing business	1.2217	ensure policies	0.4467	checking employee	0.1375
make changes	1.2214	comply state	0.4383	check_in merchandise	0.1374
maintain custom	1.2155	include program	0.4380	check acceptance	0.1371
existing customers	1.1991	ensure restaurant	0.4377	skating carhop	0.1368
on_going training	1.1942	may merchandise	0.4361	maintain prescription	0.1365
including nights	1.1743	may floor	0.4279	sustained periods	0.1365
work projects	1.1730	put customer	0.4249	pulls deposits	0.1360
develop planning	1.1620	scheduling appointments	0.4193	apprehend company	0.1358
stand walk	1.1526	assisting team	0.4184	document cash	0.1356
maximize sale	1.1489	providing coaching	0.4137	adapting store	0.1355
sells products	1.1478	have merchandise	0.4125	secure change	0.1352
written oral_communication	1.1286	including support	0.4115	identify shoplifters	0.1350
ensure customer_satisfaction	1.1274	causing discomfort	0.4102	react program	0.1350
operate equipment	1.1250	provides performance	0.4035	in_store repairs	0.1350

meet goals	1.1221	processing transactions	0.4030	resolve rejections	0.1350
use hands	1.1209	offer products	0.3978	organized pharmacy	0.1348
analyzing data	1.1207	include client	0.3976	signing crew	0.1348
meet sales	1.1067	containing materials	0.3974	react shoplifters	0.1347
prepare reports	1.1062	may slippery	0.3958	using enhancements	0.1346
assigned management	1.1047	maintain area	0.3946	execute walk_through	0.1346
according company	1.0815	receives service	0.3945	intern communication	0.1344
including management	1.0743	transforming delivery	0.3921	according hipaa	0.1344
engage customers	1.0722	maintain files	0.3918	locking setting	0.1340
provides input	1.0682	become slippery	0.3917	sweep room	0.1339
perform maintenance	1.0614	causing walking	0.3916	adjust facings	0.1335
prioritize tasks	1.0197	causing drafts	0.3916	trash rest	0.1335
managing teams	1.0034	appear floor	0.3915	dcr photofinishing	0.1335
ensure accuracy	1.0017	floors work	0.3912	bulletins action	0.1335
improving quality	1.0000	passing emit	0.3910	maintain pull	0.1335
team members	0.9907	include customer_service	0.3894	comply cvs	0.1332
establish policies	0.9903	focus team_work	0.3883	pharmacist communicate	0.1331
assisting management	0.9799	as_needed assist	0.3864	needed inventory_management	0.1330
maintain records	0.9741	retrieving information	0.3735	according cvs	0.1330
ensure delivery	0.9489	assist staff	0.3715	cvs workflow	0.1330
working store	0.9374	maintaining business	0.3691	greeting operations	0.1274
meet business	0.9364	include order	0.3660	sorting merchandise	0.1226
using equipment	0.9115	generating business	0.3639	delegated photo	0.1214
protect company	0.8972	staffing needs	0.3632	merchandising directives	0.1102
carry pounds	0.8943	establish priorities	0.3496	preventing terrorists	0.1075
ensuring merchandising	0.8941	bagging merchandise	0.3460	supervisor team	0.0957
following policies	0.8890	handling cash	0.3437	driving culture	0.0908
ensure operation	0.8781	procedures cash	0.3257	drive_in employees	0.0902
responding customer	0.8579	using eye	0.3249	identifying conditions	0.0699
ensure service	0.8539	taking vehicle	0.3210	assigned reading	0.0413
including cash	0.8443	maintained times	0.3133	customer_service culture	0.0241

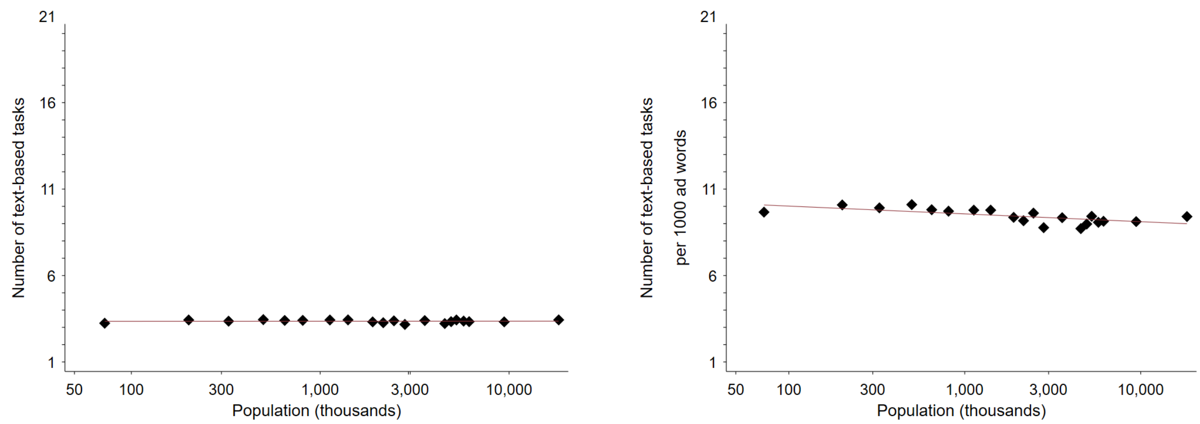
As described in the text, we exclude 101 tasks from the original list of 500 most common verb-noun pairs, using our judgment to select pairs that do not correspond to tasks. These excluded verb-noun pairs are presented below and describe worker skills (e.g., "high school diploma," "ged years," "required 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").

Table B.2 (continued): Verb-Noun Pair Drop List

be years	be doors	is job
is equal_opportunity	can doors	be company
arc bach	are business	perform duties
must years	requested react	be part
high_school diploma	are store	work environment
demonstrated ability	including evenings	perform functions
required employee	is law	required knowledge
bachelor degree	is customer	have experience
meet needs	earned degree	are position
required ability	is ability	have years
required years	send resume	required qualifications
required skills	s journal	is service
according state	eas program	includes ability
include customers	is delivery	committed diverse
work hours	are company	are sales
are customers	ged years	knowledge skills
be customer	include duties	working business
preferred years	required position	desired skills
required experience	be duties	providing product
s degree	pass drug	be lbs
arc setter	required bachelor	are manages
end caps	are accordance	are duties
preferred experience	sporting goods	is walks
including products	have ability	will career
is position	based business	are reporting
work part	ensuring aspects	according needs
are time	assigned job	permitted law
ensure execution	be ability	performing tasks
bach bets	may duties	playing role
be team	are fast_growing	preferred knowledge
travel travel	requires state	achieve results
is experience	must_have driver	completing tasks
may materials	will business	performing work
are drafts	s level	

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 399 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.2 Comparing Tasks in a Firm's Headquarters v. Other Establishments

In this section, we compare the intensity of tasks in a firm's headquarters relative to its other establishments. We find that systematic differences align with our priors about the tasks that take place in headquarters. As a result, we conclude that the EMSI data and our extraction of granular tasks provide a useful new characterization of differences in work activities between geographies.

For this exercise, we limit ourselves to the 10 largest firms, measured by total job postings, and exclude chains and postings by government agencies. We identify the headquarters location for each firm as the CZ with the largest number of the firm's postings and then validate this list against public records. The list of firms used for this validation exercise, along with their headquarters location, are: Amazon (Seattle, WA), Genesis HealthCare (Kennett Square, PA), UnitedHealth Group (Minnetonka, MI), IBM (Armonk, NY), HCA Healthcare (Nashville, TN), Lockheed Martin Corporation (Bethesda, MD), Aramark Corporation

(Philadelphia, PA), Providence Health and Services (Renton, WA), Citigroup Incorporated (New York, NY), Parallon (Nashville, TN). This subsample includes 136,324 ads.We run a regression at the job-ad level for each task. We regress task intensity on an indicator for the job being in the firm headquarters, along with six-digit SOC fixed effects, and with the standard errors clustered at the CZ-level. In Table B.3, we report the tasks with the largest positive and negative coefficient estimates on headquarters (after standardizing, by dividing by the task standard deviation). The list of largest positive and negative gradients are presented. A clear pattern emerges, which is that the headquarters locations require management, teamwork, or span of control: managing projects, communication (both written and oral), analyzing data, and identifying opportunities are all tasks that reflect these types of work activities. The tasks with the largest negative gaps—i.e., tasks that are common in non-headquarters locations relative to the headquarters—involve training and working with or assisting clients. Overall, we view these intuitive differences as an additional validation of our approach.

Table B.3: Tasks with Largest Gap Between Headquarters and Non-Headquarters Locations

Positive gradient		Negative gradient	
Task	$\hat{\beta}_{hq}$	Task	$\hat{\beta}_{hq}$
written communication	0.2714	on_going training	-0.3949
managing projects	0.2041	provides training	-0.3694
growing business	0.1910	work projects	-0.3409
written oral_communication	0.1717	requires travel	-0.2930
detail ability	0.1603	existing clients	-0.1394
will teams	0.1359	include client	-0.1310
analyzing data	0.1292	work others	-0.1098
support business	0.1243	assisting clients	-0.1039
working team	0.1235	meet clients	-0.1011
sells products	0.1106	provide assistance	-0.0897
identify opportunities	0.0923	stand walk	-0.0800
provides quality	0.0910	written skills	-0.0785
meet deadlines	0.0857	assume responsibilities	-0.0772
serving customers	0.0851	providing solutions	-0.0593
seek customer	0.0847	increase sales	-0.0591

The table above is based on a subsample of 136,324 ads from the following multi-establishment firms: Amazon, Genesis HealthCare, UnitedHealth Group, IBM, HCA Healthcare, Lockheed Martin Corporation, Aramark Corporation, Providence Health and Services, Citigroup Incorporated, and Parallon. We run a regression at the job-ad level for each of the 399 tasks. We regress the task intensity on an indicator for the job being in the firm headquarters, along with six-digit SOC fixed effects, and with the standard errors clustered at the CZ-level. We report the tasks with the largest positive and negative coefficient estimates on headquarters (after standardizing, by dividing by the task standard deviation). The list of the 15 largest positive and negative gradients are presented. All estimates are significant at the 5 percent level.

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 ($\times$ 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 that the words appear in order.

For the three social media technologies in the list (Facebook, YouTube, and LinkedIn), we explicitly search for and exclude false positives in our analysis. To identify false positives, we search for phrases that strongly suggest the ad is directing the reader to visit or follow the firm on social media. For example, any of the following bracketed phrases along with the mention of "facebook" would be flagged as a false positive for the Facebook technology: "[fan us][visit us][like us][connect with us][follow us][check us out][for more information][please visit][share this job][how did you hear][look for us][learn more about] ... facebook." We perform the analogous exercise to create false positive flags for YouTube and LinkedIn. We conducted robustness to our method of identifying false positives, such as creating a "true positive" flag that explicitly identifies the phrase "social media" along with other words, such as "knowledge," "experience," or "proficiency" in the ad, and the results are unchanged.

Table B.4: Technologies Extracted from Job Vacancy Data (with Frequency per 100)

microsoft excel	2.0566	apache hive	0.0135
sap	1.4853	geographic information system gis software	0.0134
linux	1.4065	microsoft dynamics gp	0.0133
microsoft project	1.3218	transact-sql	0.0132
microsoft word	1.1720	unified modeling language uml	0.0125
javascript	1.1669	apache cassandra	0.0119
unix	1.0452	apache pig	0.0097
microsoft office	1.0363	extensible markup language xml	0.0077
microsoft access	0.8903	cascading style sheets css	0.0077
microsoft windows	0.8149	oracle business intelligence enterprise edition	0.0076
react	0.7996	apache kafka	0.0071
microsoft outlook	0.7230	spring boot	0.0071
python	0.7208	integrated development environment ide software	0.0068
c++	0.7007	delphi technology	0.0065
microsoft powerpoint	0.6548	apache groovy	0.0060
microsoft sql server	0.5013	adobe systems adobe creative cloud	0.0057
oracle java	0.4844	enterprise resource planning erp software	0.0054
chef	0.4732	atlassian bamboo	0.0053
sas	0.4551	virtual private networking vpn software	0.0046
ruby	0.4071	node.js	0.0045
tax software	0.3962	ibm spss statistics	0.0045
ajax	0.3503	google angularjs	0.0037
mysql	0.3412	hypertext markup language html	0.0036
git	0.2910	job control language jcl	0.0030
swift	0.2735	apache subversion svn	0.0019
microsoft sharepoint	0.2653	oracle hyperion	0.0015
citrix	0.1815	backbone.js	0.0014
microsoft visio	0.1793	customer information control system cics	0.0013
facebook	0.1707	oracle primavera enterprise project portfolio management	0.0013

nosql	0.1579	adobe systems adobe aftereffects	0.0009
tableau	0.1526	microsoft asp.net	0.0007
linkedin	0.1426	practical extraction and reporting language perl	0.0007
bash	0.1416	ca erwin data modeler	0.0006
microsoft visual studio	0.1412	microsoft active server pages asp	0.0002
microsoft dynamics	0.1411	common business oriented language cobol	0.0001
relational database management software	0.1397	salesforce software	0.0001
microsoft exchange server	0.1342	google analytics	0.0001
google drive	0.1230	computer aided design cad software	0.0001
epic systems	0.1166	qlik tech qlikview	0.0000
objective c	0.1140	ibm websphere	0.0000
microsoft sql server reporting services	0.1110	junit	0.0000
selenium	0.1097	oracle peoplesoft	0.0000
puppet	0.1069	microsoft .net framework	0.0000
spring framework	0.1022	microsoft asp.net core mvc	0.0000
apache tomcat	0.1010	yardi	0.0000
data entry software	0.0952	oracle taleo	0.0000
microsoft visual basic	0.0860	national instruments labview	0.0000
symantec	0.0858	oracle pl/sql	0.0000
mongodb	0.0846	splunk enterprise	0.0000
youtube	0.0825	marketo marketing automation	0.0000
red hat enterprise linux	0.0769	healthcare common procedure coding system hcpcs	0.0000
ruby on rails	0.0690	adobe systems adobe indesign	0.0000
postgresql	0.0617	microsoft powershell	0.0000
microsoft azure	0.0549	c#	0.0000
shell script	0.0532	the mathworks matlab	0.0000
scala	0.0508	aws redshift	0.0000
teradata database	0.0492	microstrategy	0.0000
drupal	0.0486	handheld computer device software	0.0000
nagios	0.0476	google adwords	0.0000
confluence	0.0466	minitab	0.0000
verilog	0.0458	netsuite erp	0.0000
adobe systems adobe acrobat	0.0457	autodesk autocad civil d	0.0000
mcafee	0.0448	oracle weblogic server	0.0000
docker	0.0442	medical procedure coding software	0.0000
oracle jdbc	0.0439	apple macos	0.0000

adobe systems adobe photoshop	0.0438	microsoft visual basic scripting edition vbscript	0.0000
intuit quickbooks	0.0433	smugmug flickr	0.0000
eclipse ide	0.0408	oracle jd edwards enterpriseone	0.0000
fund accounting software	0.0348	enterprise javabeans	0.0000
apache hadoop	0.0337	dassault systemes catia	0.0000
adobe systems adobe illustrator	0.0325	apache solr	0.0000
oracle fusion applications	0.0322	trimble sketchup pro	0.0000
google docs	0.0314	wireshark	0.0000
ubuntu	0.0307	red hat wildfly	0.0000
apache maven	0.0298	ibm infosphere datastage	0.0000
django	0.0282	adobe systems adobe dreamweaver	0.0000
structured query language sql	0.0282	github	0.0000
apache http server	0.0250	medical condition coding software	0.0000
hibernate orm	0.0245	javascript object notation json	0.0000
meditech software	0.0237	elasticsearch	0.0000
apache ant	0.0231	oracle javaserver pages jsp	0.0000
ansible software	0.0229	php: hypertext preprocessor	0.0000
autodesk autocad	0.0219	supervisory control and data acquisition scada software	0.0000
ibm notes	0.0186	advanced business application programming abap	0.0000
atlassian jira	0.0182	oracle solaris	0.0000
adp workforce now	0.0178	blackbaud the raiser's edge	0.0000
apache struts	0.0156	bentley microstation	0.0000
sap crystal reports	0.0148	dassault systemes solidworks	0.0000
esri arcgis software	0.0146	autodesk revit	0.0000
jquery	0.0140	ibm cognos impromptu	0.0000

B.4 Tasks and Market Size

Tables B.5 and B.6 reproduce Tables 2 and 3, respectively, except with a continuous measure of market size on the right-hand side—log population—rather than market size decile indicators. The tasks with the largest positive and negative gradients are similar to those presented in Section 3. Table B.7 re-estimates equation (1) using a predetermined list of verbs from Michaels et al. (2018) instead of our task list extracted from the text itself. The takeaway is quite similar. Using only the Michaels et al. (2018) 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.5: Tasks with the Steepest Gradient in Log Population

Positive Gradient				Negative Gradient			
No SOC f.e.		SOC f.e.		No SOC f.e.		SOC f.e.	
Task-Population	$\hat{\beta}$	Task	$\hat{\beta}$	Task-Population	$\hat{\beta}$	Task	$\hat{\beta}$
written communication	0.0451	written skills	0.0134	maintain store	-0.0414	maximizes profitability	-0.0303
managing projects	0.0367	achieving sales	0.0121	operating cash_register	-0.0393	protect company	-0.0288
providing support	0.0294	ensure safety	0.0115	provide customer_service	-0.0380	maintain store	-0.0272
develop solutions	0.0269	stand walk	0.0113	protect company	-0.0357	operating cash_register	-0.0256
problem_solving skills	0.0268	prioritize tasks	0.0111	maximizes profitability	-0.0350	make changes	-0.0236
meet deadlines	0.0263	providing coaching	0.0106	greeting customers	-0.0321	greeting customers	-0.0217
work projects	0.0234	driving sales	0.0105	assist store	-0.0313	procedures cash	-0.0215
support business	0.0230	supervising store	0.0104	make changes	-0.0298	skating carhop	-0.0198
written skills	0.0219	providing environment	0.0102	maintaining inventory	-0.0274	unloading trucks	-0.0186
developing strategies	0.0217	meet deadlines	0.0101	procedures cash	-0.0273	ensure employees	-0.0185
provide guidance	0.0212	written communication	0.0100	following company	-0.0264	drive_in employees	-0.0177
identify opportunities	0.0207	identify opportunities	0.0099	preventing trafficking	-0.0264	assigned store	-0.0176
develop business	0.0205	exceed sales	0.0097	unloading trucks	-0.0257	maintaining inventory	-0.0175
will teams	0.0200	provide feedback	0.0097	skating carhop	-0.0255	provide customer_service	-0.0173
working team	0.0198	managing projects	0.0096	assigned store	-0.0251	working store	-0.0161

The table above reproduces Table 2 except replaces the right-hand side market size decile indicators with a continuous log population measure. The coefficients above present the tasks with the steepest positive and negative gradients with respect to market size, as captured by $\hat{\beta}$ on the continuous population measure. All coefficients are statistically significant at the 1 percent level, except "supervising store" in column 3 ($p = 0.091$).

Table B.6: Technologies with the Steepest Gradient in Log Population

All		**College**		**High School**	
Technology	$\hat{\beta}$	Technology	$\hat{\beta}$	Technology	$\hat{\beta}$
Linux	0.0361	Python	0.0352	Microsoft Excel	0.0174
JavaScript	0.0324	Linux	0.0327	Microsoft Outlook	0.0127
Python	0.0319	JavaScript	0.0317	Microsoft Office	0.0108
Unix	0.0309	Unix	0.0283	Microsoft Word	0.0102
Microsoft Excel	0.0288	Ruby	0.0249	Chef	0.0097
Microsoft Project	0.0272	C++	0.0246	Microsoft Powerpoint	0.0088
C++	0.0256	SAS	0.0222	Microsoft Access	0.0086
Oracle Java	0.0220	Microsoft Project	0.0219	Linux	0.0068
SAP	0.0215	Oracle Java	0.0215	Epic Systems	0.0065
Microsoft Access	0.0209	Microsoft Excel	0.0203	Swift	0.0061
SAS	0.0205	Git	0.0198	Citrix	0.0060
MySQL	0.0202	Ajax	0.0198	Tax Software	0.0056
Git	0.0199	MySQL	0.0196	Facebook	0.0056
Microsoft Office	0.0196	Tableau	0.0195	Microsoft Sharepoint	0.0056
Microsoft Powerpoint	0.0193	NoSQL	0.0187	Python	0.0054

We reproduce Table 3 by re-estimating equation 1 except replacing the right-hand side market size decile indicators with a continuous log population measure. All estimates are statistically significant at the 1 percent level, with the following exceptions: C++ in the college column ($p = 0.025$) and Swift in the high school column ($p = 0.017$).

Table B.7: Verbs with the Steepest Gradient

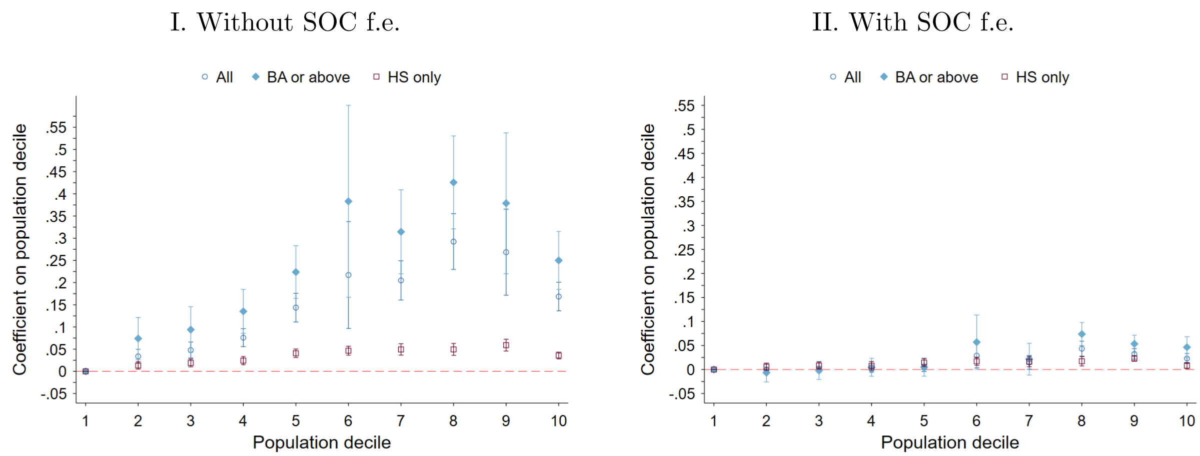
Positive gradient		Negative gradient	
Task	$\hat{\beta}_{10}$	Task	$\hat{\beta}_{10}$
design	0.0812	pay	-0.0625
project	0.0797	truck	-0.0623
experience	0.0660	store	-0.0559
research	0.0632	earn	-0.0513
develop	0.0616	clean	-0.0506
manage	0.0581	license	-0.0452
web	0.0560	fuel	-0.0448
finance	0.0499	get	-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 2, except it uses the list of verbs from Michaels et al. (2018). This exercise is conducted on a 1 percent sample of all job ads, rather than 5 percent, for computational speed, since the verb list includes 1,665 verbs. All estimates are statistically significant at the 1 percent level.

B.5 Technology Requirements and Market Size

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.

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, 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 ad-level wages and these are regressions of mean wages on mean tasks, using variation across geography-education cells. 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 capture only 0.1 percent of wage variation beyond occupation categories. Column 4 adds to the regression the granular task measures averaged to the occupation-education-market cell. 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 diploma 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 diploma 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***	0.043***	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***	-0.003	-0.025***	-0.014
	(0.004)		(0.004)	(0.004)	(0.005)	(0.011)
Routine manual	0.059***		-0.018***	-0.009*	-0.020***	-0.056***
	(0.005)		(0.006)	(0.005)	(0.006)	(0.011)
Non-routine manual	0.040***		0.010*	0.001	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	0.803	0.552	0.694
Adjusted R^2		0.784	0.785	0.802	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 occupation-education-market cell. Education category dummies are included in columns 1-4. 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 four-digit SOC f.e., tasks, log population, and log population interacted with tasks. Tasks are standardized to have mean zero and 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 Appendix to Sections 3.1 and 3.2

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

Within-Between Decompositions

To further evaluate how much of 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, $t_{kq} = \sum_{o \in \mathcal{O}} t_{koq} s_{oq}$, where the average task content of each occupation o in quartile q, t_{koq}, is multiplied by occupation o's share of quartile q's employment, s_{oq}. We express the difference in task content between two quartiles, q and $\tilde{q}$, as

$$t_{kq} - t_{k\tilde{q}} = \sum_{o \in \mathcal{O}} (t_{koq} - t_{ko\tilde{q}}) \bar{s}_{oq\tilde{q}} + \sum_{o \in \mathcal{O}} \bar{t}_{koq\tilde{q}} (s_{oq} - s_{o\tilde{q}}), \tag{5}$$

where $\bar{s}_{oq\tilde{q}} = (s_{oq} + s_{o\tilde{q}})/2$ and $\bar{t}_{koq\tilde{q}} = (t_{koq} + t_{ko\tilde{q}})/2$. The first term on the right-hand side of equation (5) 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, 14 percent of the variation between 1st quartile and 4th quartile CZs is within occupation. For non-routine interactive tasks, the corresponding figure is 22 percent. This result implies that standard data sources fail to capture much of the variation in tasks between small and large labor markets.

We perform the decomposition on each of our granular task measures to understand how much of the variation in these tasks across markets occurs within occupations versus between occupations. We calculate the decomposition shares for each of the granular tasks and report the median. We find that 18 percent of the variation from smallest to largest quartile CZs occurs within occupations. Applying the equation (5) decomposition to the number of technologies, we find that about 79 percent of the variation in technologies between 1st quartile CZs and 4th quartile CZs occurs between occupations and about 21 percent within occupations.

Table C.2 examines the sensitivity of our within-between decomposition of Table C.1 to measurement error. To do so, we randomly assign ads to population quartiles, in proportion to the actual distribution of six-digit SOCs across quartiles. We then reproduce the within-between decomposition exercise of Table C.1. Since we randomize ads to quartiles, the within shares should be close to zero or exactly zero. Table C.2 shows that most of the within shares are close to zero and most of the between shares are close to one, as we would expect if measurement error were not a major concern. This pattern notably holds for the Q4-Q1 decompositions. There are a couple of exceptions to this pattern, such as the routine cognitive decompositions for the Q4-Q3 and Q3-Q2 differences. However, in those cases the within shares are less meaningful, as the differences in average tasks (the denominators in the decompositions) are near-zero. Overall, the main takeaway from Table C.2 is that our within-between decompositions are robust to measurement error.

Evidence for Jobs Being Jointly Intensive in Interactive and Analytic Tasks

We consider whether jobs that are jointly intensive in interactive and analytic tasks appear predominantly in large markets. 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 CZ population 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.1. We find that jobs that are intensive in *both* analytic and interactive tasks make up 12.3 percentage points more of jobs in the highest decile compared with the lowest decile. Jobs that are intensive in only analytic tasks but not interactive tasks make up only about 3.4 percentage points more of jobs in the highest decile, 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.1.

Internal and External Interactive Tasks by Education

In Figure C.2, 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 fixed effects, the difference between these gradients is minor.

Sensitivity to Time Period

In Figure C.3, we explore whether the key tasks and technologies gradients of Figures 1 and 3 might be sensitive to the time period studied. Specifically, a potential concern is that a rapidly changing labor market in large v. small CZs might generate changing gradients over time. To explore this issue, we divide the sample period into two approximately equal periods, 2012-2014 and 2015-2017, and re-estimate panel I of each of the two figures. The results are highly stable across the two time periods.

Market Density and Tasks and Technologies

We next assess whether the task and technology patterns with respect to CZ population size, observed in Figures 1-4, also hold with respect to CZ population density. We construct CZ population density deciles. Then, we reproduce Figures 1-4 using population density deciles, as opposed to population size deciles, as the explanatory variables of interest. The coefficient estimates are presented in Figures C.4-C.7.

Table C.1: Task Decomposition Across Markets

	NR-Analytic		NR-Interactive		NR-Manual		R-Cognitive		R-Manual		Granular Tasks		Technologies	
Q1	5.30		5.66		0.81		0.71		2.74				0.12	
Q2	6.40		6.10		0.82		0.78		2.49				0.20	
Q3	7.51		6.49		0.77		0.78		2.20				0.32	
Q4	7.78		6.84		0.74		0.76		2.09				0.34	
	Between and Within Occupational Decomposition													
	Between	Within	Between	Within	Between	Within	Between	Within	Between	Within	Between	Within	Between	Within
Q2-Q1	0.80	0.20	0.79	0.21	1.35	-0.35	1.06	-0.06	0.50	0.50	0.57	0.43	0.84	0.16
Q3-Q2	0.92	0.08	0.83	0.17	0.53	0.47	-4.12	5.12	0.59	0.41	0.61	0.39	0.79	0.21
Q4-Q3	1.01	-0.01	0.70	0.30	0.72	0.28	0.07	0.93	0.72	0.28	0.04	0.96	0.36	0.64
Q4-Q1	0.86	0.14	0.78	0.22	0.51	0.49	1.68	-0.68	0.57	0.43	0.82	0.18	0.79	0.21

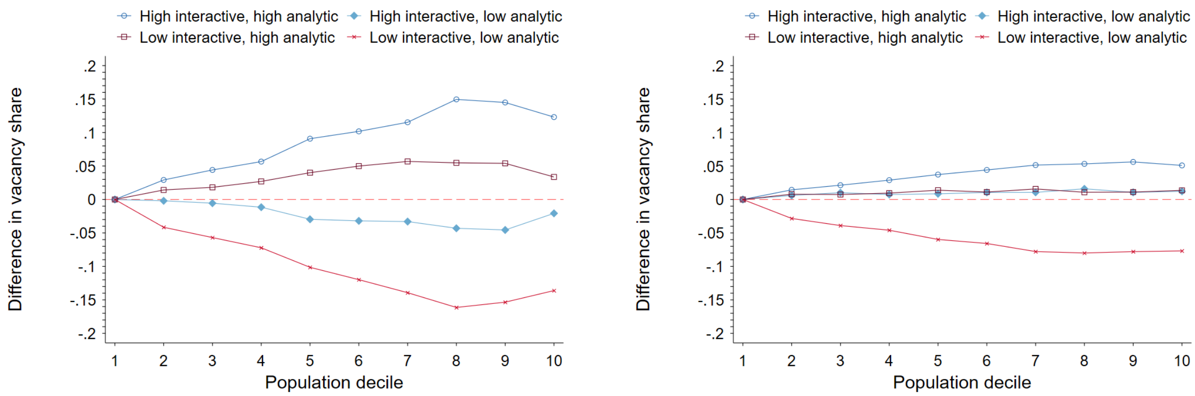
This table presents the results of the decomposition of equation (5) for each of the five task categories (non-routine analytic, non-routine interactive, non-routine manual, routine cognitive, and routine manual), each granular task (reporting the median), and the number of technologies. The top panel reports the average task content and the average number of technologies in each of four market size quartiles. Tasks in the top panel are expressed as number of task-word mentions per 1,000 ad words, and technologies are expressed as the average number of technologies mentioned per ad. The bottom panel presents the within and between shares of the total difference between population quartiles. The decomposition for granular tasks is constructed by first calculating the decomposition for each of the 399 tasks separately and then taking the median of the within and between shares.

Table C.2: Task Decomposition Across Markets: Sensitivity to Measurement Error

	NR-Analytic		NR-Interactive		NR-Manual		R-Cognitive		R-Manual		Granular Tasks		Technologies	
Q1	5.51		5.78		0.80		0.69		2.58				0.14	
Q2	6.41		6.14		0.81		0.77		2.45				0.21	
Q3	7.41		6.46		0.78		0.79		2.27				0.31	
Q4	7.70		6.71		0.75		0.79		2.19				0.32	
	Between and Within Occupational Decomposition													
	Between	Within	Between	Within	Between	Within	Between	Within	Between	Within	Between	Within	Between	Within
Q2-Q1	1.00	0.00	0.99	0.01	0.93	0.07	1.01	-0.01	0.99	0.01	0.98	0.02	1.00	-0.00
Q3-Q2	1.01	-0.01	1.03	-0.03	0.94	0.06	0.85	0.15	0.97	0.03	0.99	0.01	1.00	0.00
Q4-Q3	0.96	0.04	0.98	0.02	1.02	-0.02	0.13	0.87	1.04	-0.04	0.86	0.14	0.99	0.01
Q4-Q1	1.00	0.00	1.00	-0.00	0.98	0.02	0.99	0.01	1.00	0.00	1.00	0.00	1.00	0.00

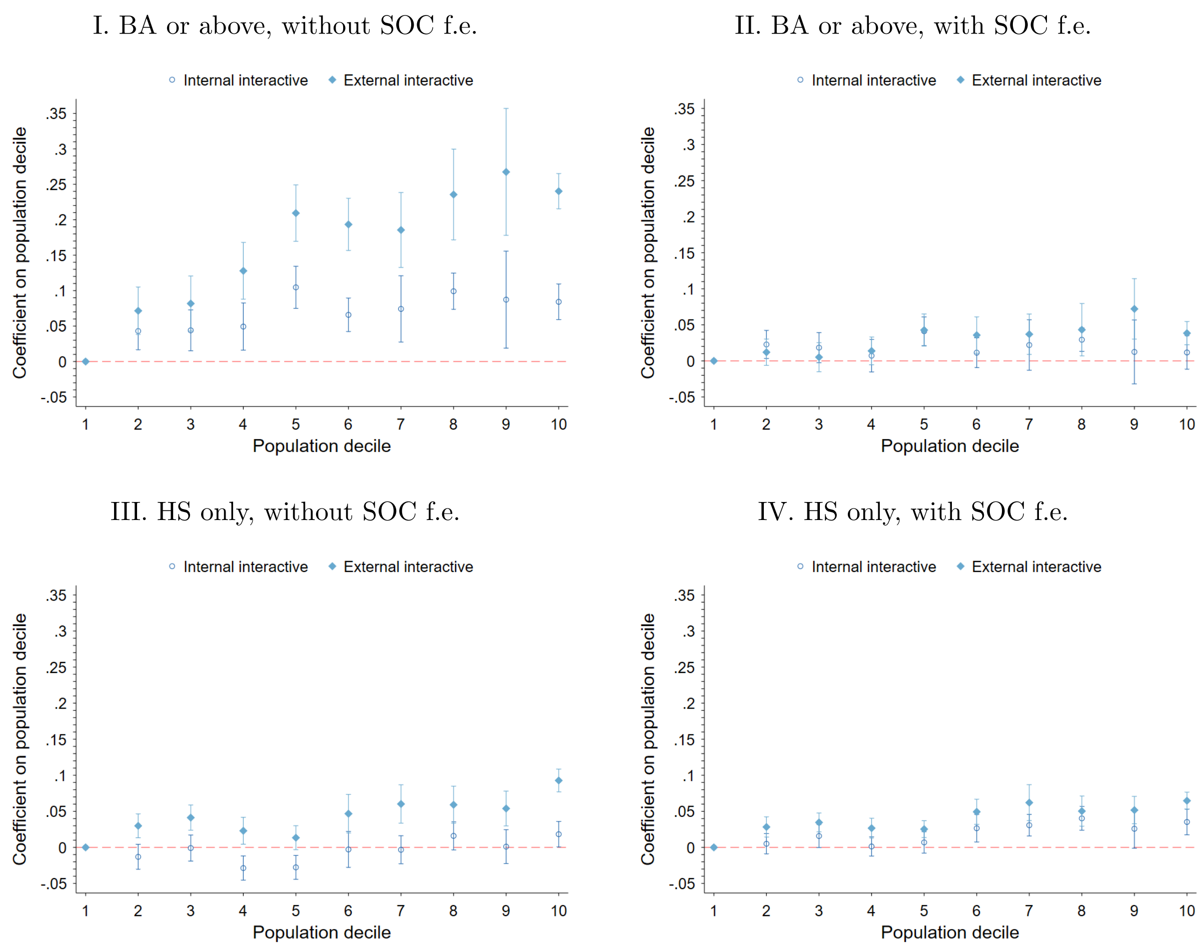
This table checks the sensitivity of Table C.1 to measurement error. To do so, we randomize job ads to population quartiles, based on the actual distribution of six-digit SOCs across quartiles. We then reproduce the decomposition of Table C.1. Since ads are randomized to quartiles, the within shares should be approximately zero.

Figure C.1: 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.2: O*NET Interactive Tasks Gradient

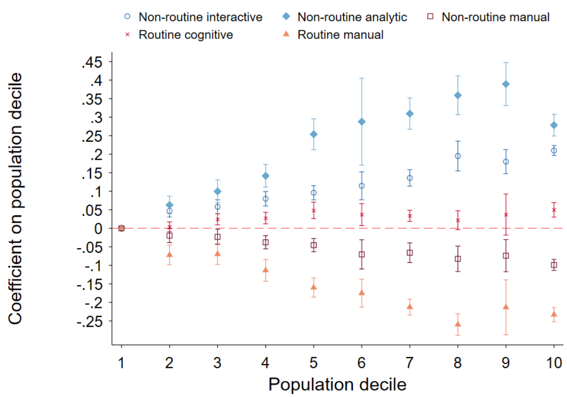


This figure reproduces Figure 2 separately by the educational requirement of the job. Panels I and II restrict the sample to ads requiring a BA or above, while panels III and IV restrict the sample to ads requiring high school only.

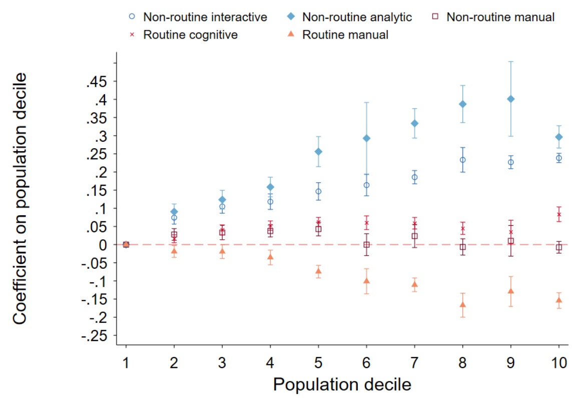
Figure C.3: Tasks and Technologies Gradient by Sample Period

A. Tasks

I. 2012-2014

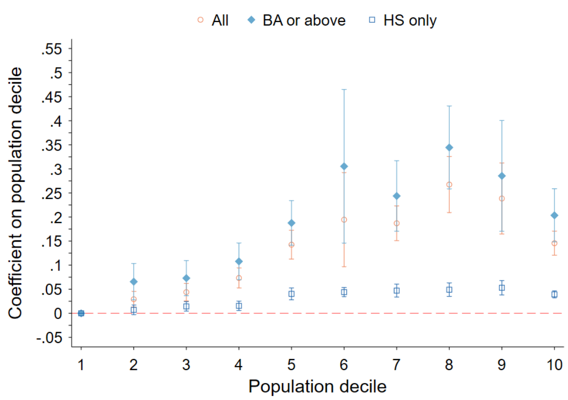


II. 2015-2017

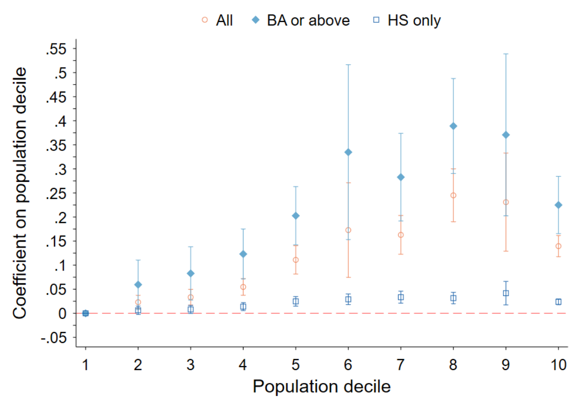


B. Technologies

III. 2012-2014



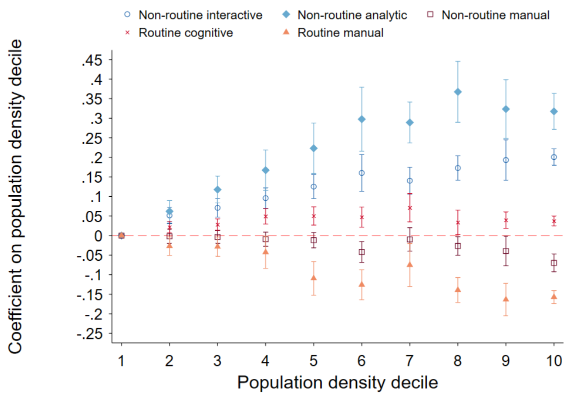
IV. 2015-2017



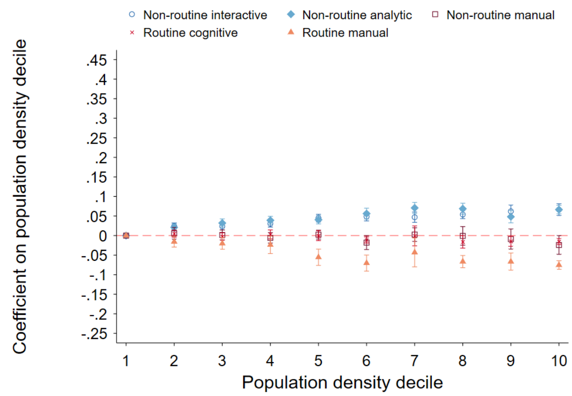
This figure presents estimates of Figure 1, panel I, and Figure 3, panel I, separately by time period. We divide the sample period into 2012-2014 and 2015-2017.

Figure C.4: Tasks and Market Density

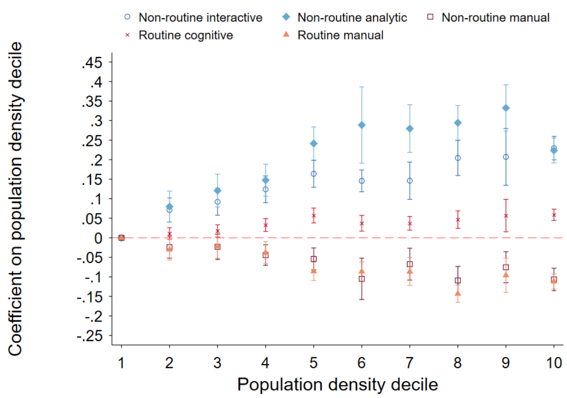
I. All ads, without SOC f.e.



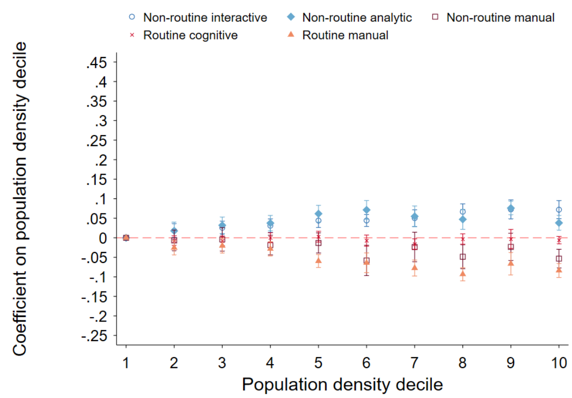
II. All ads, with SOC f.e.



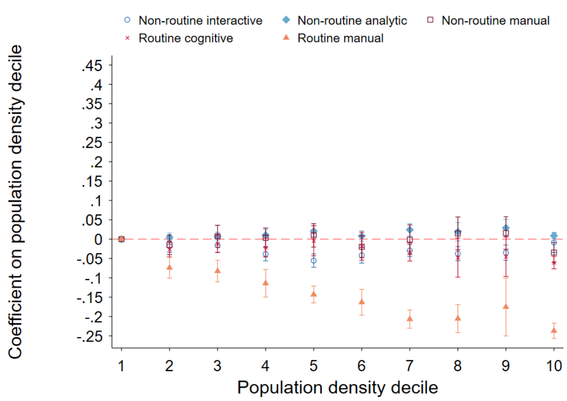
III. BA or above, without SOC f.e.



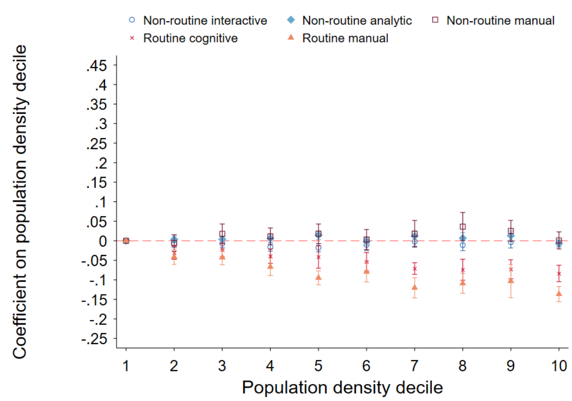
IV. BA or above, with SOC f.e.



V. HS only, without SOC f.e.

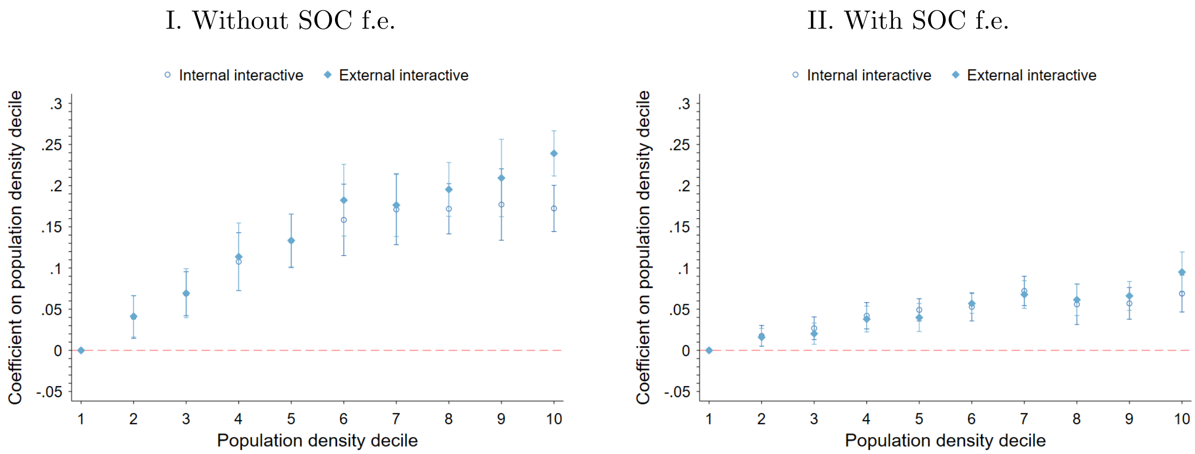


VI. HS only, with SOC f.e.



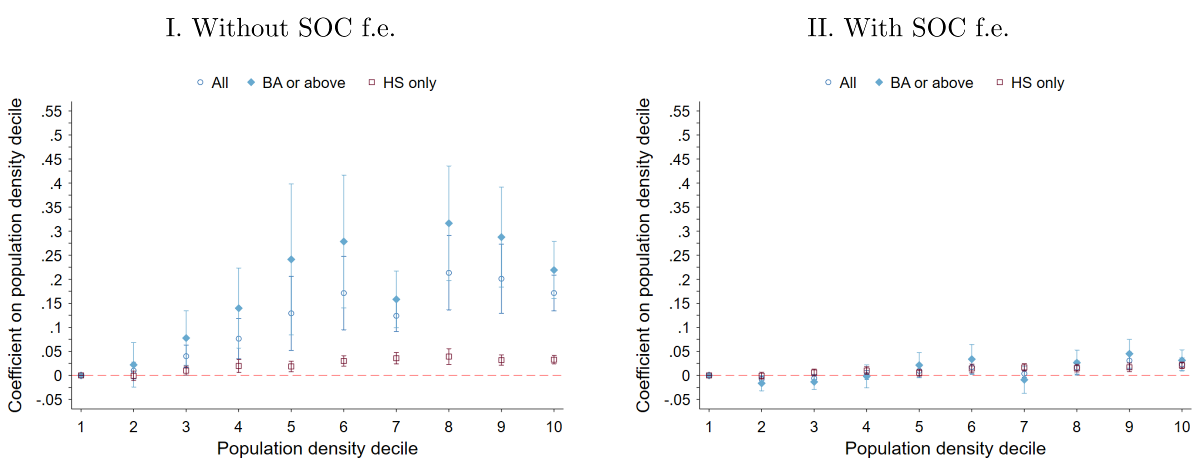
This figure reproduces Figure 1 but substitutes CZ population density deciles for CZ population deciles.

Figure C.5: O*NET Interactive Tasks and Market Density



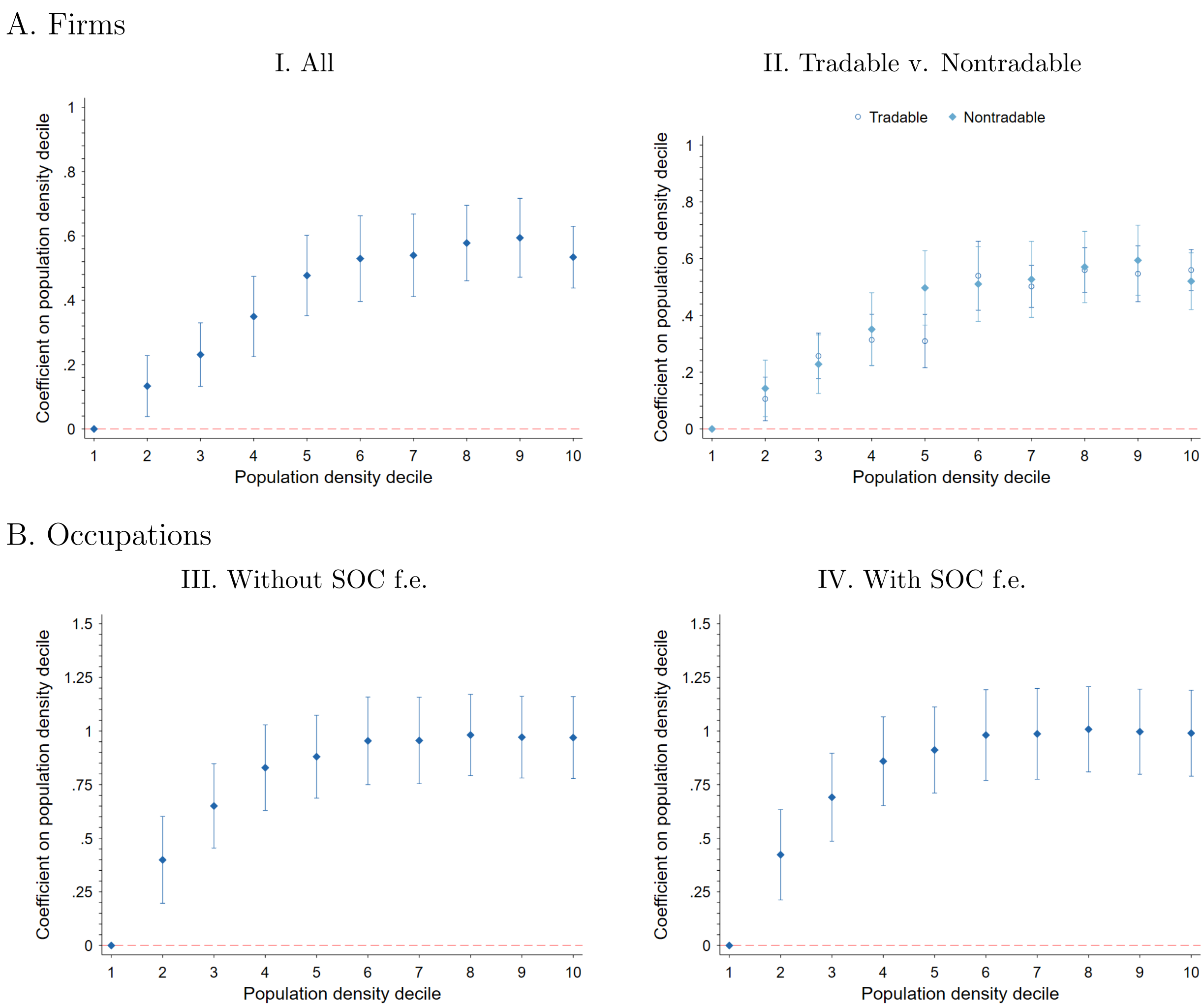
This figure reproduces Figure 2 but substitutes CZ population density deciles for CZ population deciles.

Figure C.6: The Technology Gradient with Market Density



This figure reproduces Figure 3 but substitutes CZ population density deciles for CZ population deciles.

Figure C.7: Specialization Gradient and Market Density



This figure reproduces Figure 4 but substitutes CZ population density deciles for CZ population deciles.

C.2 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 399 by excluding those that, according to our judgment, do not reflect job tasks.

In this section, we increase the number of tasks to 2,000—a higher resolution—and reproduce Figure 4, the main figure using these granular task measures to study the relationship between specialization and market size. Figure C.8 shows that the results are not sensitive to increasing the number of tasks to 2,000. Figure C.9 reproduces Figure 4 where the specialization measures are based on a task vector of 300—i.e., keeping the most common 300 of our main specification's 399 tasks. Figure C.9 shows that the results are not sensitive to reducing the number of tasks to 300.

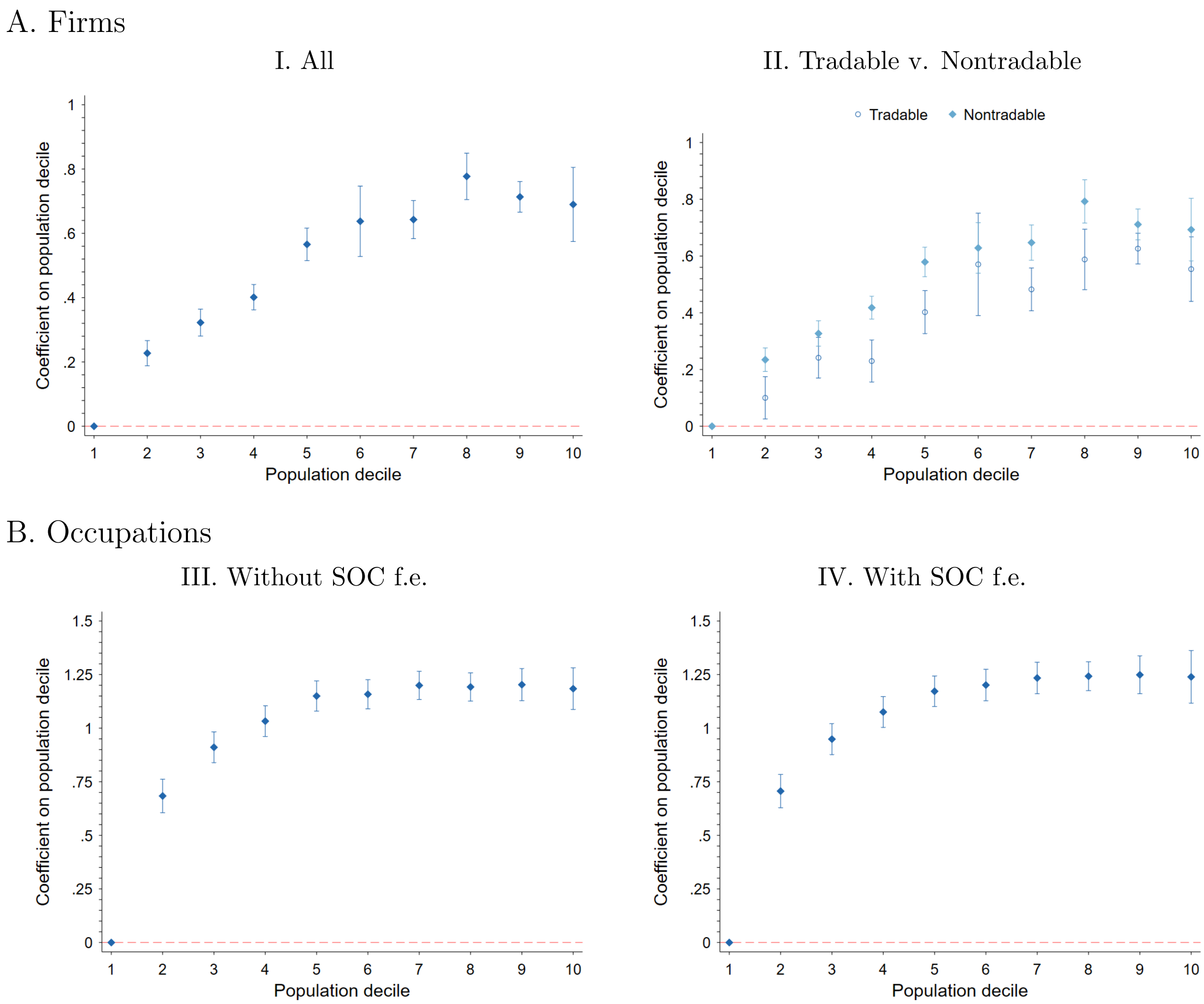
We lastly check the sensitivity of our specialization results to aggregating granular tasks that have similar meanings. The task extraction algorithm produces some distinct tasks that are similar, such as "provide feedback" and "provide recommendations." Rather than rely on our judgment to determine the similarity of different tasks, we use a natural language processing approach that uses word contexts (from a separate corpus) to aggregate similar tasks. To group our verb-noun tasks into a smaller number of clusters, we follow two steps. In the first step, we convert our list of tasks into a vector representation, using the Word2Vec implementation in the Gensim library for Python. Specifically, we use one of Gensim's pre-trained models, which was trained on the Google News data set. Using this model, we recover each task's underlying representation as a 100-element vector.[42] We next add the vector representation of the verb to that of the noun, so as to obtain a single vector for each task. In the second step, we cluster tasks in this vector space, using Stata's implementation of k-means clustering. The algorithm requires specifying the number of clusters ex-ante, and we experiment with 50, 75, and 100, all of which entail a substantial dimensionality reduction relative to our original list of tasks. We choose 75 task clusters for the exercise presented here, but we note that we examined the sensitivity to this choice to using 50 or 100 task clusters and it has little effect on the results.

Tables C.3-C.4 illustrate how the aggregation works. The tables show the original list of 399 tasks and the corresponding 75 task clusters. The first task cluster the algorithm creates is "identifies problems," "resolve issue," and "resolve problems." The second task cluster includes intuitively similar tasks such as "provide feedback," "provide recommendations," "provide guidance," and "provides leadership," but also includes tasks we may not a priori view as similar, such as "offer products" or "make offer." We believe the clustering method does a reasonably good job of aggregating similar tasks in a way that minimizes bias from excess reliance on researcher discretion. Using the 75 task clusters, we calculate task cluster dissimilarities in firm-market and occupation-market cells and reproduce the key

[42]This library can be found at https://radimrehurek.com/gensim/index.html. The library's documentation states that the Google News model was trained on "about 3 million words and phrases." We adopt Gensim's default for the size of vector representation.

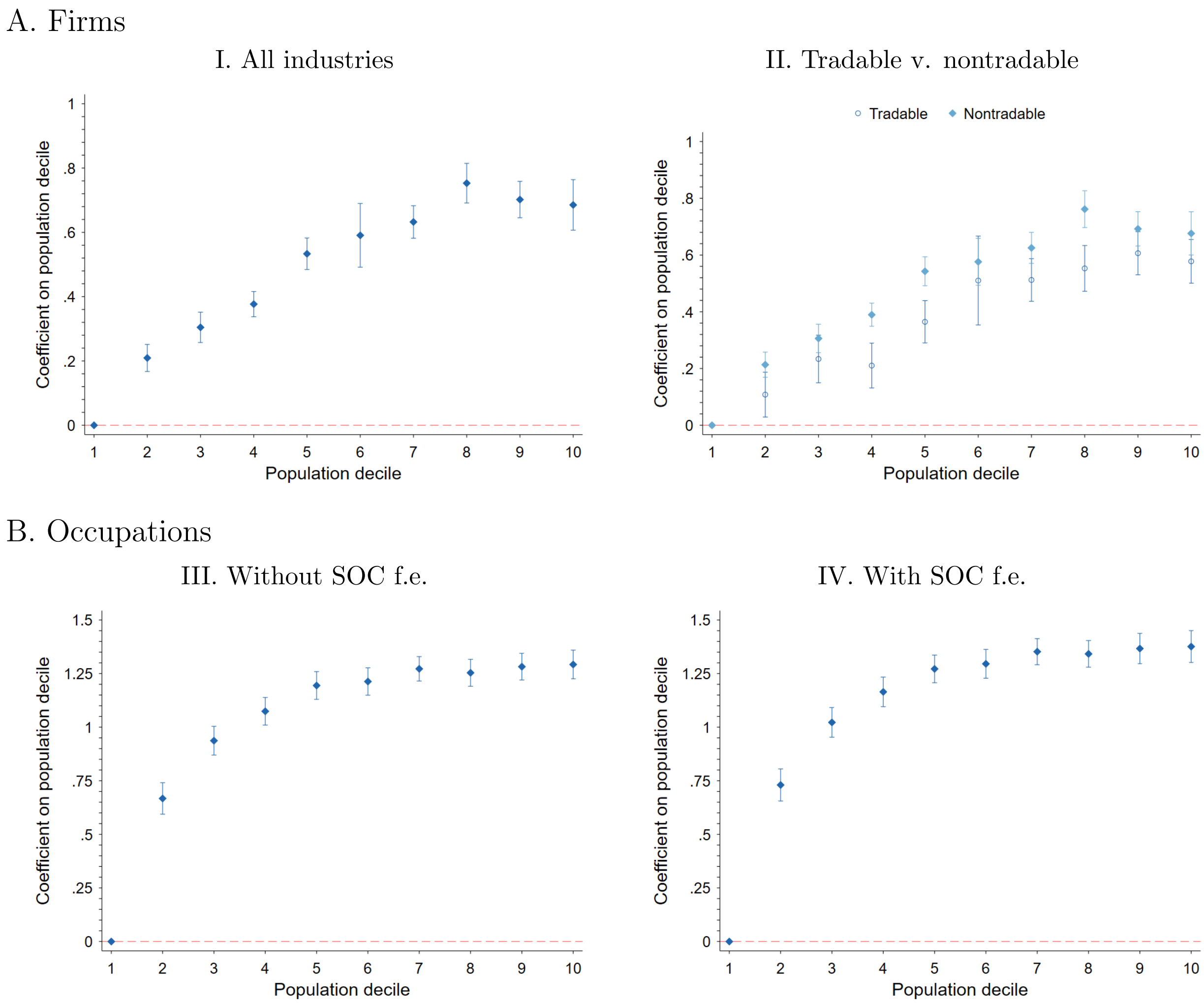
specialization gradients in the paper. Figure C.10 shows the results of this exercise. This figure shows a similar specialization gradient as Figure 4, reinforcing the finding of increased specialization in larger markets.

Figure C.8: Specialization Gradient: Task Dissimilarity Within Firms and Occupations (with 2,000 Tasks)



The figure above reproduces Figure 4, except the task dissimilarity measures in the occupation-CZ are constructed based on extracting 2,000 tasks, a higher resolution vector of verb-noun tasks per job ad. For reference, the 1st population decile mean for the top left panel is -0.51, and for the top right panel is -0.54 for the nontradable sample and -0.06 for the tradable sample. The 1st population decile mean for the bottom two panels is -1.00.

Figure C.9: Specialization Gradient: Task Dissimilarity Within Firms and Occupations (300 Tasks)



This figure reproduces Figure 4 using a task list of 300 verb-noun pairs. For reference, the 1st population decile mean for the top left panel is -0.51, and for the top right panel is -0.53 for the nontradable sample and -0.06 for the tradable sample. The 1st population decile mean for the bottom two panels is -1.06.

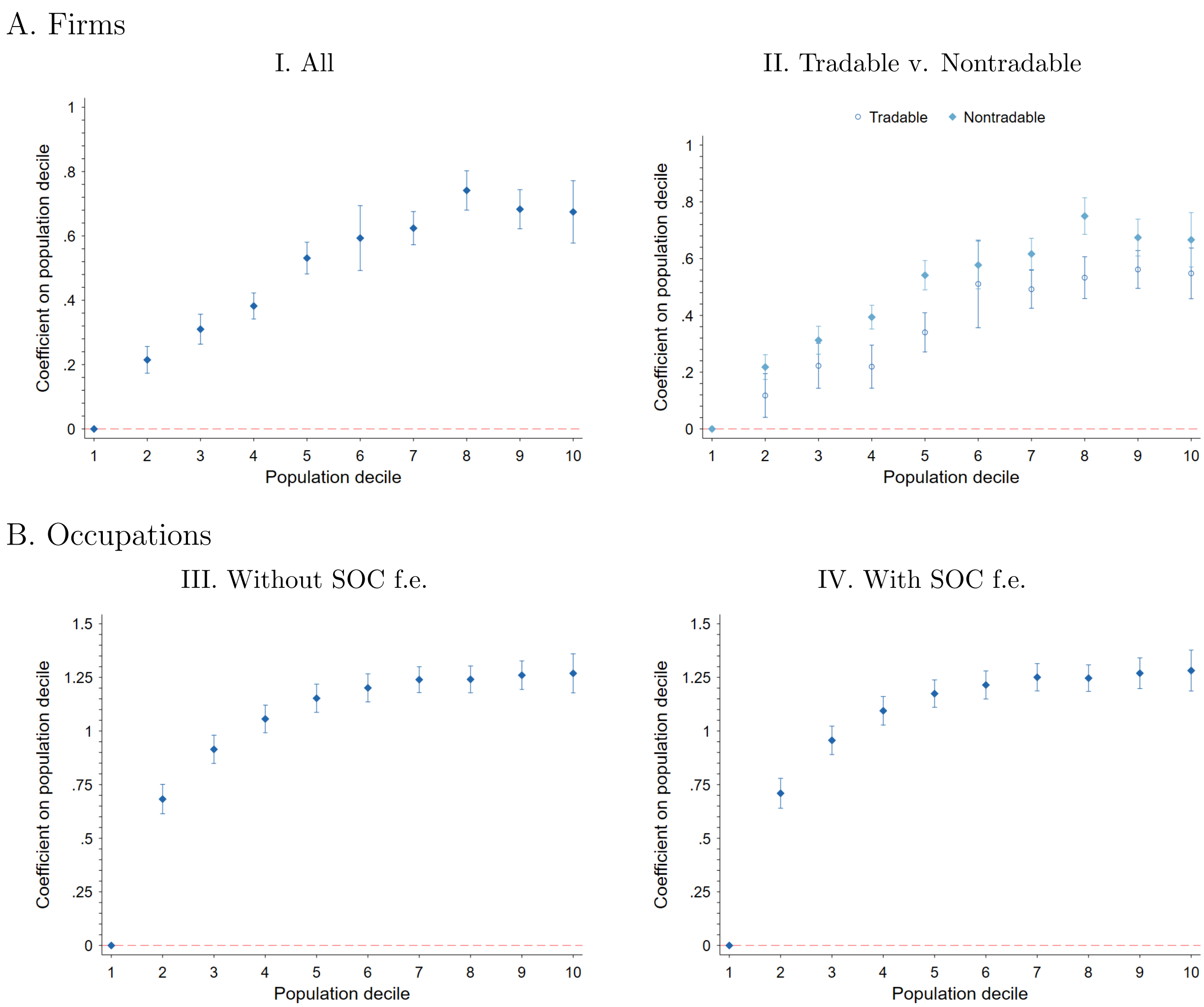
Table C.3: Task Clusters: Part I

identifies problems	1	assist store	8	sustained work	14	written skills	24	existing customers	29
resolve issue	1	bags counter tops	8	reaching pulling	15	assigned management	25	interact customers	29
resolve problems	1	dump baskets	8	assume responsibilities	16	assigned reading	25	leads customers	29
make offer	2	in store repairs	8	closing duties	16	assigned store	25	meet clients	29
offer products	2	lead store	8	opening duties	16	assigned supervisor	25	serving customers	29
provide client	2	maintain store	8	responsibilities duties	16	assigns directs	25	will customers	29
provide feedback	2	may store	8	analyzing data	17	executes store	26	answer questions	30
provide guidance	2	move store	8	conducting research	17	continuing education	27	asking questions	30
provide recommendations	2	needed in store	8	reset departments	18	generating business	27	according company	31
provide technical support	2	signing shelves	8	build relationships	19	growing business	27	following company	31
provided information	2	supervising store	8	maintaining relationships	19	has client	27	following pogs	31
provides input	2	traveling store	8	manage relationships	19	including business	27	following policies	31
provides leadership	2	working store	8	working relationships	19	manages business	27	following vendor	31
provides performance	2	improving quality	9	ensure delivery	20	managing operations	27	pay vendors	31
provides quality	2	maintaining business	9	ensure quality	20	managing projects	27	including nights	32
providing care	2	maintaining environments	9	ensures quality assurance	20	meet business	27	including performance	32
providing direction	2	maintaining inventory	9	ensuring communications	20	processing transactions	27	sweep room	32
providing expertise	2	maintaining program	9	ensuring food	20	securing company	27	ensuring merchandising	33
providing solutions	2	maintaining standards	9	assisting management	21	staffing needs	27	include merchandising	33
providing support	2	providing environment	9	assisting team	21	support business	27	merchandising directives	33
preparing foods	3	achieve goals	10	communicating field	21	supporting activities	27	merchandising product	33
crews customer service	4	resolve rejections	11	crew directing	21	ensure accuracy	28	execute completion	34
preventing terrorists	5	include client	12	developing team	21	ensure adherence	28	execute display	34
preventing trafficking	6	include design	12	ensuring team	21	ensure client	28	execute walk through	34
containing materials	7	include development	12	focus team work	21	ensure completion	28	executing set	34
discontinued items	7	include hand	12	leading team	21	ensure compliance	28	according cvs	35
measuring drugs	7	include knowledge	12	managing teams	21	ensure employees	28	appear floor	35
operate equipment	7	include order	12	providing coaching	21	ensure guests	28	are compliance	35
remove items	7	include program	12	signing crew	21	ensure operation	28	capping vials	35
retrieving information	7	include sales	12	supervisor team	21	ensure policies	28	check acceptance	35
stored areas	7	include service	12	team members	21	ensure product	28	comply cvs	35
taking orders	7	include shelves	12	will teams	21	ensure projects	28	comply state	35
taking vehicle	7	include staff	12	working team	21	ensure restaurant	28	delegated photo	35
unloading trucks	7	closes store	13	as needed assist	22	ensure safety	28	detail ability	35
using computer	7	opens store	13	assist development	22	ensure service	28	document counts	35
using enhancements	7	causing discomfort	14	assist staff	22	ensure stores	28	drive in employees	35
using equipment	7	causing drafts	14	provide assistance	22	assisting clients	29	driving culture	35
using knowledge	7	causing walking	14	request help	22	assisting customers	29	floors work	35
using orders	7	required driver	14	written oral communication	23	engage customers	29	follow instructions	35
adapting store	8	returned check	14	writing skills	24	existing clients	29	including work	35

Table C.4: Task Clusters: Part II

make recommendations	35	adjust facings	41	working variety	49	answering telephone	56	maintain pull	66
making decisions	35	guided values	41	meet deadlines	50	using phone	56	maintain records	66
may floor	35	locking setting	41	meet goals	50	trash rest	57	maintain work environment	66
pay policy	35	set displays	41	set goals	50	develop relationship	58	maintain working	66
perform job	35	build customer	42	labeling prescriptions	51	establish relationships	58	maintained times	66
perform maintenance	35	communicate customer	42	maintain pharmacy	51	stand periods	59	signing maintain	66
perform register	35	ensure customer	42	maintain prescription	51	sustained periods	59	document cash	67
pulls deposits	35	handling customer	42	organized pharmacy	51	ensure customer satisfaction	60	operating cash register	67
putting drug	35	helping customer	42	pharmacist communicate	51	increase customer satisfaction	60	procedures cash	67
react program	35	meets customer	42	requires merchandise	52	maximizing customer satisfaction	60	required paperwork	67
require walking	35	put customer	42	requires travel	52	needed customer satisfaction	60	exceed sales	68
return system	35	resolving customer	42	windows ceilings	53	checking employee	61	generate sales	68
scheduling activities	35	responding customer	42	windows removal	53	conducting employee	61	increase sales	68
scheduling appointments	35	seek customer	42	apprehend company	54	evaluates employees	61	intern communication	69
securing door	35	meet requirements	43	communicate information	54	carry pounds	62	written communication	69
skating carhop	35	meets standards	43	demonstrate knowledge	54	lift lbs	62	written instructions	69
stand walk	35	appropriate use	44	develop business	54	lifting pounds	62	make adjustments	70
taking actions	35	move trays	44	develop planning	54	weighing pounds	62	make changes	70
work others	35	passing emit	44	develop productivity	54	prioritize tasks	63	secure change	70
work projects	35	preferred ability	44	develop solutions	54	bagging merchandise	64	work flexible schedule	70
work schedule	35	seal trays	44	develop test	54	check in merchandise	64	work shift	70
work week ends	35	use hands	44	developing implement	54	damaged merchandise	64	sweeping stock	71
working departments	35	using eye	44	developing strategies	54	handle merchandise	64	work stock	71
works custom	35	vacuum face	44	establish policies	54	have merchandise	64	prepare returns	72
outdated merchandise	36	change bulbs	45	establish priorities	54	lifting merchandise	64	become slippery	73
customer service culture	37	assigned skills	46	established guidelines	54	may merchandise	64	may slippery	73
include customer service	37	on going training	47	identify opportunities	54	move merchandise	64	hiring training	74
provide customer	37	achieving sales	48	identify shoplifters	54	react shoplifters	64	including maintenance	74
provide customer service	37	assisted sales	48	identifying conditions	54	recalled merchandise	64	including management	74
provide service	37	dcr photofinishing	48	maximize sale	54	sorting merchandise	64	including preparation	74
receives service	37	developed sales	48	maximizes profitability	54	problem solving skills	65	including support	74
execute cash	38	driving sales	48	obtain information	54	maintain appearance	66	including systems	74
handling cash	38	managing sales	48	promote shopping	54	maintain area	66	including training	74
including cash	38	meet sales	48	protect company	54	maintain awareness	66	including translation	74
regarding cash register	38	needed inventory management	48	transforming delivery	54	maintain card	66	provides training	74
handle tasks	39	photofinishing orders	48	greeting card	55	maintain communication	66	requiring security	74
bulletins action	40	sells products	48	greeting customers	55	maintain custom	66	training sessions	74
following reports	40	working sales	48	greeting operations	55	maintain files	66	cvs workflow	75
prepare reports	40	perform variety	49	using greet	55	maintain knowledge	66		
writing reports	40	serving quality	49	answering phones	56	maintain productivity	66		

Figure C.10: Specialization Gradient: 75 Task Clusters



The figure above reproduces Figure 4, except the task dissimilarity measures in the firm-market and occupation-market cells are constructed using 75 task clusters. These task groupings are constructed from the original 399 tasks, which are reduced to 75 task clusters using a natural language processing approach described in the text. The 1st population decile mean for the top left panel is -0.50, and for the top right panel is -0.52 for the nontradable sample and -0.04 for the tradable sample. The 1st population decile mean for the bottom two panels is -1.02.

Measurement Error and Robustness to Controls

Small markets have fewer job ads per occupation-market (or firm-market) cell. Since the resulting within-cell sampling error may systematically vary with market size, one may worry that sampling error may lead us to spuriously conclude that job dissimilarity is increasing in market size. To assess the validity of this concern, we reproduce the key specialization figure in the analysis (Figure 4) with an additional control for the number of ads in the cell. Reassuringly, the estimates of this exercise, reported in Figure C.11 below, are virtually identical to Figure 4. We also check the sensitivity of the specialization gradients to measurement error by reproducing panel A of Figure C.12 for firm-market cells with strictly more than the median number of job ads, which is 5, and below or including the median number of job ads.

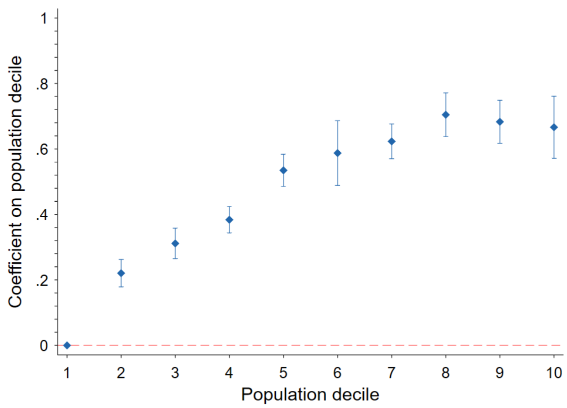
Placebo-Type Exercise: Analysis of National Chains

We also perform a placebo-type analysis of our specialization gradients for a subset of firms—national chains. To come up with a list of these national chains, we first identify the top 20 company names that have the most job postings. From this list, we identify the chains, which include: Advance Auto Parts, CVS Caremark, Dollar General, Family Dollar, Harbor Freight, Home Depot, Lowes, Macys, McDonalds, Pizza Hut, Sears, Taco Bell, and Wells Fargo. We reproduce panel A of Figure 4 for these retailers (which are all in non-tradable industries) and compare it to all other non-tradable sector firms, which is presented in Figure C.13. The results show a flattened specialization gradient, as we might expect, given the relative homogeneous types of organizational structure of these chains across space. Note that we would not expect the specialization gradient to flatten entirely, since even the workforce of national chains may become more specialized in large markets. Nevertheless, it is reassuring to see a flattened gradient for these national chains.

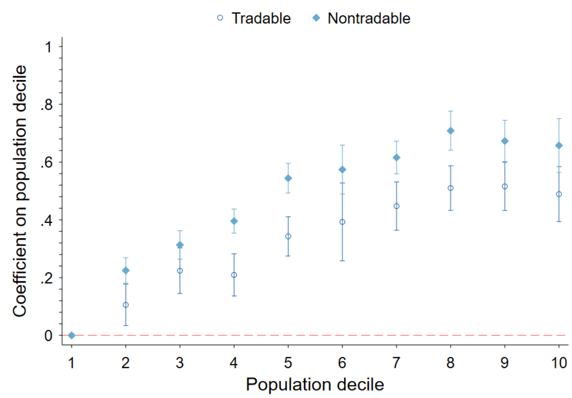
Figure C.11: Specialization Gradient: Task Dissimilarity Within Firms and Occupations

A. Firms

I. All

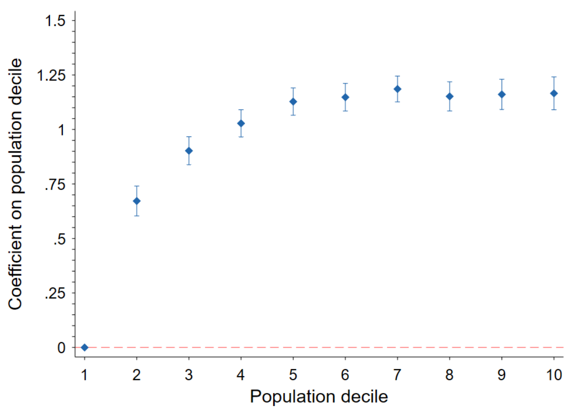


II. Tradable v. Nontradable

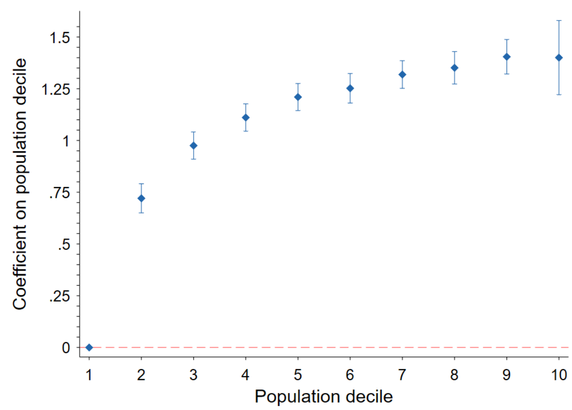


B. Occupations

III. Without SOC f.e.



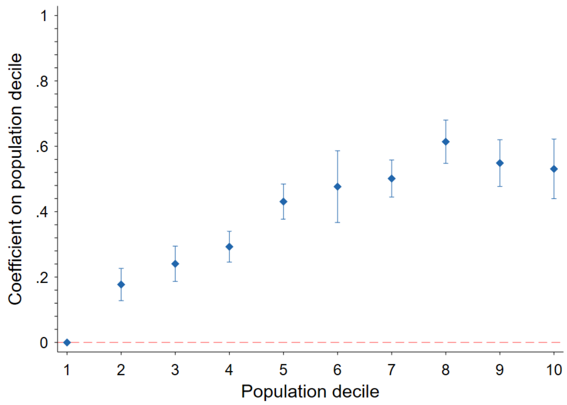
IV. With SOC f.e.

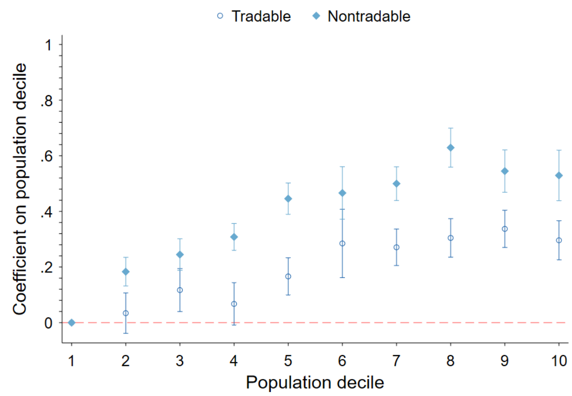


The figures above reproduce Figure 4 with an additional control for the number of ads in the cell.

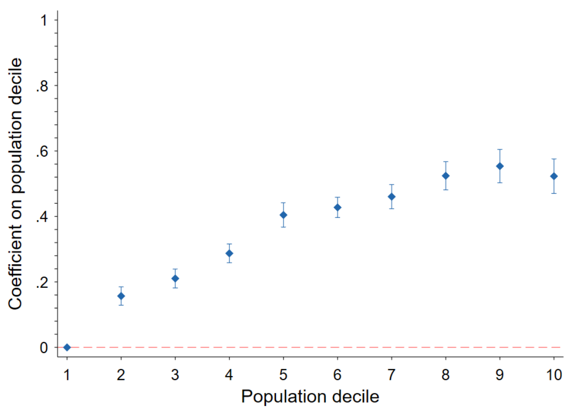
Figure C.12: Specialization Gradient: Above v. Below Median Firm Size

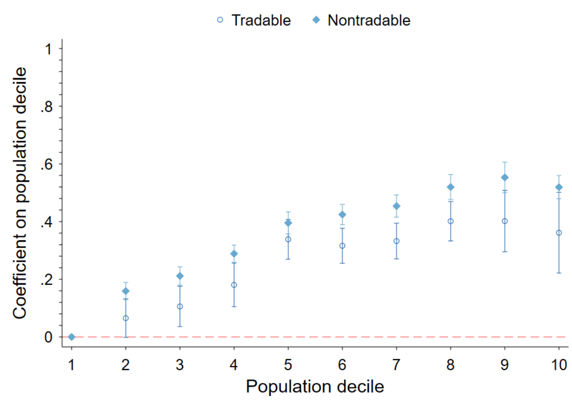
I. Above Median





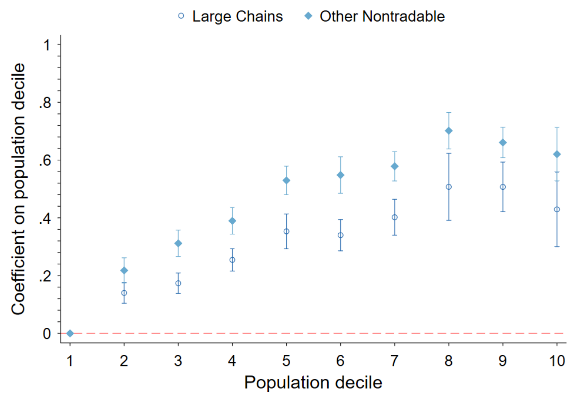
II. Below Median





The figures above reproduce panel A of Figure 4 separately for firm-markets with strictly more than the median number of postings (5) and below or including the median number of postings.

Figure C.13: Specialization Gradient: Large Chains v. Other Non-Tradable



The figures above reproduce panel A.I of Figure 4 separately for large chains (which are all in non-tradable industries) and other non-tradable firms. We control for both log ad words in the cell and two-digit NAICS fixed effects.

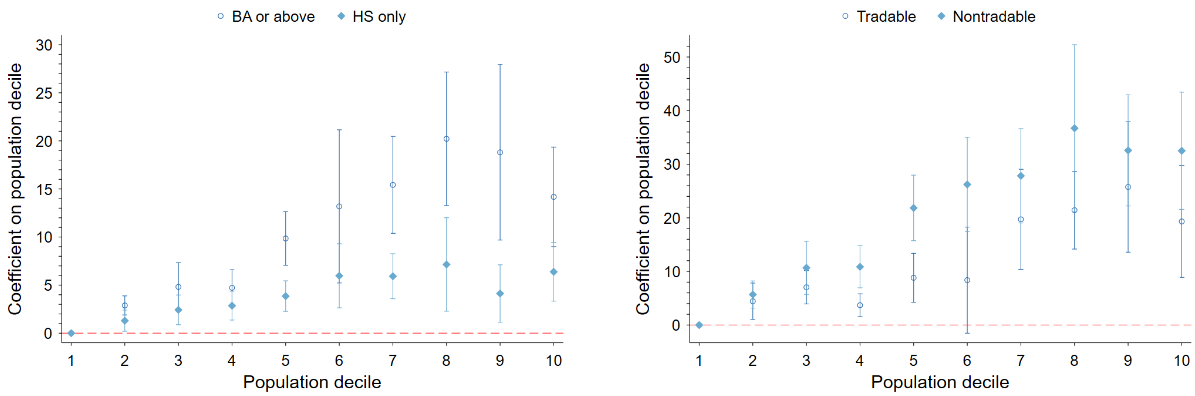
Number of Job Titles

Prior research—notably, Tian (2019)—examines 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.14, 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 tradable and nontradable 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 nontradable sector firms.

Figure C.14: Specialization Gradient: Number of Job Titles

I. BA or above v. HS only

II. Tradable v. Nontradable



The unit of observation is the firm-market (CZ). We regress the number of distinct job titles on market size deciles, controlling for the 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 diploma, 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 tradable and nontradable sector firms. All regressions are weighted by the number of ads in the firm-market. Standard errors are robust and clustered at the CZ level. The figure plots the coefficients on the CZ size deciles. For reference, in the left panel, the 1st population decile mean for BA or above is 2.58 and for HS only is 3.11. In the right panel, the 1st population decile mean for tradable is 9.96 and for nontradable is 10.68.

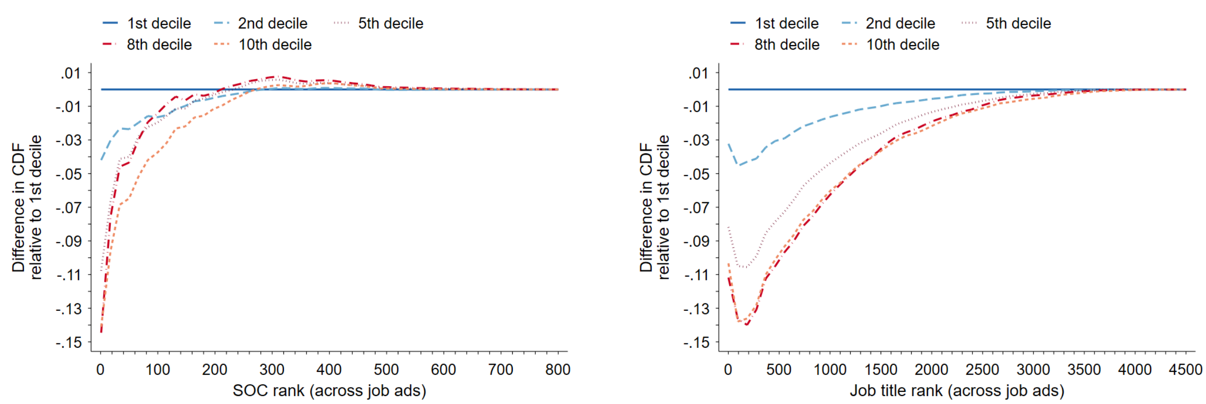
The Distribution of Common and Rare Occupations

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 1st population decile. The left panel of Figure C.15 shows that the most common occupations are overrepresented in small markets, while more rare occupations are overrepresented in large markets. For example, of the 10 most common occupations economy-wide, the 10th decile market has 11-13 percentage points lower share of these occupations compared with the 1st population decile. For the 300-400 most common occupations, the 10th decile market has about a 0.3 percentage point greater share relative to the 1st population decile.

This finding—that rare jobs represent a larger share of total jobs in larger markets—is even more pronounced when we perform the analysis at the job title level. Note that the job title is not observed in standard datasets such as the ACS or the 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 overrepresented in smaller markets (as a share of total jobs).

Figure C.15: 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 1st 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.

C.3 Robustness of Wage Regressions

This section evaluates the sensitivity of the wage regressions presented in Table 5. We then examine the sensitivity of the wage regression table to two-way clustering of the standard errors, by CZ and four-digit SOC. We then examine the sensitivity of Table 5 to using an alternative task dissimilarity measure—one in which dissimilarity is based on a task vector with 2,000 tasks, a higher resolution of tasks per job ad. Lastly, we examine the sensitivity to including CZ fixed effects.

Table C.5 reproduces Table 5, the main wage regression table, except uses two-way clustering of the standard errors, by CZ and four-digit SOC. The estimated coefficient on task dissimilarity in column 1—the full sample excluding SOC fixed effects—and columns 6 and 7—the sample of blue-collar occupations—lose statistical significance but the overall takeaways of Table 5 are unchanged.

Table C.6 reproduces Table 5 except the task dissimilarity measures in the occupation-CZ are constructed based on a task vector with 2,000 tasks, a higher resolution of tasks per job ad. The results are nearly identical to those in Table 5. Note that the number of observations is slightly higher compared to Table 5. One difference, since longer task vectors are more likely to have a non-zero element, is that there are slightly more occupation-CZ cells with more than 2 job ads that have non-zero task vectors, which is required for the task dissimilarity to be defined and for the occupation-CZ cell to enter the regression. Table C.7 reproduces Table 5 with task dissimilarity measures in the occupation-CZ based on 300 tasks, a lower resolution, and shows similar results.

Table C.8 reproduces Table 5 with CZ fixed effects. The goal is to understand whether specialization and technologies have an effect on wages after controlling for CZ size and other unobserved features of the labor market. Table C.8 shows that with CZ f.e., the coefficient on specialization diminishes. This result is precisely what Smith's theory would predict: It is *through* market size that specialization affects productivity; after controlling for CZ size, the link between specialization and productivity is muted. Nevertheless, the specialization coefficient remains significant with CZ and SOC fixed effects for white-collar occupations in column 5. The interactive tasks coefficient is also diminished once we control for CZ f.e., which is consistent with market size enhancing the relationship between worker interactions and productivity. The technologies coefficient remains statistically significant even with CZ f.e., for the full sample and for white collar occupations.

Table C.5: Task Dissimilarity, Technologies, Interactive Tasks, and Wages (with Two-Way Clustering)

	All			White-collar		Blue-collar	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interactive tasks	0.123***	0.034**	0.021**	0.053**	0.028	0.027**	0.024**
	(0.043)	(0.014)	(0.010)	(0.023)	(0.017)	(0.010)	(0.010)
Technology requirements	0.385***	0.340***	0.212***	0.367***	0.214***	0.001	-0.004
	(0.074)	(0.073)	(0.044)	(0.076)	(0.044)	(0.042)	(0.040)
Task dissimilarity	0.033	0.036***	0.029***	0.062***	0.047***	0.007	0.006
	(0.022)	(0.006)	(0.005)	(0.011)	(0.006)	(0.005)	(0.004)
BA or above			0.859***		0.933***		0.474***
			(0.090)		(0.101)		(0.079)
SOC f.e.	No	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	44,956	44,956	44,956	24,370	24,370	11,247	11,247
R^2	0.240	0.811	0.840	0.768	0.819	0.568	0.580
Mean of dependent var.	10.784	10.784	10.784	10.983	10.983	10.576	10.576
Mean task dissimilarity	0.000	0.000	0.000	0.152	0.152	-0.178	-0.178
Mean technology requirements	0.157	0.157	0.157	0.223	0.223	0.043	0.043
Mean interactive tasks	0.000	0.000	0.000	0.435	0.435	-0.915	-0.915
Mean BA or above	0.363	0.363	0.363	0.517	0.517	0.076	0.076

This table reproduces Table 5, except uses two-way clustering by CZ and four-digit SOC.

Table C.6: Task Dissimilarity, Technologies, Interactive Tasks, and Wages (with 2,000 Tasks)

	All			White-collar		Blue-collar	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interactive tasks	0.118***	0.036***	0.022***	0.053***	0.028***	0.028***	0.025***
	(0.007)	(0.007)	(0.005)	(0.010)	(0.007)	(0.007)	(0.006)
Technology requirements	0.374***	0.337***	0.210***	0.361***	0.211***	0.005	-0.000
	(0.013)	(0.042)	(0.026)	(0.046)	(0.029)	(0.029)	(0.028)
Task dissimilarity	0.054***	0.035***	0.028***	0.066***	0.049***	0.005*	0.004
	(0.003)	(0.003)	(0.002)	(0.006)	(0.004)	(0.003)	(0.003)
BA or above			0.856***		0.926***		0.478***
			(0.069)		(0.076)		(0.059)
SOC f.e.	No	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	45,602	45,602	45,602	24,681	24,681	11,476	11,476
R^2	0.246	0.810	0.839	0.768	0.819	0.567	0.579
Mean of dependent var.	10.783	10.783	10.783	10.983	10.983	10.576	10.576
Mean task dissimilarity	0.000	0.000	0.000	0.178	0.178	-0.232	-0.232
Mean technology requirements	0.156	0.156	0.156	0.223	0.223	0.043	0.043
Mean interactive tasks	0.000	0.000	0.000	0.434	0.434	-0.912	-0.912
Mean BA or above	0.363	0.363	0.363	0.518	0.518	0.076	0.076

This table reproduces Table 5, except the task dissimilarity measures in the occupation-CZ are constructed based on extracting 2,000 tasks, a higher resolution vector of verb-noun tasks per job ad.

Table C.7: Task Dissimilarity, Technologies, Interactive Tasks, and Wages (with 300 Tasks)

	All			White-collar		Blue-collar	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interactive tasks	0.124***	0.035***	0.021***	0.054***	0.028***	0.027***	0.024***
	(0.007)	(0.007)	(0.005)	(0.011)	(0.007)	(0.007)	(0.007)
Technology requirements	0.384***	0.342***	0.214***	0.367***	0.214***	0.002	-0.004
	(0.013)	(0.042)	(0.026)	(0.047)	(0.029)	(0.029)	(0.028)
Task dissimilarity	0.041***	0.037***	0.029***	0.062***	0.048***	0.005	0.004
	(0.004)	(0.004)	(0.003)	(0.006)	(0.004)	(0.003)	(0.003)
BA or above			0.857***		0.933***		0.474***
			(0.069)		(0.076)		(0.059)
SOC f.e.	No	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	44,393	44,393	44,393	24,281	24,281	11,100	11,100
R^2	0.242	0.812	0.841	0.768	0.819	0.569	0.581
Mean of dependent var.	10.784	10.784	10.784	10.983	10.983	10.577	10.577
Mean task dissimilarity	0.000	0.000	0.000	0.142	0.142	-0.213	-0.213
Mean technology requirements	0.157	0.157	0.157	0.223	0.223	0.043	0.043
Mean interactive tasks	0.000	0.000	0.000	0.434	0.434	-0.919	-0.919
Mean BA or above	0.363	0.363	0.363	0.517	0.517	0.076	0.076

This table reproduces Table 5, except the task dissimilarity measures in the occupation-CZ are constructed based on extracting 300 tasks.

Table C.8: Task Dissimilarity, Technologies, Interactive Tasks, and Wages: Adding CZ Fixed Effects

	All			White-collar		Blue-collar	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interactive tasks	0.122***	0.006	0.004	0.002	-0.000	0.004	0.004
	(0.007)	(0.004)	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)
Technology requirements	0.338***	0.145***	0.108***	0.099***	0.072***	-0.015	-0.018
	(0.011)	(0.020)	(0.015)	(0.018)	(0.015)	(0.024)	(0.024)
Task dissimilarity	0.008***	-0.002	0.001	0.002	0.006***	-0.001	0.000
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
BA or above			0.509***		0.483***		0.311***
			(0.027)		(0.024)		(0.040)
SOC f.e.	No	Yes	Yes	Yes	Yes	Yes	Yes
CZ f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	44,956	44,956	44,956	24,370	24,370	11,247	11,247
R^2	0.308	0.871	0.879	0.871	0.880	0.696	0.700
Mean of dependent var.	10.784	10.784	10.784	10.983	10.983	10.576	10.576
Mean task dissimilarity	0.000	0.000	0.000	0.152	0.152	-0.178	-0.178
Mean technology requirements	0.157	0.157	0.157	0.223	0.223	0.043	0.043
Mean interactive tasks	0.000	0.000	0.000	0.435	0.435	-0.915	-0.915
Mean BA or above	0.363	0.363	0.363	0.517	0.517	0.076	0.076

This table reproduces Table 5 with CZ fixed effects.

C.4 Robustness to Data Source

In this appendix, we reproduce some of our main empirical exercises using a sample of ads from Burning Glass. The EMSI dataset has its own advantages for our purpose. In particular, it contains the ads' raw text, allowing us to isolate the tasks that employers list. In contrast, Burning Glass commingles 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 can replicate Figure 2, we compute measures of internal-to-the-firm interactive tasks[43] and external-to-the-firm interactive tasks.[44] As in Section 3.1,

[43] 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."

[44] 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,"

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 can 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, we 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-) market as $d_{jcm} = 1 - V_{jcm} \cdot \overline{V}_{(-j)cm}$.

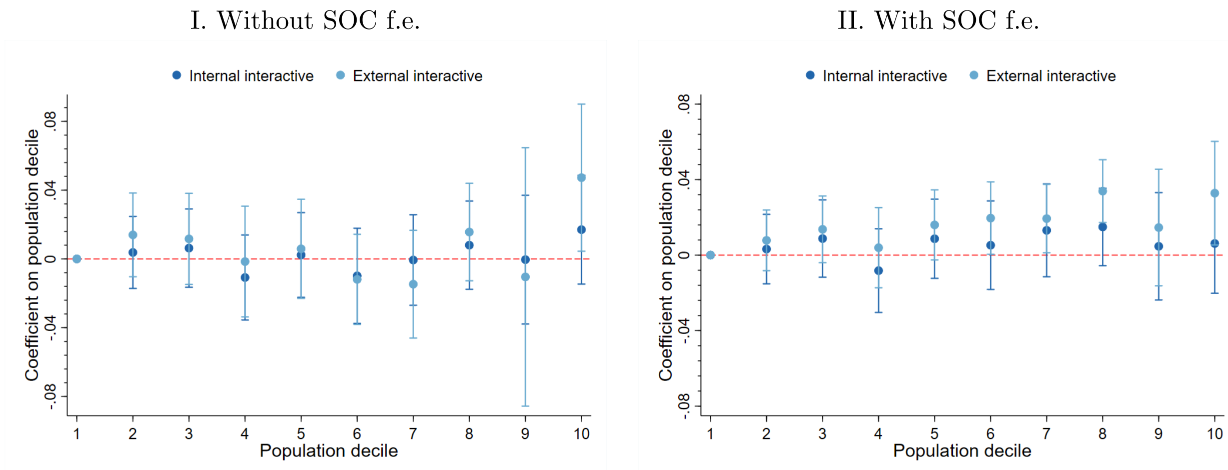
First, Figure C.16 replicates Figure 2. As in Section 3.1, external tasks increase in CZ size, both within and between six-digit SOCs. However, potentially due to the smaller sample size, the relationship between CZ size and internal tasks is no longer statistically significant.

Second, we reproduce Figure 4. As in Figure 4, Figure C.17 indicates that within-occupation and within-firm specialization is greater in more populous commuting zones, with a steeper gradient for firms in nontradable industries than for firms in tradable industries (panel II).

Finally, we reproduce Table 5. As in Table 5, Table C.9 indicates that wages are higher in markets with greater specialization, with greater technology usage, with greater interactive task intensity, and with a greater share of workers with a college degree. Furthermore, also as in Table 5, the relationships between wages and within-occupation × market specialization, technology intensity, and interactive task intensity are each stronger in white-collar than in blue-collar occupations.

"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.16: O*NET Interactive Tasks Gradient

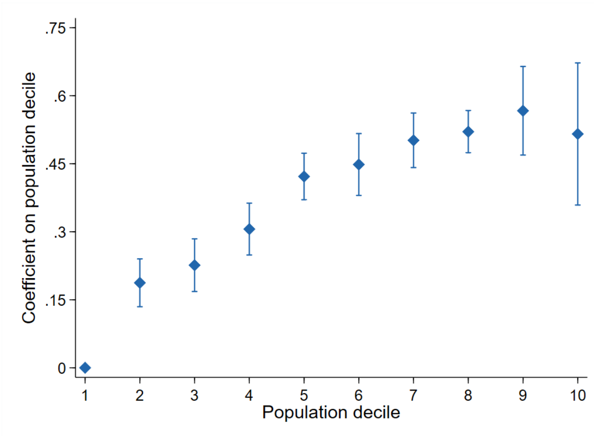


See the caption for Figure 2. In contrast, our task measures here come from Burning Glass data.

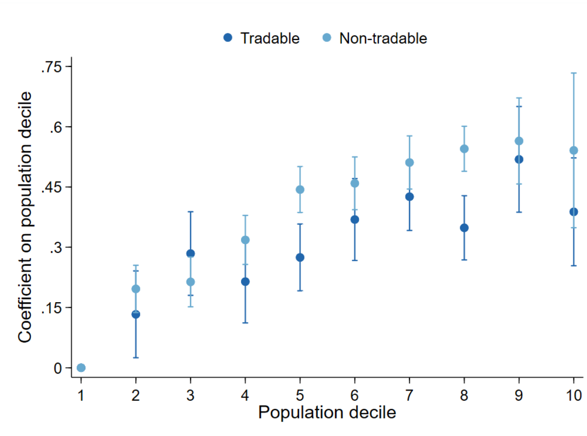
Figure C.17: Specialization Gradient: Task Dissimilarity Within Firms and Occupations

A. Firms

I. All

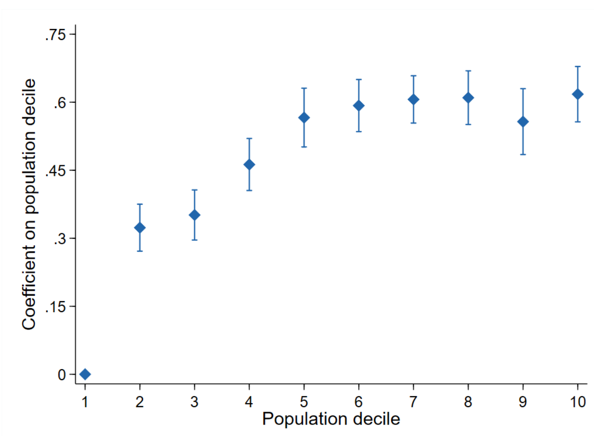


II. Tradable v. Nontradable

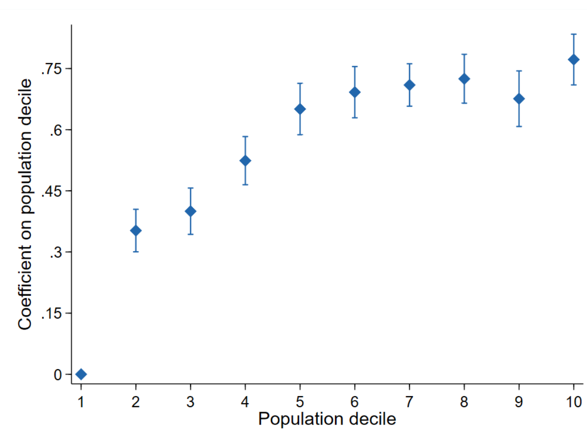


B. Occupations

III. Without SOC f.e.



IV. With SOC f.e.



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

Table C.9: Task Dissimilarity, Technologies, and Wages

	All			White collar		Blue collar	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Task dissimilarity	0.069***	0.022***	0.017***	0.053***	0.044***	0.010***	0.010***
	(0.005)	(0.003)	(0.002)	(0.011)	(0.008)	(0.003)	(0.003)
Technology requirements	0.285***	0.174***	0.114***	0.224***	0.142***	0.010	0.004
	(0.007)	(0.021)	(0.014)	(0.038)	(0.016)	(0.015)	(0.015)
Interactive Tasks	0.060***	0.007*	0.004	0.010**	0.005	0.001	0.001
	(0.006)	(0.004)	(0.004)	(0.003)	(0.004)	(0.006)	(0.005)
Education			0.518***		0.647***		0.087*
			(0.076)		(0.135)		(0.040)
SOC f.e.	No	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	32,623	32,623	32,623	20,194	20,194	7,099	7,099
R^2	0.200	0.823	0.833	0.774	0.795	0.577	0.578
Mean of dependent var.	10.783	10.783	10.783	10.971	10.971	10.567	10.567
Mean task dissimilarity	-0.000	-0.000	-0.000	0.078	0.078	-0.083	-0.083
Mean technology requirements	0.573	0.573	0.573	0.771	0.771	0.244	0.244
Mean interactive tasks	0.000	0.000	0.000	0.309	0.309	-0.691	-0.691
Mean BA or above	0.384	0.384	0.384	0.553	0.553	0.069	0.069

See the caption for Table 5. In contrast, the task dissimilarity and technology measures here come from Burning Glass data.

C.5 Robustness to Using Posted Wages

To close, we examine the sensitivity of our results to the use of wage data from the American Community Survey. Unfortunately, we are unable to extract information on ads' wages from our EMSI job ad data. Given this, we rely on data from Burning Glass for this sensitivity analysis.

To begin, Tables C.10 and C.11 summarize the share of Burning Glass ads with a posted salary as well as the average (annual) posted salary in the American Community Survey. We find that the share of ads with a posted salary varies systematically by occupation — with 45 percent of Farming, Fishing, and Forestry jobs with a posted salary, compared to 8 percent in Food Preparation and Service — and by CZ population decile — with a larger share of ads with a posted salary in lower population CZs. Second, average wages in the two datasets are highly correlated with one another: Across the 23 two-digit SOCs presented in Table C.10, the raw (unweighted) correlation between log wages in the ACS and in Burning Glass equals 0.95. Across the 10 CZ population deciles, the analogous correlation equals

0.97.

Table C.10: Posted Wages by Occupation

Occupation	Share with Posted Salary	Log Salary	
		ACS	Burning Glass
Management (11)	0.13	11.29	11.14
Business and Financial Operations (13)	0.16	11.18	10.92
Computer and Mathematical (15)	0.12	11.29	11.22
Architecture and Engineering (17)	0.14	11.28	11.10
Life, Physical, and Social Science (19)	0.19	11.14	10.92
Community and Social Science (21)	0.20	10.71	10.71
Legal (23)	0.12	11.49	11.13
Educational Instruction and Library (25)	0.18	10.80	10.65
Arts, Design, and Entertainment (27)	0.13	10.87	10.72
Healthcare Practitioners and Technical (29)	0.10	11.16	11.01
Healthcare Support (31)	0.09	10.27	10.32
Protective Service (33)	0.35	10.85	10.58
Food Preparation and Serving (35)	0.08	10.07	10.24
Building, Grounds Cleaning and Maintenance (37)	0.17	10.15	10.26
Personal Care and Service (39)	0.17	9.96	10.34
Sales and Related (41)	0.07	10.84	10.70
Office and Administrative Support (43)	0.19	10.55	10.44
Farming, Fishing, and Forestry (45)	0.45	10.07	10.28
Construction and Extraction (47)	0.22	10.55	10.65
Installation, Maintenance, and Repair (49)	0.15	10.70	10.58
Production (51)	0.24	10.56	10.42
Transportation and Material Moving (53)	0.19	10.51	10.41
Military (55)	0.39	10.80	10.86

For each two-digit occupation, this table lists the share of Burning Glass ads with a posted salary, average annual (log) wages in the American Community Survey, and (for the subset of ads with a posted salary) average annual wages in the Burning Glass data.

Table C.11: Posted Wages by CZ Population Decile

CZ Population Decile	Share with Posted Salary	Log Salary	
		ACS	Burning Glass
1	0.18	10.55	10.49
2	0.18	10.62	10.53
3	0.17	10.65	10.56
4	0.18	10.69	10.61
5	0.16	10.77	10.64
6	0.16	10.79	10.70
7	0.15	10.85	10.72
8	0.14	10.95	10.81
9	0.12	10.95	10.83
10	0.15	10.86	10.83

For each CZ population decile, this table lists the share of Burning Glass ads with a posted salary, average annual (log) wages in the American Community Survey, and (for the subset of ads with a posted salary) average annual wages in the Burning Glass data.

Finally, Table C.12 reproduces Table 5, replacing ACS with Burning Glass as the source of (log) wages by occupation and commuting zone. (Since only a fraction—less than one-fifth—of ads have a posted wage, both the number of observations and the number of underlying ads represented in this regression table are smaller than in Table 5.) As in Table 5, we find that wages are increasing in task dissimilarity and technology usage, with greater slopes in white-collar occupations. Also as in Table 5, wages are higher in occupation-CZ cells with greater mentions of interactive tasks, though now the coefficients are no longer statistically significant in specifications with occupation fixed effects.

Table C.12: Task Dissimilarity, Technologies, Interactive Tasks, and Wages

	All			White collar		Blue collar	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Task dissimilarity	0.089***	0.025***	0.017***	0.045**	0.030*	0.022***	0.020***
	(0.006)	(0.005)	(0.004)	(0.015)	(0.015)	(0.006)	(0.006)
Technology requirements	0.232***	0.219***	0.126***	0.234***	0.123***	0.162***	0.112***
	(0.006)	(0.025)	(0.016)	(0.040)	(0.023)	(0.038)	(0.030)
Interactive Tasks	0.045***	0.022***	0.016***	0.025***	0.016**	0.029***	0.024***
	(0.004)	(0.006)	(0.006)	(0.007)	(0.006)	(0.005)	(0.005)
Education			0.729***		0.798***		0.627***
			(0.059)		(0.087)		(0.086)
SOC f.e.	No	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	21,851	21,851	21,851	14,115	14,115	4,456	4,456
R^2	0.195	0.670	0.692	0.667	0.699	0.406	0.421
Mean of dependent var.	10.694	10.694	10.694	10.822	10.822	10.498	10.498
Mean task dissimilarity	0.064	0.064	0.064	0.115	0.115	0.033	0.033
Mean technology requirements	0.597	0.597	0.597	0.787	0.787	0.246	0.246
Mean interactive tasks	0.003	0.003	0.003	0.305	0.305	-0.730	-0.730
Mean BA or above	0.397	0.397	0.397	0.557	0.557	0.065	0.065

See the caption for Table 5. In contrast, the cell-level average wages are computed as the average posted salary within Burning Glass among the subset of ads for which this salary exists.