sociological science

Supplement to:

Labussière, Marie, Thijs Bol. 2026. “Are Occupations “Bundles of Skills”? Identifying Latent Skill Profiles in the Labor Market Using Topic Modeling” Sociological Science 13: 362-407.

Online supplement

Are occupations "bundles of skills"? Identifying latent skill profiles in the labor market using topic modeling

Marie Labussière[1] and Thijs Bol[2]

[1]Sciences Po, marie.labussiere@sciencespo.fr
[2]University of Amsterdam

Contents

1 Descriptive statistics

1.1 Comparison with employment figures from the Annual Population Survey

Figure S1: Comparison between the distribution of occupations in the 2019 Starlight job postings data and employment figures from the 2019 Annual Population Survey (APS), at the major (1-digit) group level. APS data was retrieved from the UK Office for National Statistics (ONS) via Nomis.

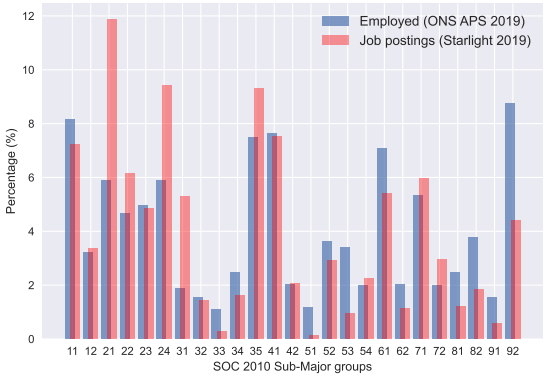

Figure S2: Comparison between the distribution of occupations in the 2019 Starlight job postings data and employment figures from the 2019 Annual Population Survey (APS), at the sub-major (2-digit) group level. APS data was retrieved from the UK Office for National Statistics (ONS) via Nomis.

1.2 Number of skill requirements per job postings

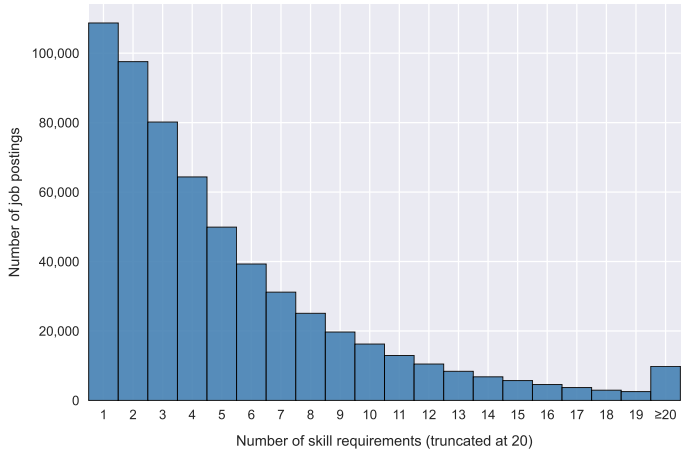

Figure S3: Histogram of the distribution of the number of skill requirements per job postings in the stratified random sample.

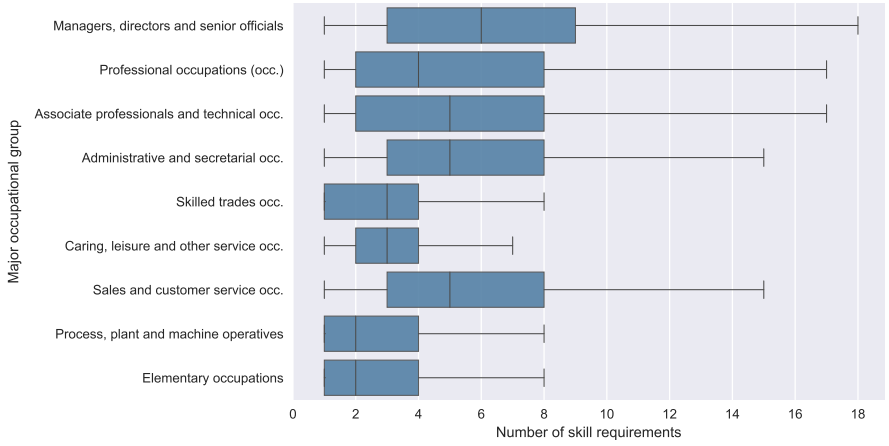

Figure S4: Boxplots of the number of skill requirements identified in job postings for the nine major occupational groups (SOC 2010).

1.3 Comparison of samples with and without wage record

	Sample without wage record		Sample with wage record	
	Mean	SD	Mean	SD
LDA topic load				
Skill set 1	0.04	0.12	0.04	0.14
Skill set 2	0.05	0.16	0.07	0.20
Skill set 3	0.09	0.15	0.10	0.16
Skill set 4	0.04	0.09	0.02	0.07
Skill set 5	0.03	0.09	0.03	0.10
Skill set 6	0.10	0.12	0.08	0.11
Skill set 7	0.05	0.16	0.05	0.15
Skill set 8	0.02	0.07	0.01	0.07
Skill set 9	0.11	0.14	0.10	0.14
Skill set 10	0.04	0.11	0.04	0.11
Skill set 11	0.11	0.21	0.10	0.20
Skill set 12	0.03	0.08	0.03	0.09
Skill set 13	0.11	0.23	0.13	0.27
Skill set 14	0.02	0.08	0.01	0.06
Skill set 15	0.09	0.17	0.08	0.17
Skill set 16	0.02	0.08	0.01	0.08
Skill set 17	0.02	0.07	0.01	0.06
Skill set 18	0.03	0.12	0.02	0.11
Skill set 19	0.02	0.11	0.04	0.15
Minimum experience required (in years)	3.39	2.85	3.24	2.97
Minimum degree level required (in years)	13.96	2.53	13.82	2.48

Table S1: Comparison between job postings with wage record and without wage record, numerical variables.

	Sample without wage record (%)	**Sample with wage record** (%)
Occupational major group (SOC 2010)		
Managers, directors and senior officials	11.97	10.99
Professional occupations	22.16	20.93
Associate professionals and technical occupations	15.47	13.92
Administrative and secretarial occupations.	8.94	10.09
Skilled trades occupations	9.34	10.68
Caring, leisure and other service occupations	8.60	9.39
Sales and customer service occupations	7.82	6.98
Process, plant and machine operatives	4.89	7.05
Elementary occupations	10.81	9.97
Industry (SIC section)		
N/A	34.26	43.73
Accommodation and food service activities	6.96	3.75
Activities of extraterritorial organisations (...)	0.01	0.01
Activities of households as employers (...)	0.02	0.06
Administrative and support service activities	3.60	3.02
Agriculture, forestry and fishing	0.15	0.10
Arts, entertainment and recreation	0.94	0.62
Construction	2.27	2.17
Education	9.16	7.66
Electricity, gas, steam and air conditioning supply	0.15	0.12
Financial and insurance activities	3.37	1.59
Human health and social work activities	8.43	13.48
Information and communication	2.82	1.06
Manufacturing	5.88	6.69
Mining and quarrying	0.34	0.08
Other service activities	1.44	1.16
Professional, scientific and technical activities	8.08	4.36
Public administration and defence (...)	1.63	2.55
Real estate activities	1.15	1.06

Table S2: Comparison between job postings with wage record and without wage record, categorical variables (Continued on next page).

	Sample without wage record (%)	**Sample with wage record** (%)
Transportation and storage	1.84	2.92
Water supply; sewerage, waste management (...)	0.28	0.40
Wholesale and retail trade (...)	7.23	3.42
Type of contract		
N/A	26.76	10.63
Intern	0.00	0.00
Permanent	63.43	76.11
Temporary	9.80	13.27
Contract hours		
N/A	28.27	11.42
Full-time	64.79	82.46
Part-time	6.94	6.13
Work from home dummy		
No	98.07	98.09
Yes	1.93	1.91
Internship dummy		
No internship	99.58	99.94
Internship	0.42	0.06
Region		
N/A	23.30	16.25
East Midlands	5.03	6.08
East of England	7.70	8.85
Greater London	17.92	15.68
North East	2.28	2.01
North West	8.37	8.75
South East	14.50	16.37
South West	7.28	7.64
West Midlands	8.10	12.36
Yorkshire and The Humber	5.51	6.01

Table S2 Continued: Comparison between job postings with wage record and without wage record, categorical variables.

1.4 Maximum and minimum hourly wages

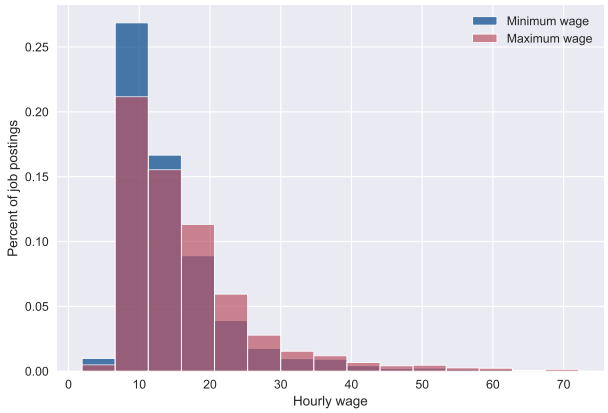

Figure S5: Distribution of the minimum and maximum hourly wage in the Lightcast data (sample with wage record, N=373,521).

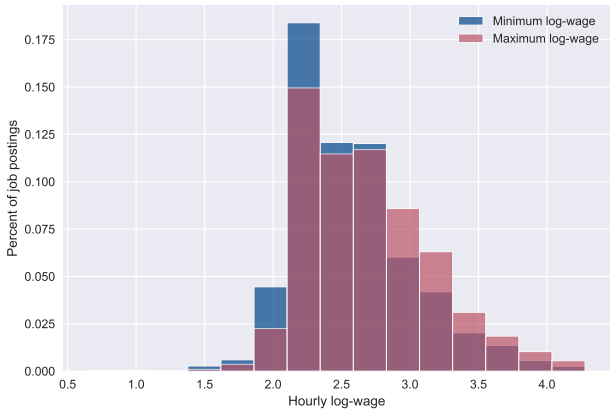

Figure S6: Distribution of the minimum and maximum hourly log-wage in the Lightcast data (sample with wage record, N=373,521).

2 Methodological Appendix

2.1 The LDA Biterm approach

2.1.1 The biterm topic model

The strengths of the LDA model also come with a limitation: it infers topics from word co-occurrence patterns at the document level, which can be sparse in short documents like those in the Lightcast data. To circumvent this issue, we use a variant designed for short texts (Yan et al., 2013), the biterm topic model (BTM). BTM differs from LDA in two important ways. First, it uses unordered word pairs, or *biterms*, as semantic units, which allow direct modeling of word co-occurrence patterns (Yan et al., 2013). For example, a job posting listing three skill requirements (Creativity, Animation, Editing) generates six associated biterms (Creativity-Animation, Creativity-Editing, Animation-Editing). Second, BTM aggregates biterms across the entire corpus rather than within documents, mitigating sparsity and improving topic learning (Yan et al., 2013). This approach models word co-occurrence patterns at the corpus level, yielding a global topic distribution; Yan et al. (2013) show that individual document topic distributions can still be derived accurately.

2.1.2 Choice of LDA model hyperparameters

Two parameters influence the learning of topics: α and β, which are the two Bayesian hyperparameters that determine the prior distributions of topics per document (α) and the prior distribution of terms per topic (β) in the LDA estimation. We assumed that both the topics and the job postings have a high level of specialization (i.e., $\alpha = \beta = 1/K$) to prevent overestimating the degree of overlap between occupations. While this fits our research purpose, we would like to point out that this assumption shapes our model: our latent topics should not be interpreted as the *true* representation of the skills space of job postings, but as a *possible* representation.

2.1.3 Choice of the number of topics

Another important parameter for LDA models is the number of topics (k). We generated LDA solutions for $k = 1, ...50$ to determine which one would best fit our analytical focus.

There are some quantitative measures to assess the predictive quality of LDA models, such as **perplexity** (see Figure S7). However, as Chang et al. (2009) argue, these measures do not examine the structure of the topics themselves. In other words, maximizing these metrics does not ensure that topics are interpretable and meaningful. Coherence measures have been developed to better quantify topic understandability (Röder et al., 2015), but

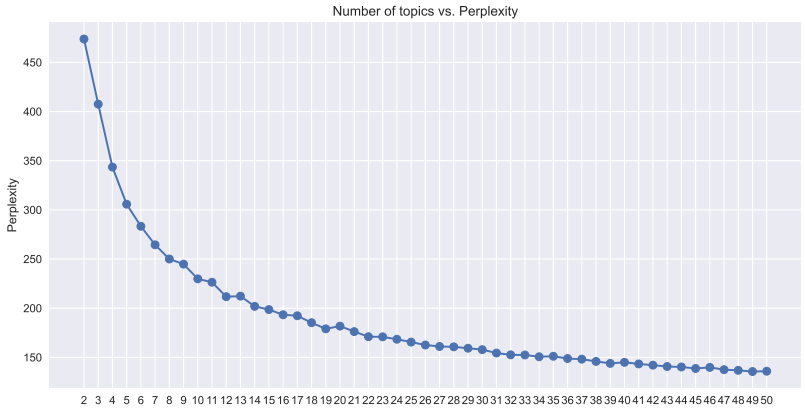

Figure S7: Perplexity by number of topics

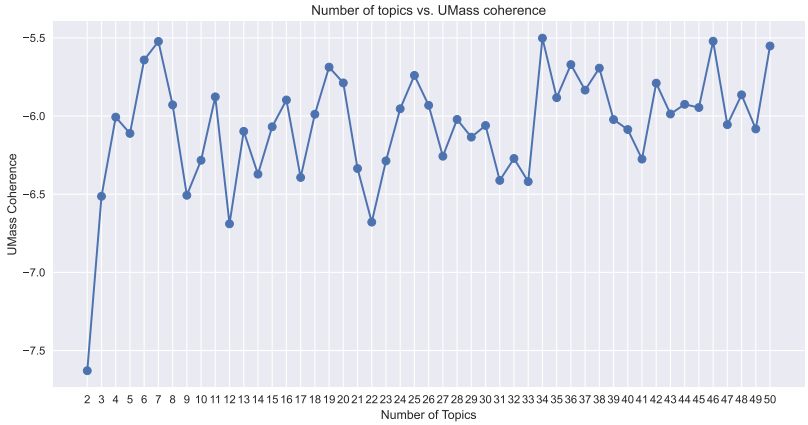

Figure S8: Umass-coherence by number of topics (see Röder et al., 2015, p. 2).

their properties are not well suited to our data. First, coherence measures assess the extent to which a given topic is supported by a reference corpus. In our case, finding a suitable reference corpus to assess the coherence of our latent skill categories has proven difficult because most existing skill classifications, such as O*NET, focus exclusively on generic skills. Second, most coherence measures are based on sliding windows, which implies that the documents of the reference corpus are of sufficient length. This prevents us from using our own dataset as a reference corpus, since job ad texts are typically very short. Therefore, we used the only available measure that is not based on a sliding window: the **Umass coherence** (see Figure S8). However, this measure assumes that the words in the documents are ordered, which is not the case in job ads where the order in which the skill requirements are listed is not informative. As a result, we chose the number of topics mainly based on a thorough examination of the different topic solutions.

Following DiMaggio et al. (2013), we evaluated the latent topic solutions based on whether they provided "substantively meaningful" and "analytically useful" topics for our research purpose (p.582). Our choice of the 19-topic solution illustrates this. On the one hand, this topic solution makes fine distinctions between skills that would otherwise be grouped together. Indeed, while some topics were already well defined with a more parsimonious number of topics ($k < 10$), such as "Office Administration and Management" or "Financial Operations", setting k to 19 ensures that some heterogeneous topics are split. This was especially the case for the topics "Healthcare and Patient Care"/"Caregiving and Support Services" and "Facilities Maintenance"/"Engineering and Technical Expertise". On the other hand, this topic solution avoids unnecessary details. For example, topic solutions with $k > 19$ tend to single out specific ICT skills, such as the Adobe software skills that are part of the "Graphic Design and Creative Media" topic. Moreover, our choice was supported by the available statistical and coherence measures: perplexity decreases only slightly after $k = 19$, and Umass coherence reaches a local maximum at $k = 19$.

An interactive visualization of the LDA 90-topic solution is available on here. See also Table S3 for the 15 most probable skills within each latent topic.

2.2 The MMD distance

For a technical presentation of the Maximum Mean Discrepancy (MMD) distance, see Gretton et al. (2012).

The MMD distance has relevant characteristics for comparing the skill profiles of occupations. This metric is nonparametric and can be used in high dimensional spaces with non-linear constraints such as ours. It is also robust, with a low sensitivity to noise. When used for two-sample testing, the MMD is conservative, that is that it tends to not reject the null hypothesis that two distributions are equal if they only slightly differ. This property is interesting for our case, as it ensures that we are not overemphasizing minor differences between occupations and thereby overestimating the variation in skill content within occupational categories.

The MMD distance is a kernel-based metric that depends on the choice of a kernel function, which is itself parameterized by a specific bandwidth γ (see Schrab et al., 2023). The role of γ is similar to that of the bandwidth h in kernel density estimation, which controls the smoothing of the estimated density. The higher γ, the lower the smoothing ($\gamma = 1/h^2$), the higher the level of granularity retained. Because the choice of the bandwidth has a strong influence on the resulting estimate, we completed multiple robustness checks to assess the stability of our results for different values of γ (see a comparison in Figures S9-S12 for the minor SOC categories). Overall, the scale of the distances changes with γ, but the identified patterns are very stable.

Our final model was obtained with a RBF Gaussian kernel and a bandwidth determined via the median heuristic (see Muandet et al., 2017, p. 54). This corresponds to $\gamma = 1/Me^2 = 1.49$, with Me the median of the pairwise distances between all pairs of occupations. This kernel-bandwidth combination has been shown to work well in many applications, especially in the context of large sample sizes such as ours (Muandet et al., 2017).

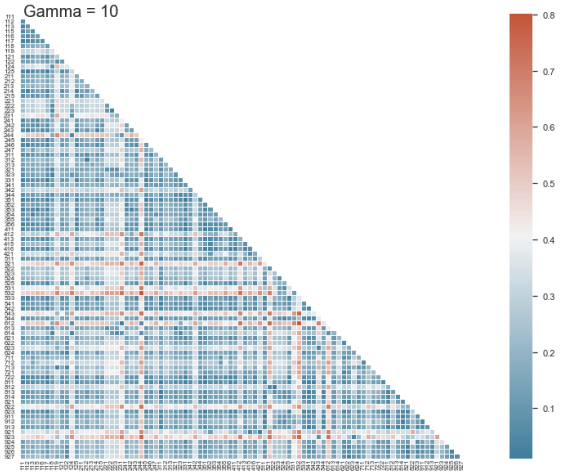

Figure S9: Empirical heat maps for the MMD distance between minor groups, with Gaussian kernel and bandwidth $\gamma = 10$

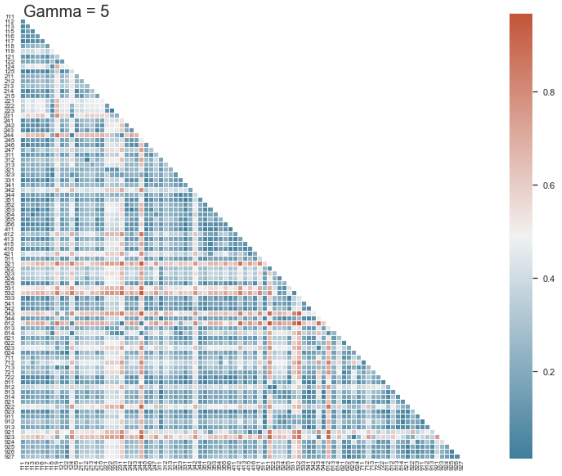

Figure S10: Empirical heat maps for the MMD distance between minor groups, with Gaussian kernel and bandwidth $\gamma = 5$

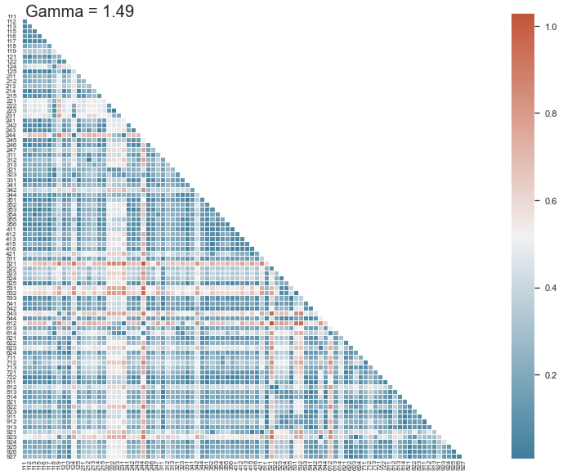

Figure S11: Empirical heat maps for the MMD distance between minor groups, with Gaussian kernel and bandwidth $\gamma = 1.49$ (median heuristic)

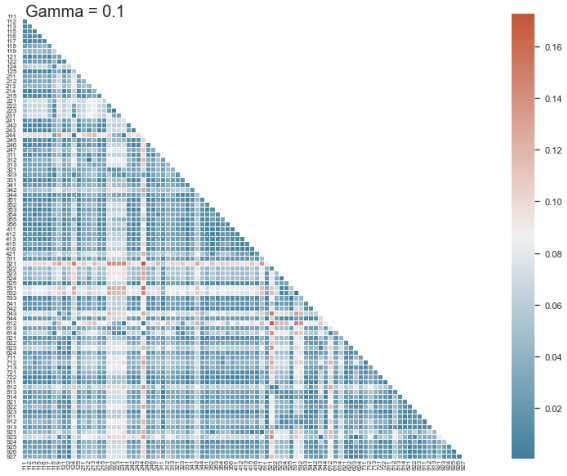

Figure S12: Empirical heat maps for the MMD distance between minor groups, with Gaussian kernel and bandwidth $\gamma = 0.1$

3 Supplementary results

3.1 LDA biterm Model

3.1.1 The chosen 19-topic solution

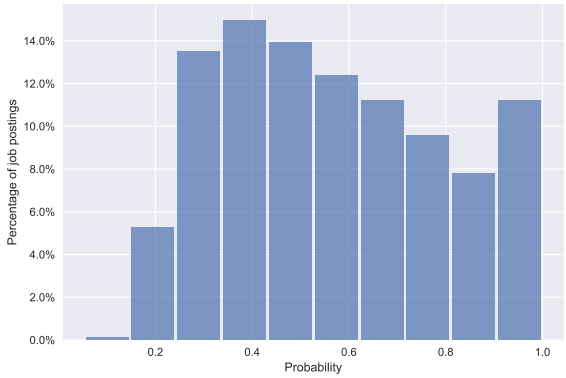

Figure S13: Distribution of the highest topic probability within the sampled job postings

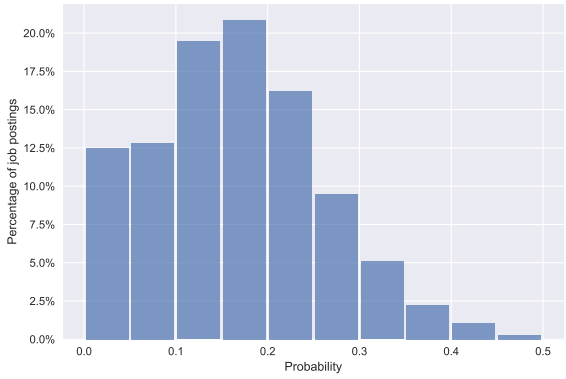

Figure S14: Distribution of the second highest topic probability within the sampled job postings

	Latent skill category	**The 15 most probable skills**, by decreasing order ($\lambda = 1$)
1	Project Management	budgeting, communication skills, planning, project management, teamwork collaboration, building effective relationships, stakeholder management, problem solving, organisational skills, key performance indicators KPIS, people management, staff management, quality management, performance management, change management
2	Office administration and management	Microsoft Excel, communication skills, organisational skills, detail orientated, Microsoft office, administrative support, customer service, Microsoft Word, secretarial skills, Microsoft PowerPoint, typing, spreadsheets, data entry, Microsoft outlook, general office duties
3	Communication and Interpersonal Abilities	communication skills, teamwork collaboration, detail orientated, organisational skills, customer service, problem solving, writing, English, Microsoft Excel, verbal oral communication, time management, listening, French, German, Spanish
4	Sales and Business Development	sales, communication skills, business development, customer service, building effective relationships, sales goals, sales management, account management, teamwork collaboration, product sales, customer contact, business to business, prospective clients, client base retention, tele sales
5	Caregiving and Support Services	teaching, communication skills, working with patient and/or condition: mental health, planning, creativity, childcare, English, care planning, organisational skills, dementia knowledge, nursing home, autism diagnosis treatment care, social services, home management, learning disability
6	Customer Service and Retail Operations	communication skills, customer service, cleaning, teamwork collaboration, organisational skills, retail industry knowledge, English, cooking, detail orientated, food safety, stock control, cash handling, store management, food preparation, housekeeping
7	Digital Marketing and Content Strategy	social media, creativity, marketing, communication skills, writing, budgeting, digital marketing, planning, research, teamwork collaboration, detail orientated, editing, marketing management, market strategy, google analytics
8	Financial Operations	accounting, Microsoft Excel, finance, budgeting, communication skills, account reconciliation, balance sheet, detail orientated, financial reporting, VAT returns, financial accounting, bank reconciliation, accruals, statutory accounts, variance analysis
9	Web Development and Software Engineering	JavaScript, software development, Microsoft hash, SQL, java, software engineering, net, git, python, devops, scrum, active server pages ASP, aspnet, continuous integration ci, angularJS
10	Logistics and Supply Chain Management	procurement, communication skills, planning, purchasing, Microsoft Excel, logistics, supply chain knowledge, supply chain management, key performance indicators KPIS, sap, enterprise resource planning ERP, procurement contracts, contract management, manufacturing resource planning MRP, material requirement planning MRP
11	Facility Maintenance	communication skills, plumbing, preventive maintenance, customer service, predictive preventative maintenance, painting, cleaning, electrical work, carpentry, HVAC, wiring, boilers, hand tools, heating systems, emergency lighting

Table S3: Description of the 19 latent skill categories obtained from the biterm topic model, using the 15 most probable skills by decreasing order, with $\lambda = 1$ (Continued on next page).

	Latent skill category	**The 15 most probable skills**, by decreasing order ($\lambda = 1$)
12	Healthcare and Patient Care	patientcare, communication skills, surgery, teamwork collaboration, teaching, research, working with patient and/or condition: trauma, rehabilitation, leadership, computer literacy, anaesthesiology, primary care, x-rays, dentistry, paediatrics
13	Business strategy	communication skills, research, teamwork collaboration, business development, Microsoft Excel, risk management, building effective relationships, planning, project management, accounting, analytical skills, due diligence, asset management industry knowledge, economics, insurance underwriting
14	Engineering and Technical Expertise	communication skills, project management, AutoCAD, planning, mechanical engineering, budgeting, commissioning, calculation, problem solving, research, civil engineering, engineering design and installation, engineering design, systems engineering, mechanical design
15	Manufacturing and Engineering	computer numerical control CNC, machining, engineering drawings, quality assurance and control, communication skills, welding, problem solving, detail orientated, quality management, teamwork collaboration, manufacturing processes, iso9001standards, lathes, lean manufacturing, MIG and TIG welding
16	Data Management and Analysis	SQL, python, communication skills, data analysis, problem solving, Microsoft Excel, machine learning, teamwork collaboration, data science, tableau, business intelligence, Microsoft Power BI, bigdata, data warehousing, extraction transformation and loading ETL
17	Technical Support and Troubleshooting	troubleshooting, Microsoft active directory, communication skills, technical support, VMware, cisco, ITIL, Windows server, ITsupport, Linux, Microsoft windows, domain name system DNS, wide area network WAN, transmission control protocol internet protocol TCPIP, dynamic host configuration protocol DHCP
18	Graphic Design and Creative Media	creativity, Adobe Photoshop, Adobe InDesign, Adobe acrobat, Adobe creative suite, Adobe illustrator, graphic design, communication skills, detail orientated, teamwork collaboration, editing, Adobe aftereffects, digital design, typesetting, animation
19	Scientific Research and Laboratory Work	research, communication skills, teamwork collaboration, chemistry, biology, teaching, clinical trials, biotechnology, quality assurance and control, English, cancer knowledge, experiments, oncology, clinical research, biochemistry

Table S3 Continued: Description of the 19 latent skill categories obtained from the biterm topic model, using the 15 most probable skills by decreasing order, with $\lambda = 1$.

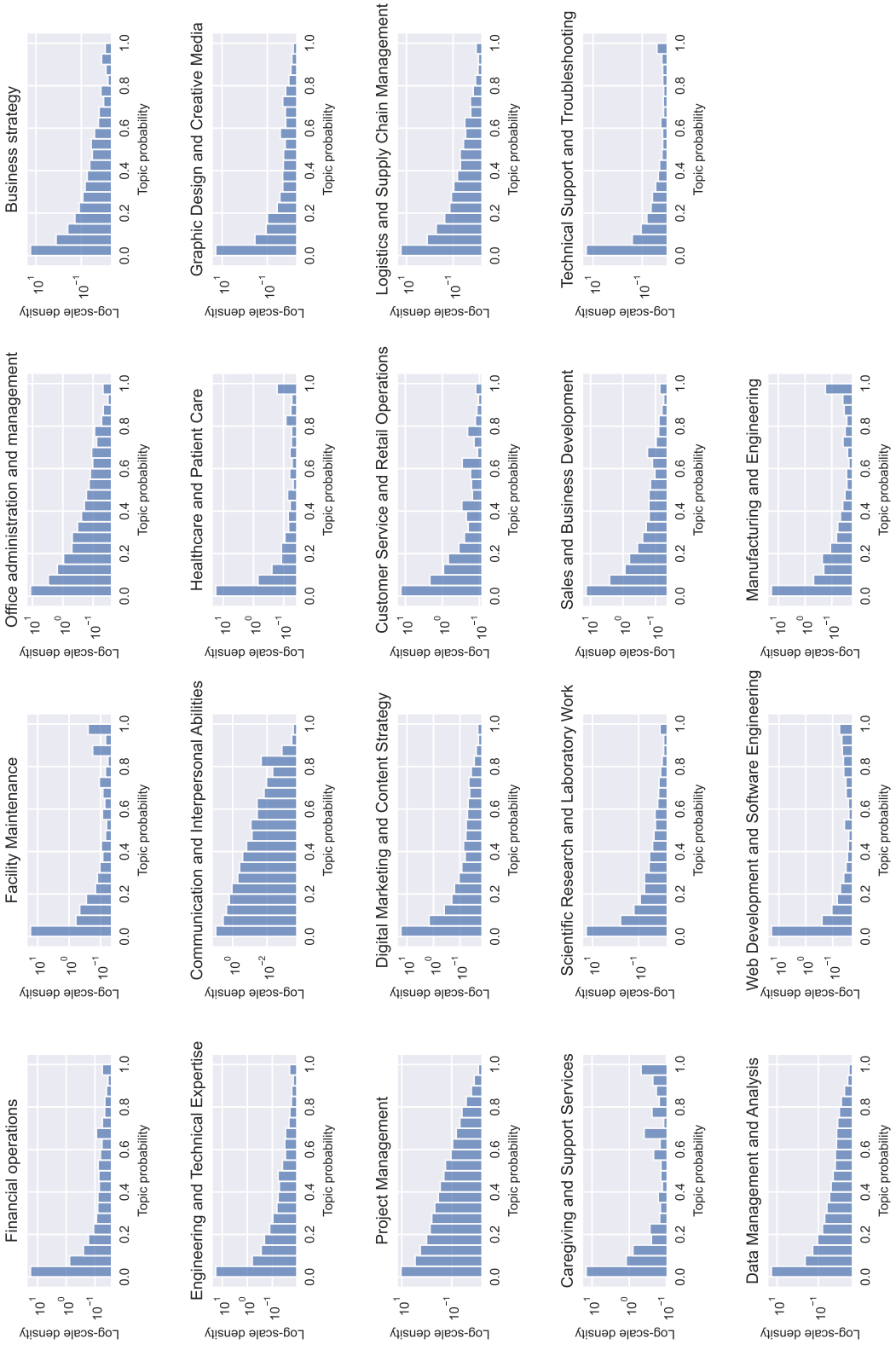

Figure S15: Probability distribution of each latent skill category within the sampled job postings, with the standard log-scale density function. *Note: Y-axes differ between latent skill categories.*

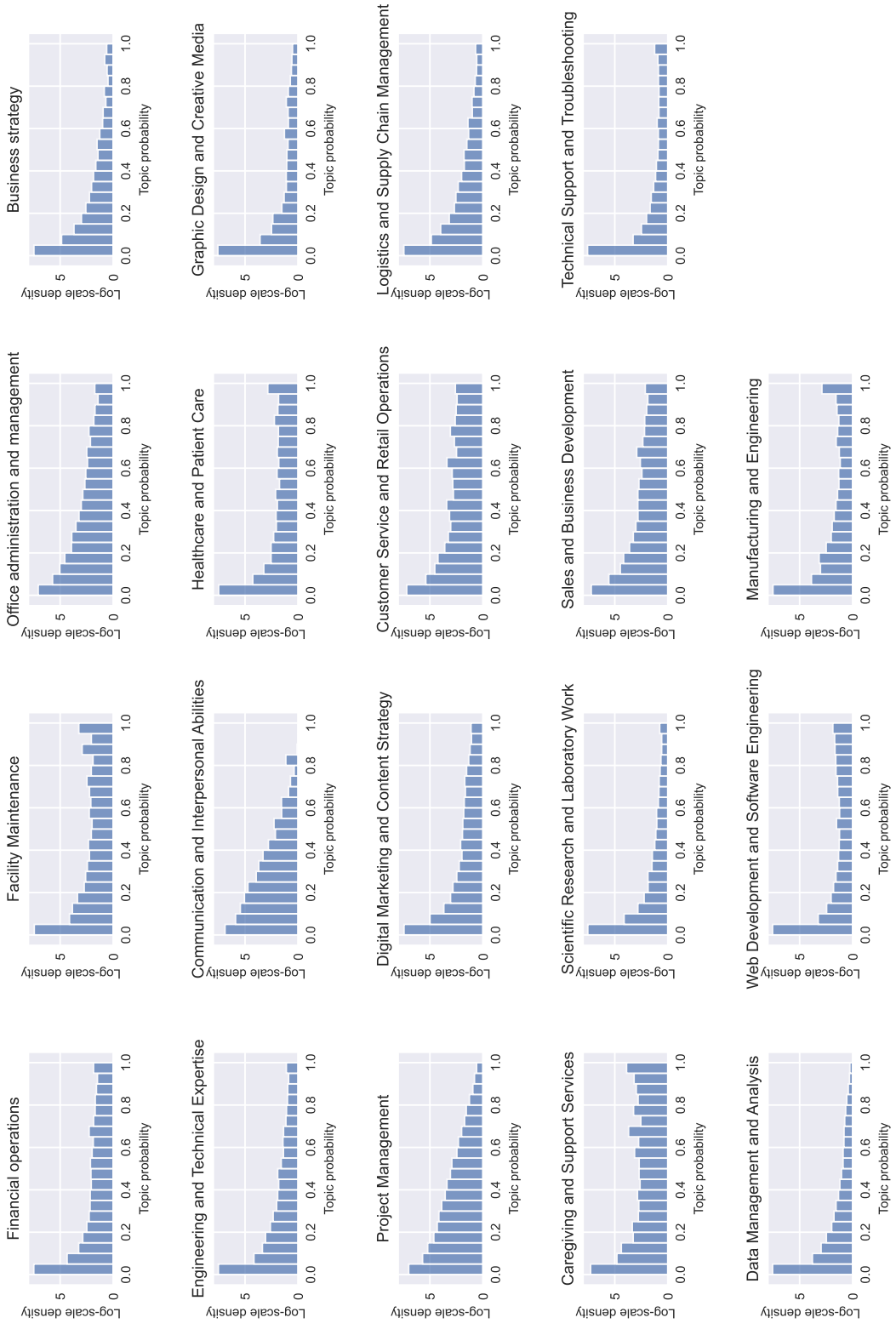

Figure S16: Probability distribution of each latent skill category within the sampled job postings, with a custom log-scale density function. *Note: We applied the log transformation* $f(x) = log(1 + 100x)$.

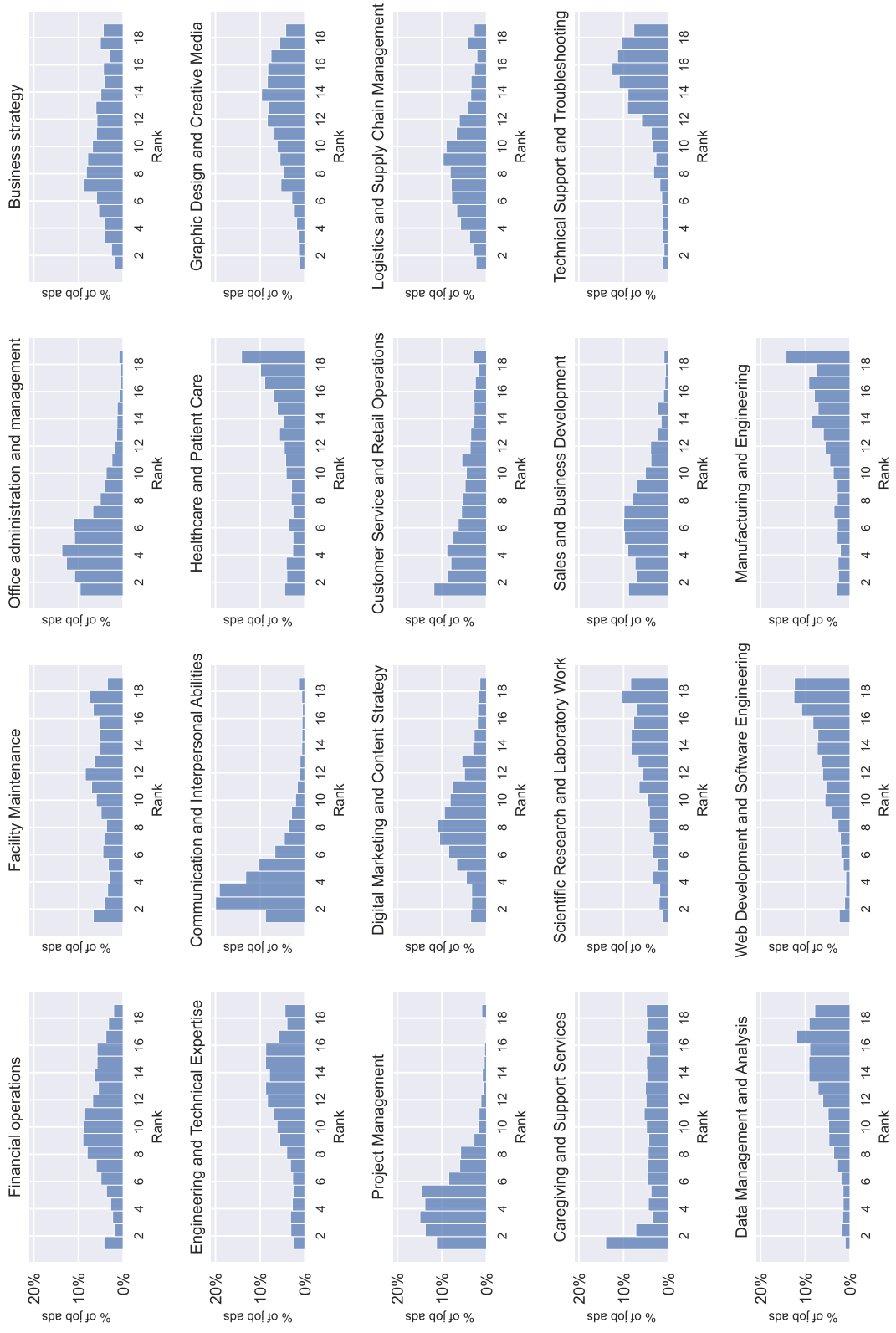

Figure S17: Distribution of the probability rank of each latent skill category within the sampled job postings. *Reading: the topic "Communication and Interpersonal skills' ranks second in 20% of the sampled job postings.*

3.2 MMD distance matrix

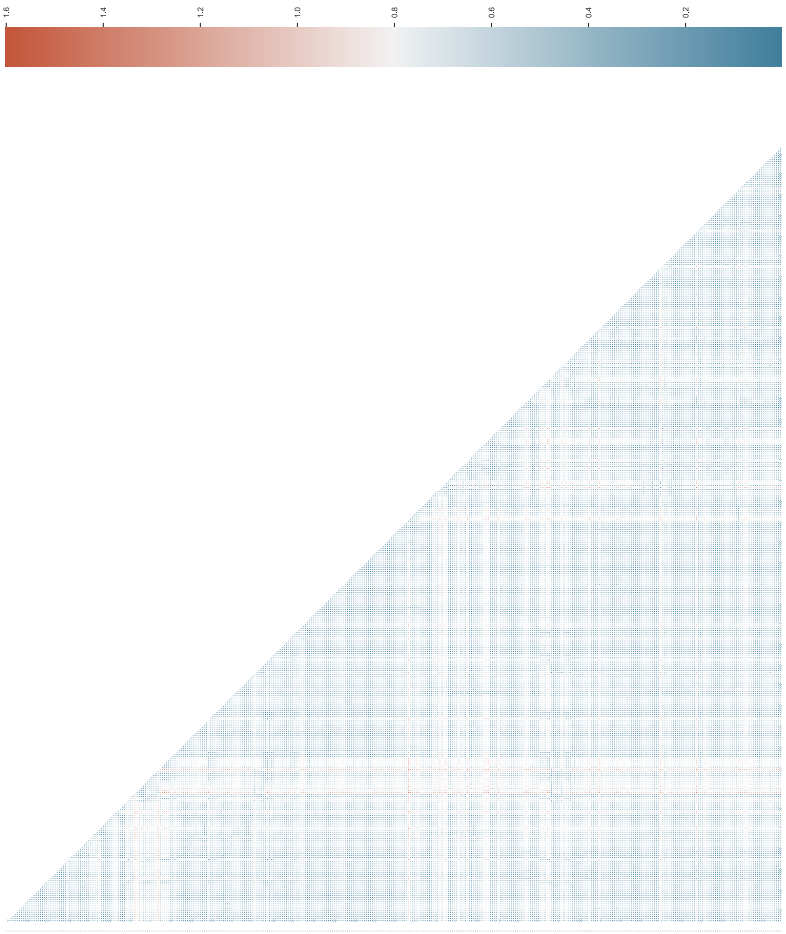

Figure S18: Empirical heat maps for the MMD distance between unit groups.
Note: The x- and y-axes are labeled from the coding in the SOC occupational classification. Each cell indicates the maximum mean discrepancy pairwise distance between unit groups based on their representation in the latent skill space. We use a kernel bandwidth of $\lambda = 1.49$. A high-quality version of the Figure is available on https://surfdrive.surf.nl/files/index.php/s/0Z7EeBbPSuCzYE3.

3.3 Wage regressions

We estimate the following OLS models:

$$log(w) = \alpha + \sum_{k=1}^{18} \beta_k p_k + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where $log(w)$ is the logged hourly wage offered in the job posting, and p_k is the probability of the job posting to contain the latent skill category k, $k = 1...K-1$ with K=19.

$$log(w) = \alpha + \sum_{s=1}^{n_{soc}-1} \gamma_s \mathbb{1}_{\{soc=s\}} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

where *soc* is the Standard Occupational Classification (SOC) coding of the job postings. We estimate model (2) with the four available levels of the occupational classification, from 1 ($n_{soc} = 9$) to 4 digits ($n_{soc} = 369$), resulting in four regressions. We finally combine (1) and (2) to assess the joint predictive power of skill profiles and occupational coding:

$$log(w) = \alpha + \sum_{k=1}^{18} \beta_k p_k + \sum_{s=1}^{n_{soc}-1} \gamma_s \mathbb{1}_{\{soc=s\}} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

As in model (2), we estimate a separate regression for each level of the occupational classification, resulting in four regressions.

Additional outputs

Tables S4-S8 provide the full regression results for the comparison of models (1)-(3), for both the logged minimum and maximum hourly wage. In the interest of space, we only include the full results of the models for the major and sub-major groups. If you are interested in the results for the minor and unit groups, please feel free to contact the corresponding author.

3.3.1 Logged minimum hourly wage

	Model (1)		Model (2)		Model (3)	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
Constant	2.448	0.004	2.903	0.002	2.718	0.004
Skill profiles						
Financial operations	0.216	0.006			0.168	0.006
Facility Maintenance	0.210	0.005			0.189	0.005
Office administration and management	-0.354	0.005			-0.195	0.006
Business strategy	0.932	0.010			0.778	0.009
Engineering and Technical Expertise	0.466	0.008			0.318	0.007
Communication and Interpersonal Abilities	-0.484	0.007			-0.338	0.007
Healthcare and Patient Care	0.372	0.006			0.273	0.006
Graphic Design and Creative Media	0.164	0.010			0.133	0.009
Project Management	0.774	0.006			0.547	0.006
Digital Marketing and Content Strategy	0.260	0.007			0.112	0.007
Customer Service and Retail Operations	-0.322	0.005			-0.185	0.005
Logistics and Supply Chain Management	0.415	0.008			0.336	0.008
Caregiving and Support Services	0.012	0.005			0.019	0.005
Scientific Research and Laboratory Work	0.281	0.010			0.087	0.010
Sales and Business Development	0.083	0.005			0.069	0.005
Technical Support and Troubleshooting	0.453	0.009			0.353	0.008
Data Management and Analysis	0.881	0.011			0.750	0.011
Web Development and Software Engineering	0.677	0.007			0.503	0.007
Manufacturing and Engineering	ref.	ref.			ref.	ref.
SOC coding						
Managers, directors and senior officials			ref.	ref.	ref.	ref.
Professional occupations (occ.)			-0.071	0.002	-0.084	0.002
Associate professionals and technical occ.			-0.292	0.002	-0.249	0.002
Administrative and secretarial occ.			-0.540	0.003	-0.389	0.003
Skilled trades occ.			-0.325	0.003	-0.231	0.003
Caring, leisure and other service occ.			-0.664	0.003	-0.503	0.003
Sales and customer service occ.			-0.506	0.003	-0.357	0.003
Process, plant and machine operatives			-0.480	0.003	-0.317	0.003
Elementary occupations			-0.670	0.003	-0.450	0.003
No. Observations:	373521		373521		373521	
Degrees of freedom Residuals:	373502		373512		373494	

Table S4: OLS regressions of the logged minimum hourly wage, using the major groups as occupational code (Continued on next page).

Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 23.

	Model (1)		**Model (2)**		**Model (3)**	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
Degrees of freedom Model:	18		8		26	
AIC:	3.15E+05		3.21E+05		2.56E+05	
BIC:	3.15E+05		3.21E+05		2.56E+05	
R-squared:	0.301		0.289		0.403	
Adj. R-squared:	0.301		0.289		0.403	
F-statistic:	8936		18980		9714	

Table S4 Continued: OLS regressions of the logged minimum hourly wage, using the major groups as occupational code.

Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 23.

	Model (1)		Model (2)		Model (3)	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
Constant	2.448	0.004	2.937	0.002	2.744	0.005
Skill profiles						
Financial operations	0.216	0.006			0.168	0.006
Facility Maintenance	0.210	0.005			0.162	0.005
Office administration and management	-0.354	0.005			-0.202	0.006
Business strategy	0.932	0.010			0.757	0.009
Engineering and Technical Expertise	0.466	0.008			0.348	0.008
Communication and Interpersonal Abilities	-0.484	0.007			-0.290	0.007
Healthcare and Patient Care	0.372	0.006			0.248	0.006
Graphic Design and Creative Media	0.164	0.010			0.107	0.010
Project Management	0.774	0.006			0.557	0.006
Digital Marketing and Content Strategy	0.260	0.007			0.097	0.007
Customer Service and Retail Operations	-0.322	0.005			-0.174	0.005
Logistics and Supply Chain Management	0.415	0.008			0.334	0.008
Caregiving and Support Services	0.012	0.005			0.074	0.005
Scientific Research and Laboratory Work	0.281	0.010			0.153	0.010
Sales and Business Development	0.083	0.005			0.047	0.006
Technical Support and Troubleshooting	0.453	0.009			0.394	0.009
Data Management and Analysis	0.881	0.011			0.761	0.011
Web Development and Software Engineering	0.677	0.007			0.537	0.007
Manufacturing and Engineering	ref.	ref.			ref.	ref.
SOC coding						
11			ref.	ref.	ref.	ref.
12			-0.120	0.004	-0.109	0.004
21			0.009	0.003	-0.145	0.004
22			-0.114	0.003	-0.073	0.004
23			-0.334	0.003	-0.227	0.004
24			-0.022	0.003	-0.088	0.003
31			-0.350	0.005	-0.348	0.005
32			-0.416	0.005	-0.332	0.005
33			-0.316	0.006	-0.263	0.006
34			-0.327	0.005	-0.235	0.005
35			-0.301	0.003	-0.267	0.003
41			-0.565	0.003	-0.434	0.003
42			-0.608	0.005	-0.356	0.005

Table S5: OLS regressions of the logged minimum hourly wage, using the sub-major groups as occupational code (Continued on next page).

Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 22.

	Model (1)		Model (2)		Model (3)	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
51			-0.565	0.006	-0.459	0.006
52			-0.343	0.004	-0.252	0.004
53			-0.236	0.004	-0.179	0.004
54			-0.518	0.005	-0.292	0.005
61			-0.710	0.003	-0.588	0.003
62			-0.646	0.005	-0.433	0.005
71			-0.505	0.003	-0.362	0.003
72			-0.626	0.005	-0.434	0.005
81			-0.468	0.004	-0.339	0.004
82			-0.542	0.004	-0.354	0.004
91			-0.604	0.005	-0.468	0.005
92			-0.725	0.003	-0.488	0.003
No. Observations:	**373521**		**373521**		**373521**	
Degree of freedom Residuals:	373502		373496		373478	
Degree of freedom Model:	18		24		42	
AIC:	3.15E+05		3.05E+05		2.49E+05	
BIC:	3.15E+05		3.05E+05		2.50E+05	
R-squared:	0.301		0.32		0.414	
Adj. R-squared:	0.301		0.32		0.414	
F-statistic:	8936		7339		6284	

Table S5 Continued: OLS regressions of the logged minimum hourly wage, using the sub-major groups as occupational code.

Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 22.

Model (2) vs. (3)	Likelihood ratio	p-value	Degrees of freedom
SOC Major groups	65507	0.0000	18
SOC Sub-Major groups	55367	0.0000	18
SOC Minor groups	47194	0.0000	18
SOC Unit groups	41816	0.0000	18

Table S6: Results of the Likelihood ratio tests for predicting of the logged minimum hourly wage, comparing the nested models Model (2) including the SOC variable and Model (3) adding the skill profiles of job postings.

3.3.2 Logged maximum hourly wage

	Model (1)		Model (2)		Model (3)	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
Constant	2.5609	0.004	3.0254	0.002	2.8185	0.005
Skill profiles						
Financial operations	0.2502	0.006			0.1976	0.006
Facility Maintenance	0.1859	0.005			0.1721	0.005
Office administration and management	-0.3984	0.006			-0.2165	0.006
Business strategy	1.0579	0.01			0.8725	0.009
Engineering and Technical Expertise	0.5714	0.008			0.3856	0.008
Communication and Interpersonal Abilities	-0.5332	0.007			-0.3734	0.007
Healthcare and Patient Care	0.4531	0.006			0.3275	0.006
Graphic Design and Creative Media	0.2414	0.01			0.1942	0.01
Project Management	0.7675	0.006			0.5334	0.006
Digital Marketing and Content Strategy	0.2733	0.007			0.1228	0.007
Customer Service and Retail Operations	-0.4039	0.005			-0.2347	0.005
Logistics and Supply Chain Management	0.4489	0.009			0.3712	0.008
Caregiving and Support Services	0.06	0.005			0.0601	0.005
Scientific Research and Laboratory Work	0.3904	0.011			0.1507	0.01
Sales and Business Development	0.1844	0.005			0.1792	0.006
Technical Support and Troubleshooting	0.49	0.009			0.3693	0.009
Data Management and Analysis	0.9236	0.012			0.754	0.011
Web Development and Software Engineering	0.7939	0.007			0.5764	0.007
Manufacturing and Engineering	ref.	ref.			ref.	ref.
SOC coding						
Managers, directors and senior officials			ref.	ref.	ref.	ref.
Professional occupations (occ.)			-0.0025	0.002	-0.0269	0.002
Associate professionals and technical occ.			-0.2709	0.003	-0.2291	0.002
Administrative and secretarial occ.			-0.5589	0.003	-0.3924	0.003
Skilled trades occ.			-0.3471	0.003	-0.2262	0.003
Caring, leisure and other service occ.			-0.6833	0.003	-0.5215	0.003
Sales and customer service occ.			-0.4945	0.003	-0.3602	0.003
Process, plant and machine operatives			-0.4871	0.003	-0.3001	0.003
Elementary occupations			-0.7346	0.003	-0.4808	0.003
No. Observations:	373,521		373,521		373,521	
Degrees of freedom Residuals:	373,502		373,512		373,494	

Table S7: OLS regressions of the logged maximum hourly wage, using the major groups as occupational code (Continued on next page).

Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 23.

	Model (1)		**Model (2)**		**Model (3)**	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
Degrees of freedom Model:	18		8		26	
AIC:	3.48E+05		3.50E+05		2.79E+05	
BIC:	3.48E+05		3.50E+05		2.80E+05	
R-squared:	0.328		0.325		0.441	
Adj. R-squared:	0.328		0.325		0.441	
F-statistic:	10120		22460		11320	

Table S7 Continued: OLS regressions of the logged maximum hourly wage, using the major groups as occupational code.

Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 23.

	Model (1)		Model (2)		Model (3)	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
Constant	2.5609	0.004	3.0662	0.002	2.8535	0.005
Skill profiles						
Financial operations	0.2502	0.006			0.1959	0.006
Facility Maintenance	0.1859	0.005			0.1417	0.006
Office administration and management	-0.3984	0.006			-0.2237	0.006
Business strategy	1.0579	0.01			0.8574	0.01
Engineering and Technical Expertise	0.5714	0.008			0.434	0.008
Communication and Interpersonal Abilities	-0.5332	0.007			-0.3403	0.007
Healthcare and Patient Care	0.4531	0.006			0.3175	0.006
Graphic Design and Creative Media	0.2414	0.01			0.1458	0.011
Project Management	0.7675	0.006			0.5441	0.006
Digital Marketing and Content Strategy	0.2733	0.007			0.0979	0.007
Customer Service and Retail Operations	-0.4039	0.005			-0.2256	0.006
Logistics and Supply Chain Management	0.4489	0.009			0.3613	0.008
Caregiving and Support Services	0.06	0.005			0.0782	0.005
Scientific Research and Laboratory Work	0.3904	0.011			0.1799	0.01
Sales and Business Development	0.1844	0.005			0.14	0.006
Technical Support and Troubleshooting	0.49	0.009			0.4226	0.009
Data Management and Analysis	0.9236	0.012			0.785	0.011
Web Development and Software Engineering	0.7939	0.007			0.6347	0.007
Manufacturing and Engineering	ref.	ref.			ref.	ref.
SOC coding						
11			ref.	ref.	ref.	ref.
12			-0.1427	0.004	-0.1171	0.004
21			0.0558	0.003	-0.1226	0.004
22			-0.0732	0.004	-0.0414	0.004
23			-0.2	0.004	-0.0777	0.004
24			0.0139	0.003	-0.0542	0.003
31			-0.3507	0.005	-0.3425	0.005
32			-0.4228	0.005	-0.3322	0.005
33			-0.333	0.006	-0.2654	0.006
34			-0.2992	0.005	-0.2041	0.005
35			-0.2771	0.003	-0.2424	0.003

Table S8: OLS regressions of the logged maximum hourly wage, using the sub-major groups as occupational code (Continued on next page).

Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 22.

	Model (1)		Model (2)		Model (3)	
	coef.	*SE*	*coef.*	*SE*	*coef.*	*SE*
41			-0.588	0.003	-0.442	0.003
42			-0.647	0.005	-0.3673	0.005
51			-0.6004	0.006	-0.4652	0.006
52			-0.3557	0.004	-0.2455	0.004
53			-0.269	0.004	-0.1832	0.004
54			-0.569	0.005	-0.2985	0.005
61			-0.7284	0.003	-0.5915	0.003
62			-0.7056	0.005	-0.4594	0.005
71			-0.4747	0.004	-0.3489	0.004
72			-0.6823	0.005	-0.4724	0.005
81			-0.4897	0.004	-0.3406	0.005
82			-0.5511	0.004	-0.334	0.004
91			-0.66	0.005	-0.4925	0.005
92			-0.7989	0.003	-0.5241	0.003
No. Observations:	373,521		373,521		373,521	
Degrees of freedom Residuals:	373,502		373,496		373,478	
Degrees of freedom Model:	18		24		42	
AIC:	3.48E+05		3.36E+05		2.74E+05	
BIC:	3.48E+05		3.36E+05		2.74E+05	
R-squared:	0.328		0.349		0.449	
Adj. R-squared:	0.328		0.349		0.449	
F-statistic:	10120		8346		7249	

Table S8 Continued: OLS regressions of the logged maximum hourly wage, using the sub-major groups as occupational code.
Note: Columns include the coefficients (coef.) and their standard errors (SE). All coefficients are statistically significant below the 1% level. Models (1)-(3) are formally defined on page 22.

Model (2) vs. (3)	**Likelihood ratio**	**p-value**	**Degrees of freedom**
SOC Major groups	70283	0.0000	18
SOC Sub-Major groups	62327	0.0000	18
SOC Minor groups	52096	0.0000	18
SOC Unit groups	46788	0.0000	18

Table S9: Results of the Likelihood ratio tests for predicting of the logged maximum hourly wage, comparing the nested models Model (2) including the SOC variable and Model (3) adding the skill profiles of job postings.

4 Robustness checks

4.1 Omission bias in job postings

Our results are based on the assumption that job postings accurately reflect the skills required for a given job. However, this may not be the case if employers omit certain skill requirements that they deem obvious or implied by other information provided in the job posting, such as the job title or educational requirements. This may be particularly true of the so-called "regulated professions", which require certain qualifications or experience by law.[1] The omission of job-specific knowledge or know-how could bias our results by artificially reducing the difference in skill content between occupations.

Although it is difficult to assess the extent of potential omission bias systematically, we provide descriptive statistics to evaluate how often job postings for regulated professions include job-specific skill requirements, even when these should be implied by the job title alone. We selected eight job titles from the Lightcast data that correspond to regulated professions in the UK: "Architect", "Dental Nurse", "Dentist", "Lawyer", "Nursery Nurse" 'Pharmacist", "Radiographer", "Veterinary Surgeon".[2] We voluntarily choose regulated professions that are close in terms of job content but differ in terms of level of expertise or required level of education (e.g., "Dental Nurse", "Dentist"). To ensure there are enough job postings for each job title, we use the full set of job postings available in 2019 (N=6,346,064). Table S10 gives the number of job postings matching the selected job titles.

	Count	**Percentage (%)**
Architect	1,369	0.022
Dental Nurse	3,073	0.048
Dentist	315	0.005
Lawyer	191	0.003
Nursery Nurse	3,716	0.059
Pharmacist	1,795	0.028
Radiographer	547	0.009
Veterinary Surgeon	1,069	0.017

Table S10: Counts and percentages of the job postings exactly matching the job titles for the eight selected regulated professions.
Source: 2019 Lightcast data.

First, we analyze the top 20 most frequent skill requirements included in the job postings advertising the selected regulated professions (see Tables S12–S13). We find that, for all job titles considered, the job postings include skills that are directly related to the advertised regulated

[1] A list of the regulated professions in the UK can be found via the following link: https://www.gov.uk/government/publications/professions-regulated-by-law-in-the-uk-and-their-regulators/uk-regulated-professions-and-their-regulators [accessed on 12/06/2015]

[2] Only job titles that exactly match the protected title as given in the Professional Qualifications Act were selected, in order to ensure that the job postings we consider advertise for regulated professions.

profession (e.g. 'Surgery', mentioned in 100% of job postings for a Veterinary Surgeon). The degree to which job-specific skill requirements are included varies. For example, while 'Dentistry' is mentioned in all the job postings advertising for a Dentist, it appears in about 40% of the job postings advertising for a Dental Nurse, in similar proportions to the skill requirement 'Surgery' (38%). However, we observe a variety of job-specific skills for all regulated professions. These can be related to professional knowledge (e.g. 'haematology' or 'biochemistry' for a Veterinary Surgeon), specialized software (e.g. 'Revit' and 'AutoCAD' for an Architect) or detailed tasks (e.g. 'contract preparation' for a Lawyer). This suggests that employers are explicit about their expectations for the position, even when the job title conveys the main content of the job. This may be because they want to emphasize certain areas of expertise over others; for example, a radiographer with specific knowledge or experience in mammography or X-rays.

The prevalence of job-specific skills can also be gauged by measuring the share of baseline skills in job postings. Baseline skills are transversal skills that are not typically taught or certified, such as communication skills. Such skills should be more prominent in job postings for regulated professions if employers mainly focus on what is *not* implied by the job title. Yet, Table S11 shows that the share of baseline skills is not higher in the selected job postings–it is, in fact, lower for a number of regulated professions in the health domain, such as Dentists. Baseline skills make up a maximum of one third of the skill requirements in the selected job postings, which confirms the idea that they include a majority of job-specific skills or software.

	Mean	**Std**
All job postings	0.3	0.3
Architect	0.3	0.3
Dental Nurse	0.2	0.3
Dentist	0.1	0.2
Lawyer	0.3	0.3
Nursery Nurse	0.3	0.3
Pharmacist	0.3	0.2
Radiographer	0.1	0.2
Veterinary Surgeon	0.1	0.2

Table S11: Mean and standard deviation of the share of baseline skills in the job postings advertising the selected regulated professions. The first row gives the values in the full sample of job postings for reference.
Source: 2019 Lightcast data.

Overall, these descriptive findings suggest that employers tend to include job-specific requirements, even when they are implied by the job title. This provides reassurance that the skill requirements provided by Lightcast are suitable for measuring the skill content of jobs, and that our results are unlikely to be biased by the omission of job-specific skills.

Architect		**Lawyer**	
Skill	Percentage	Skill	Percentage
Revit	64.9	Litigation	36.1
AutoCAD	49.5	Communication Skills	26.7
Planning	32.4	Teamwork / Collaboration	13.1
Communication Skills	31.7	Procurement	12.0
Writing	13.5	Legal Documentation	11.5
SketchUp	12.6	Planning	11.5
Creativity	12.3	Problem Solving	10.5
Project Management	10.2	Creativity	9.9
Adobe Photoshop	9.3	Local Government	9.4
Professional Architect	7.9	Research	7.9
Budgeting	6.9	Organisational Skills	7.9
Project Architecture	6.9	Contract Preparation	7.9
Teamwork / Collaboration	6.4	Procurement Contracts	7.3
Adobe Indesign	6.1	Contract Draughting	7.3
Presentation Skills	5.2	Analytical Skills	6.3
Business Development	4.9	Legal Research	5.8
Organisational Skills	4.2	Customer Service	5.8
Technical Drawings	3.7	Negotiation Skills	5.8
Time Management	3.6	Claims Knowledge	5.8
English	3.4	Writing	5.2

Dental Nurse		**Dentist**	
Skill	Percentage	Skill	Percentage
Dentistry	41.6	Dentistry	100.0
Surgery	37.8	X-Rays	22.5
Patient Care	25.3	Dental Care	19.0
Communication Skills	23.7	English	18.1
Infection Control	16.4	Surgery	14.6
English	13.2	Dental Hygiene	12.4
X-Rays	12.6	Communication Skills	11.1
Dental Hygiene	11.9	Patient Care	9.2
Organisational Skills	11.7	Orthodontics	6.0
Dental Care	11.1	Dentures	5.1
Orthodontics	10.2	Mentoring	4.8
Computer Literacy	6.6	Clinical Experience	4.4
Cleaning	6.6	Rehabilitation	3.8
Teamwork / Collaboration	6.3	Administrative Support	3.8
Dental Equipment	5.7	Anaesthesiology	3.8
Radiography	5.3	Treatment Planning	3.5
Customer Service	4.8	Patient/Family Education (...)	3.2
Working With (...): Hepatitis B	4.5	Patient Treatment	3.2
Social Services	4.4	Infection Control	2.5

Table S12: Top 20 skill requirements in the job postings advertising the job titles "Architect", "Lawyer", "Dental nurse", "Dentist". The percentages show the proportion of job postings for a given job title that include a particular skill requirement. Example: 64.9% of the job postings advertising for an Architect include the skill requirement "Revit". *Source: 2019 Lightcast data.*

Nursery Nurse		**Pharmacist**	
Skill	Percentage	Skill	Percentage
Child Care	73.3	Patient Care	57.4
Teaching	26.9	Communication Skills	55.3
Communication Skills	19.3	Listening	48.1
Planning	19.1	Store Management	47.9
Child Protection	11.3	Patient Safety	43.3
Creativity	10.5	Diabetes Diagnosis / Treatment	11.6
Teamwork / Collaboration	6.3	Public Health and Safety	6.4
Organisational Skills	5.4	Customer Contact	5.5
External Auditing	4.7	Retail Management	5.2
Child Development	4.6	Retail Sales	4.2
English	4.5	Customer Service	3.6
Writing	4.2	Primary Care	3.3
Positive Disposition	3.9	Social Services	2.7
Staff Management	3.4	Organisational Skills	2.7
Energetic	3.4	Dispensing Patients Medication	2.3
Lesson Planning	2.6	Teamwork / Collaboration	2.2
Atlassian Bamboo	2.3	Surgery	2.1
Cleaning	2.3	Staff Management	1.8
Record Keeping	2.2	Problem Solving	1.8
Building Effective Relationships	1.8	Detail-Orientated	1.6

Radiographer		**Veterinary Surgeon**	
Skill	Percentage	Skill	Percentage
Radiography	63.4	Surgery	100.0
Radiology	45.5	Communication Skills	28.2
Communication Skills	23.0	X-Rays	21.6
X-Rays	20.7	Patient Care	19.4
Diagnostic Imaging	17.0	Ultrasound	18.1
Patient Care	16.3	Teamwork / Collaboration	10.4
Ultrasound	11.2	Client Base Retention	9.4
Radiation Protection	9.9	Animal Health	8.2
Organisational Skills	7.9	Dentistry	8.0
Teaching	7.1	Veterinary Medicine	6.7
Staff Management	7.1	Blood Pressure Measurement	5.8
Mammography	6.6	Biochemistry	5.0
DEXA	6.0	Haematology	4.9
Customer Service	5.7	Endoscopy	4.5
Policy Development	5.7	Preventive Services	4.1
Working With (...): Trauma	5.7	Customer Service	3.8
English	5.1	Clinical Experience	3.6
Teamwork / Collaboration	5.1	Orthopaedics	3.5
Nuclear Medicine	4.8	Microsoft Excel	3.2
Quality Assurance and Control	4.0	Microsoft Office	3.1

Table S13: Top 20 skill requirements in the job postings advertising the job titles "Nursery Nurse", "Pharmacist", "Radiographer", "Veterinary Surgeon". The percentages show the proportion of job postings for a given job title that include a particular skill requirement. *Source: 2019 Lightcast data.*

4.2 Alternative wage specifications

4.2.1 Wage models with control variables

To further assess the robustness of our regression results, we add control variables to the models defined in (1)–(3). Although the Lightcast data contain only limited auxiliary information, the available job characteristics should help reduce the risk of capturing spurious correlations between skills and wages. For instance, jobs advertised in rural areas may both require specific skills and offer lower wages than comparable jobs in cities (Rouwendal and Koster, 2025). To mitigate such omitted-variable bias, we include four job characteristics: the type of contract (temporary | permanent | intern), contract hours (full-time | part-time), the industry code of the advertised job based on the Standard Industrial Classification (SIC), and a detailed geographic variable distinguishing 249 counties. Table S14 provides an overview of these variables, including their share of missing values.

Control variable	Number of categories	Share of missing values (%)
Type of contract	3	16.7
Job hours	2	17.8
Industrial code SIC	21	40.2
Geographical county	249	13.3

Table S14: Number of categories and share of missing values of the control variables included in equations (4)–(6)

We then re-estimate the models in (1)–(3) with this set of controls, yielding the specifications in (4)–(6). While there is no clear evidence of selective missingness, we treat missing values as a separate category to avoid loss of sample size.

$$log(w) = \alpha + \sum_{k=1}^{18} \beta_k p_k + controls + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

$$log(w) = \alpha + \sum_{s=1}^{n_{soc}-1} \gamma_s \mathbb{1}_{\{soc=s\}} + controls + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

$$log(w) = \alpha + \sum_{k=1}^{18} \beta_k p_k + \sum_{s=1}^{n_{soc}-1} \gamma_s \mathbb{1}_{\{soc=s\}} + controls + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

able **??** reports the results for the three models. As expected, the R^2 increases after adding the control variables. The increase is around 0.04 for models (4) and (5)—reaching up to 0.07 at the major-group level—and about 0.03 for model (6). The best-performing specification, which

combines topic loads, unit groups, and control variables, attains an R^2 of 0.49. For comparison, a model that includes only the control variables yields an R^2 of 0.12.

Although the R^2 rises with the inclusion of controls, the overall patterns remain unchanged. Topic loads alone perform similarly to the major groups, and they substantially increase explanatory power when combined with any occupational-level variable. The improvement is smaller at the major-group level compared to the model without controls (a premium of 24% versus 40%), but is only slightly reduced at the other levels (for example, an increase of 13% versus 16% at the unit-group level). Combining topic loads with the major groups still provides at least as much explanatory power as using unit groups alone.

		Model (4) *Skill profiles*		**Models (5)** *SOC coding*		**Models (6)** *(1) and (2) combined*	
		R^2	Df	R^2	Df	R^2	Df
		0.347	283				
SOC level	Major groups			0.356	273	0.443	291
	Sub-major groups			0.371	289	0.451	307
	Minor groups			0.399	354	0.467	372
	Unit groups			0.436	633	0.492	651

Table S15: Adjusted coefficient of determination R-squared (R^2) and degrees of freedom (Df) for predicting the logged minimum hourly wage in job postings, comparing the models (4)-(6) including the additional control variables.
Notes: Degrees of freedom represent the number of effective parameters in the models.

4.2.2 Out-of-sample R^2

Our main analysis is an in-sample evaluation, that is the R^2 is calculated on the same data as the one used for the estimation. While our number of estimated coefficients is small compared to the sample size ($N = 373,500$ observations with recorded wage), this approach carries the risk of overfitting and of inadvertently interpreting explanatory power as predictive power (Shmueli, 2010). To address this potential bias, we partitioned the data into disjunct estimation and evaluation sets using K-fold cross-validation (see Verhagen, 2022, for a presentation of the method), with $K = 5$. Table **??** below gives the results for models (1)-(3) where we average the out-of-sample R^2 across the five folds. The standard deviations are very small (below 0.01). As expected given the large sample size, the in-sample and out-of-sample R^2 values are nearly identical for most models, differing by no more than 0.001.

		Model (1) *Skill profiles*		**Models (2)** *SOC coding*		**Models (3)** *(1) and (2) combined*	
		R^2	Df	R^2	Df	R^2	Df
		0.301	18				
SOC level	Major groups			0.289	8	0.403	26
	Sub-major groups			0.320	24	0.414	42
	Minor groups			0.356	89	0.432	107
	Unit groups			0.395	368	0.459	386

Table S16: Out-of-sample coefficient of determination R-squared (R^2) and degrees of freedom (Df) for predicting the logged minimum hourly wage in job postings, comparing the models (1)-(3) and estimated using 5-fold cross-validation.
Notes: Degrees of freedom represent the number of effective parameters in the models.

4.2.3 Sensitivity to different numbers of topics

The partition of the skill space into $k = 19$ latent topics represents only one possible representation of the skill profiles of jobs. In the absence of standard measures that quantify topic interpretability (see section 2.1.3), the choice of the number of topic necessarily involves some degree of arbitrariness. To assess the sensitivity of our regression results to this choice, we re-estimated the models defined in equations (1) and (3) using alternative values of $k \in 5, 10, 20, 50, 100$. While $k = 5$ yields broad, heterogeneous topics, $k = 100$ produces a highly granular representation with very specialized and cohesive topics.

Figure S19 shows that $k = 5$ yields a considerably lower R^2 than $k \geq 10$. The increase in predictive power is substantial from $k = 10$ to $k = 20$ and from $k = 20$ to $k = 50$, whereas it is negligible from $k = 50$ to $k = 100$. These results indicate that the predictive value of the skill profiles increases with the level of granularity, albeit with decreasing returns. Choosing $k = 19$ thus represents a balance between capturing sufficient detail in the skill topics while avoiding an unmanageably large number of topics to interpret. With less emphasis on interpretability, a more granular topic solution (e.g., $k = 50$) would further improve the predictive power of the model.

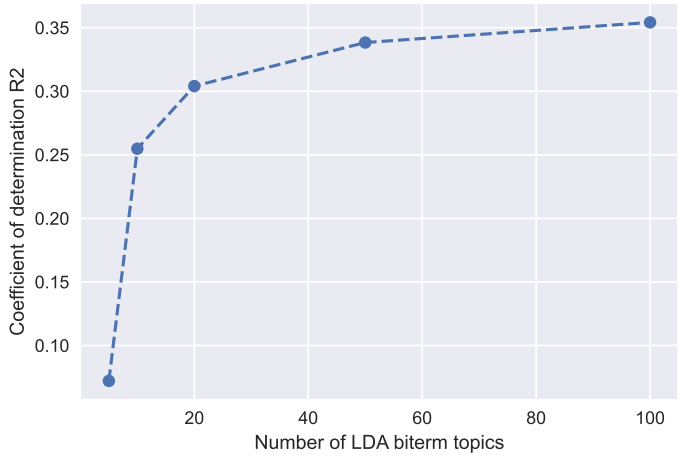

Figure S19: Coefficient of determination R-squared (R^2) for predicting the logged minimum hourly wage in job postings, comparing different numbers of LDA latent topics ($k \in 5, 10, 20, 50, 100$, shown on the x-axis). All models correspond to equation (1) and performance is evaluated using 5-fold cross-validation.

We then assess the robustness of our conclusions regarding the respective contributions of skill profiles and occupations to different values of k. Figure S20 shows the increase in R^2 when the topic loads for a given k are added to the model including occupational variables. Consistent with Figure S19, $k = 5$ provides only a limited improvement in model fit. In contrast, a larger number

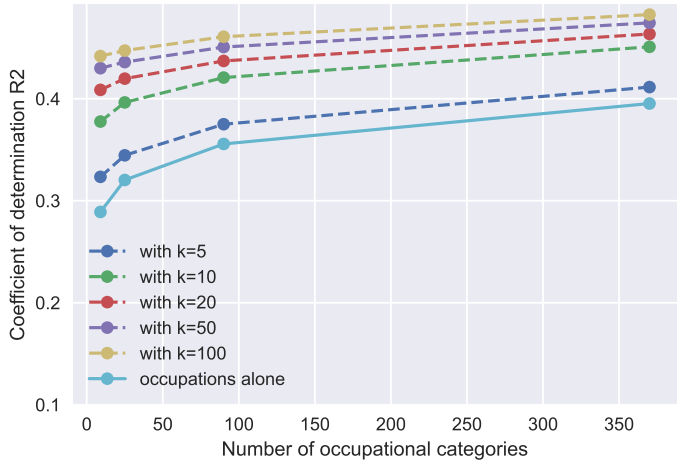

Figure S20: Coefficient of determination (R^2) for predicting the logged minimum hourly wage in job postings, comparing LDA topic models with different numbers of latent topics ($k \in 5, 10, 20, 50, 100$). The horizontal axis reports the number of occupational categories included as controls ($s \in 9, 25, 90, 369$). All models correspond to equation (3) and performance is evaluated using 5-fold cross-validation.

of topics substantially increases R^2 at all levels of the occupational classification. Also in line with Figure S19, the predictive gain is smaller between $k = 50$ and $k = 100$, indicating that increasing the granularity beyond 100 topics yields only marginal improvements. Overall, although higher predictive power could be achieved with more granular skill profiles, the patterns observed for $k = 19$ remain valid for $k \geq 10$.

References

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei (2009). "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pp. 288–296.

DiMaggio, P., M. Nag, and D. Blei (2013). "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding". en. In: *Poetics* 41.6, pp. 570–606. DOI: 10.1016/j.poetic.2013.08.004.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). "A Kernel Two-sample Test". In: *The Journal of Machine Learning Research* 13.1, pp. 723–773. DOI: 10.5555/2188385.2188410.

Muandet, K., K. Fukumizu, B. Sriperumbudur, and B. Schölkopf (2017). "Kernel Mean Embedding of Distributions: A Review and Beyond". en. In: *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141. DOI: 10.1561/2200000060.

Röder, M., A. Both, and A. Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". en. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399–408. DOI: 10.1145/2684822.2685324.

Rouwendal, H. J. and S. Koster (2025). "Does it take extra skills to work in a large city?" en. In: *Regional Science and Urban Economics* 112, p. 104094. DOI: 10.1016/j.regsciurbeco.2025.104094.

Schrab, A., I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton (2023). "MMD Aggregated Two-Sample Test". In: *Journal of Machine Learning Research* 24. Publisher: arXiv Version Number: 4, pp. 1–81. DOI: 10.48550/ARXIV.2110.15073.

Shmueli, G. (2010). "To Explain or to Predict?" In: *Statistical Science* 25.3. DOI: 10.1214/10-STS330.

Verhagen, M. D. (2022). "A Pragmatist's Guide to Using Prediction in the Social Sciences". en. In: *Socius: Sociological Research for a Dynamic World* 8, p. 23780231221081702. DOI: 10.1177/23780231221081702.

Yan, X., J. Guo, Y. Lan, and X. Cheng (2013). "A Biterm Topic Model for Short Texts". en. In: *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445–1456. DOI: 10.1145/2488388.2488514.