Migration, innovation and productivity in UK firms

Max Nathan, Tom Kemeny, Ceren Ozgen, Guido Pialli, Anna Rosso, Mateo Seré & Anna Valero

max.nathan@ucl.ac.uk
max-nathan.github.io

January 2026

Work in progress, please do not share

Comments welcome!

Economic and Social Research Council

[Related literatures]

Summary

- **Q:** Does firm migrant share affect firm performance? How?
- **Why this matters:**
  - Big public debate about economic and social impacts of migration
  - Productivity is an important margin; international evidence is pretty mixed
  - Knowledge gaps: mechanisms, and in countries (like UK) lacking register data
- **What we do:** explore using new worker-firm panel from rich web data + microdata, and an IV based on historic Census info
- **Findings so far:**
  - Positive but non-significant link between firm TFP and migrant share
  - Migrants carry a) distinctive and b) more complex skills, compared to UK-born
  - Migrant tech specialisation improves firm innovation

A very simple framework

- ***A priori*, firm migrant share ~ productivity is ambiguous**

- **Positive mechanisms** – selection on human capital; improved task specialisation; gains from cognitive diversity (~ birth country)

- **Negative / confounders** – discrimination, workplace frictions, ‘McKinsey multiculturalism’

- **Heterogeneity** – plausibly, bigger effects in more ‘knowledge-intensive’ sectors, workflows, workers, big cities
  - Implies *average* effects may be small, b/c pooling across these dimensions

Data stack

- **Diffbot**: graph database of the global public web [more] [validation]
- As of 2025, 35.2m active firms, 221.5m workers worldwide
  - Sources incl: multiple public databases, media, social media platforms
  - Similar to: Revelio, Cognism, but larger and more high-dimensional
- Diffbot gives us rich info on **person profiles, education + career histories + worker skills**; and clean **linkage to firms**
  - Contains company identifiers (CRNs) from UK's open company register
  - Use CRNs to build the stack
- **OpenCorporates + Historical Orbis:** jobs, financials [more]
- **PATSTAT Global, Orbis IP:** patenting counts, quality [more]

Worker-firm panel

- **Frame: larger UK companies.** Defined by turnover / balance sheet / workforce thresholds. Have complete, audited accounts [more]
- **Approach:** Search in Diffbot using CRNs, retain matches [more]
- **Diagnostics:** share of workforce observed in Diffbot varies. Explore distribution, test for predictors of coverage ratio [more]

- **Sample build** runs as follows:
  - Starting sample: **55,187** medium/large companies active 2007-2023
  - **33,081** companies **(59.9%)** matched on CRN, >=1 employee in Diffbot in any year
  - Keep **11,233** companies with coverage ratios ≥ 0.25; ~800k workers
  - Merge + trim to Orbis => unbalanced panel of **9,007** companies + **535k** workers

[Fitting topics] [Selection test]

Worker skills data

- **Diffbot assigns 'skills' to workers** Clean typology of ~32k terms, from self-description, endorsements etc
- Use LDA to bundle skills => topics, then assign topics to workers [more]
- 454k workers, observed *once*

- **Framework**. Tasks ~ generalised problems, skills ~ capabilities to complete tasks (Dorn et al 2025)
- **Test**. In any SOC4 bin, skills have +ve link with years of experience. Mixed link to qualifications [more]

Source: Diffbot. 25-topic model based on 454k workers for which we observe Diffbot skills. Labelled by LLM and numbered from the prompt: *"Following the ISCO guidelines, order these topics according to their level of complexity."*

[Descriptives]

Company-level regression

- Panel dataframe: company *i*, sector *j*, area *a* and year *t*
- Overall link between firm migrant share and productivity:

$$Y_{ijat} = b1 + b2MIG_{it-1} + \mathbf{X}_{ijat-1} + \mathbf{X'}_{at-1} + I_{ija} + T_t + e \qquad (1)$$

  - Y_{ijat} is log estimated total factor productivity growth, per Parotta et al (2014) [more]
  - In extensions, Y is patent counts / quality
  - MIG_{it} is share migrants in the workforce. Country of birth ~ country of lowest observed education, per Jin et al 2025, Lee & Glennon 2023 and others [more]
  - $\mathbf{X}_{ijat}$ is a rich array of company-level controls
  - I_{ija} is a firm-industry-area fixed effect – no firms switch industry or location

Identification

- **Challenges: unobservables, simultaneity, reverse causation**

- Run robustness tests on the company panel results
  - **Lee Bounds-type test:** re-build MIG_{it} assuming all unobserved are migrants / natives
  - **Placebo tests:** 1) regress MIG_{it} on 1-10 period lags of Y; 2) randomise MIG_{it}

- **Use instruments** based on historic name info, from 1881 UK Census
  - **Intuition:** deep persistence of names & name families => workers with 'historic UK' surnames more likely to be UK-born. Surnames unlikely to directly influence company-level outcomes, except through migrant/UK-born status
  - **Noisy:** mis-ascribes UK-born descendants of post-1881 arrivals; mis-ascribes Anglophone migrants who happen to have 'historic UK' surnames
  - **Shift-share IV**: 1 – (historic UK surnames $_{i07}$ * historic UK_t) ~> MIG_{it} [more]

[Sensitivity] [Placebos]

Positive but non-significant MIG ~ TFP effect

TFP growth	(1)	(2)	(3)	(4)	(5)
L. Migrant share	0.0178	0.0236	0.0225	0.0687	0.345
	(0.0176)	(0.0182)	(0.0183)	(0.0575)	(1.098)
Controls	No	Yes	Yes	Yes	Yes
Year-Region FE	Yes	Yes	Yes	Yes	Yes
Industry FE	No	No	Yes	-	-
Firm FE	No	No	No	Yes	Yes
N firms	6373	6373	6373	6373	6373
Observations	46906	46906	46906	46906	46906
R^2	0.00786	0.0105	0.0113	0.177	0.177
Pr(migrant share)					-0.0021***
First stage F					10.71

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT.
Regressions are run on an unbalanced panel of 6,373 companies. Year FE = 16 years. Region FE = 12 UK regions. Industry FE = 11 1-digit SIC07 bins. Controls: share graduates, share female, share non-execs, log(workforce mean age); foreign subsidiary dummy, number of subsidiaries, weighted patent count, citation count, log(revenues) + Diffbot coverage rate. All explanatory variables are 1-period lags. Standard errors in parentheses clustered at the firm level. * p<0.1 ** p<0.05 *** p<0.01.

Extensions

- **We can't rule in/out a migrant ~ TFP link** (cf Hall & Manning 2025)
- We test for heterogeneity at company level. We find positive, but not significant, coefficients on migrant shares
  - **Industry variation:** MIG * mf / services / high-tech
  - **Occupational structure:** MIG * skills / STEM intensity
  - **Workforce:** skilled migrant or migrant manager shares
  - **Workforce:** split migrants / natives by qualifications and experience
  - **Cultural proximity.** share Commonwealth origin migrants
  - **Cultural diversity**. Use birth country and linguistic fractionalisation indices. Possible U-shape from linguistic diversity to TFP [build] [results]

[Descriptives]

Worker-level analysis

- **Cognitive diversity** – build on Page (2007), using Diffbot skills
  - Compare Diffbot skills for migrants vs. UK-born: control for observables, SOC4 FE
  - Instrument for migrant status using historic UK names info
  - Migrants and UK-born carry distinct skills [more]
- **Selection on skills** – using LLM-assigned skills complexity
  - Compare migrants vs. UK born in same role *k*, using IV specification as above
  - Overall, and especially in tech/STEM roles, migrants carry more complex skills [more]
- **Specialisation** – adapt framework from Mayda et al (2022)
  - Migrant vs UK-born specialisation in managerial vs technical roles
  - Test migrant specialisation ~ firm-level innovation, productivity

Complex skills

- **We compare ISCO-complexity of Diffbot skills for migrants vs. natives doing the same types of job**. We define job types as bundles of SOC4 occupations. We estimate for worker *i*, job type *k*, SOC4 bin *o*:

$$Y_{io} = a + b1\{k\}\_MIGRANT_{io} + \mathbf{X}c_i + e_{io} \qquad (2)$$

  - Y is the ISCO-complexity of worker *i*'s dominant topic (ranked 1-25, where 25 is the most complex)
  - MIGRANT is a dummy for migrants doing *k* = all, tech, STEM or managerial roles. The omitted category is natives doing *k* roles
  - **X** is individual-level controls, including qualifications and experience; *o* indexes SOC4 bins
  - We instrument for migrant status: $Pr(MIGRANT_i) = 1 – (\text{historic UK name})_i$

- **Significant *b1* ~>** migrants in *k* carry more complex skills than natives in *k*

Migrants carry more complex skills

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
migrant	0.623***		2.019***									
	(0.0256)		(0.0893)									
Predicted migrant, historic name		0.539***										
		(0.0238)										
tech_migrant				0.895***		1.483***						
				(0.0812)		(0.221)						
Predicted migrant in tech occupation					0.525***							
					(0.0783)							
stem_migrant							0.795***		1.201***			
							(0.0737)		(0.199)			
Predicted migrant in stem occupation								0.427***				
								(0.0709)				
manager_migrant										0.625***		2.634***
										(0.0466)		(0.176)
Predicted migrant in managerial occupation											0.669***	
											(0.0443)	
Observations	437630	437630	437630	37224	37224	37224	43532	43532	43532	138052	138052	138052
R^2	0.241	0.241	0.0562	0.215	0.214	0.00749	0.278	0.277	0.0102	0.190	0.191	-0.00775
Oster delta / migrant	5.230			3.396			3.125			2.006		
K-P under-identification test			24069.2			3513.7			4244.6			6091.4
K-P weak instrument test			27377.2			4251.6			5134.9			6919.7

Source: Diffbot. Sample = workers with observed skills and observables. Regression compares the ISCO-complexity of skills for migrants and natives doing the same jobs, versus migrants and natives doing any other jobs. All specifications include year dummies, SOC4 dummies and controls (female, foreign language speaker, graduate or above, degree type, years of experience). Robust standard errors in parentheses.

Selection on skills: sensitivity

	(1) OLS	(2) IV	(3) gpt_o3	(4) claude	(5) _2023	(6) _1123	(7) _20223	(8) _20123	(9) _20203	(10) _20193	(11) placebo
tech_migrant	0.895***	1.483***	1.604***	1.328***	1.477***	1.500***	1.530***	1.502***	1.456***	1.478***	-0.0636
	(0.0812)	(0.221)	(0.211)	(0.233)	(0.353)	(0.222)	(0.329)	(0.319)	(0.301)	(0.287)	(0.0919)
Observations	37224	37224	37224	37275	14216	36460	16105	17383	19312	21258	32812
R^2	0.215	0.00749	0.0100	0.00802	0.0106	0.00761	0.0111	0.0101	0.0101	0.0108	0.00802
Oster delta / migrant	3.396										
Under-ID test		3513.7	3513.7	3546.5	1391.2	3470.0	1602.8	1705.2	1907.9	2104.7	
Weak instrument test		4251.6	4251.6	4327.9	1693.4	4203.2	1957.8	2078.8	2332.3	2570.1	

Source: Diffbot. Cols 1-2 per main table. Cols 3-4 use alternative topic complexity classifiers. Cols 5-10 use alternative time windows. Col 11 fits a placebo test where we randomize topics across workers.

[Descriptives]

Specialisation

- Focus: specialisation in managerial vs technical roles. Adapt framework from Mayda et al (2022), Peri and Sparber (2011):
  - Skilled migrants / natives are imperfect substitutes
  - Natives have comparative advantage in managerial/language-intensive roles
  - We should observe migrants' relative specialisation in technical roles

- Define migrant/native ratio, e.g. for tech roles:

$$\text{MNRtech}_{it} = (\text{Migrant share tech})_{it} \,/\, (\text{Native share tech})_{it} \qquad (3)$$

- Then define migrant (relative) tech specialisation:

$$\text{Migrant tech specialisation}_{it} = (\text{MNRtech}_{it} \,/\, \text{MNRman}_{it}) \qquad (4)$$

[MNRs][TFP results]

Migrant tech specialisation ~ higher patenting

	(1) #patents	(2) #patents	(3) #patents	(4) # citations	(5) # citations	(6) # citations
L.migrant tech specialisation	0.0889*			-0.00110		
	(0.0471)			(0.0853)		
L.migrant stem specialisation		0.00712			-0.00251	
		(0.0162)			(0.0547)	
L.migrant management spec			-0.0301			0.173
			(0.0640)			(0.141)
N groups	222	221	182	166	166	135
Observations	1264	1271	974	954	962	743
Pseudo R^2	0.601	0.597	0.633	0.700	0.701	0.705

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT. Poisson pseudo-maximum likelihood estimators. Patents are weighted by applicants. Cites are lifetime citations. All regressions fit controls, year and area-SC4-firm FE. Other notes per main table.

The link is stronger for firms that already patent

A. Extensive margin	#patents	#patents	#cites	#cites
L.migrant tech role specialisation	0.00000544		0.000194	
	(0.000456)		(0.000388)	
L.migrant stem role specialisation		-0.0000480		0.000406
		(0.000499)		(0.000538)
N groups	2484	2429	2484	2429
Observations	15080	14743	15080	14743
R^2	0.486	0.485	0.446	0.445

B. Intensive margin	#patents	#patents	#cites	#cites
L.migrant tech role specialisation	0.0981**		-0.0642	
	(0.0463)		(0.0906)	
L.migrant stem role specialisation		0.000213		-0.0322
		(0.0148)		(0.0596)
N groups	133	134	117	117
Observations	463	467	422	424
Pseudo R^2	0.596	0.591	0.754	0.753

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT. Panel A fits LPM regressions. Panel B fits Poisson pseudo-maximum likelihood estimators. Patents are weighted by applicants. Cites are lifetime citations. All regressions fit controls, area, region, year and firm FE. Other notes per main table.

Reflections

- **We explore effects of migration on UK firm performance** using **a novel worker-firm data stack**. Main findings:
  - Migrant ~ firm TFP effect is positive but small, not significant
  - Explore within-industry and within-firm variation, using worker-level data
  - Migrants carry more complex skills than UK-born doing the same job, especially in tech and STEM roles
  - Migrant carry distinctive skills, compared to UK-born doing the same job
  - Migrant tech specialisation is associated with higher firm innovation
- Next steps:
  - **Identification:** extend IV to specialisation analysis
  - **Firms vs. places:** demographics in surrounding region / city-region

Thanks!

max.nathan@ucl.ac.uk
max-nathan.github.io

Appendix

[back]

Related literatures

- **Migration and firm productivity:** Closest: Exadaktylos et al 2024 (+ve effect for mf), Hall and Manning 2024 (zero effect), Ottaviano et al 2018 (+ve effect for services). Peri 2012, Paserman 2013, Garnero et al 2014, Parrotta et al 2014, Mitaritonna et al 2017, Fabling et al 2022, Nam & Portes 2023, Ek 2024
- **Skilled migration, innovation:** Page 2007, Hunt and Gauthier-Loiselle 2010, Parrotta et al 2014, Nathan 2015, Kim and Starks 2016, Ferruci and Lissoni 2019, Dale-Olsen and Finserras 2020, Hofstra et al 2020, Mayda et al 2022
- **Migrant diversity and urban economic performance:** Ottaviano and Peri 2006 and 2007, Kerr and Lincoln 2010, Nathan and Lee 2013, Kemeny 2014, Nathan 2015, Peri et al 2015, Kemeny and Cooke 2018 and others
- **Linking web and company data to make worker-firm panels:** Tambe et al 2020, Fedyk & Hodson 2022, Rock 2022, Breithaupt et al 2023, Babina et al 2023, Jeffers 2023, Lee & Glennon 2023, Gagliardi et al 2024, Dahlke et al 2025, Dorn et al 2025

[back]

Diffbot

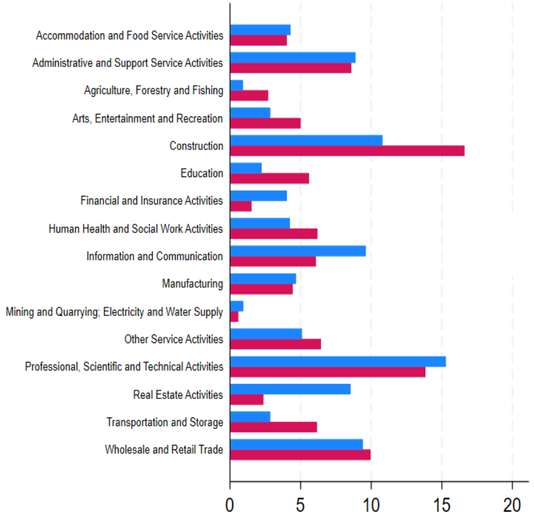
- **Diffbot**: commercial graph database of the global public web
- We use Diffbot to give us person profiles, education / career histories, and additional data on UK companies

- Key features of Diffbot's platform
  - Scrapes internet => feature extraction => supervised learning to link entities
  - Clear ontology of organisations, people, locations, etc. and their characteristics and histories. Also covers news media and news events
  - Very large, rich firm and worker data
  - Globally, 30.2m firms, 199m individuals in 2022
  - Transparent sourcing: for the UK, draws on multiple public databases (including UK companies House registry), media and social media platforms

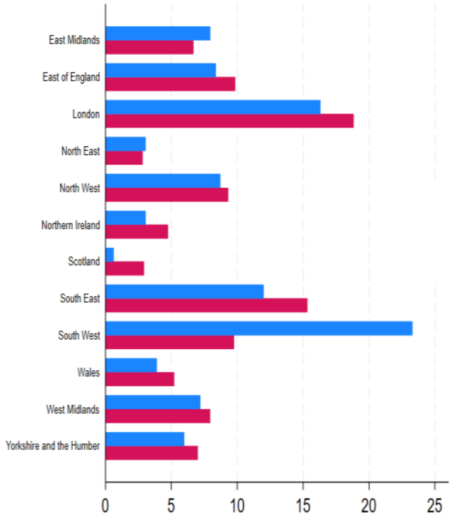
[back]

Diffbot validation

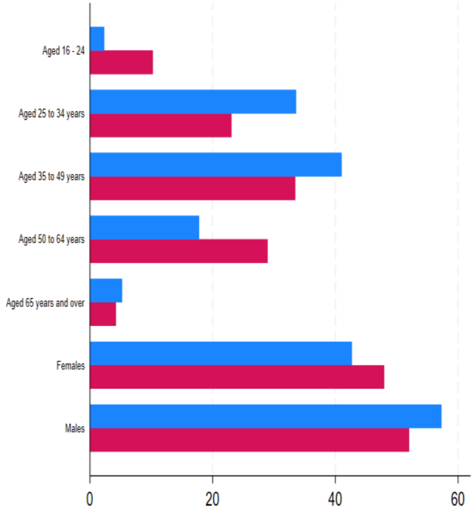
Industries



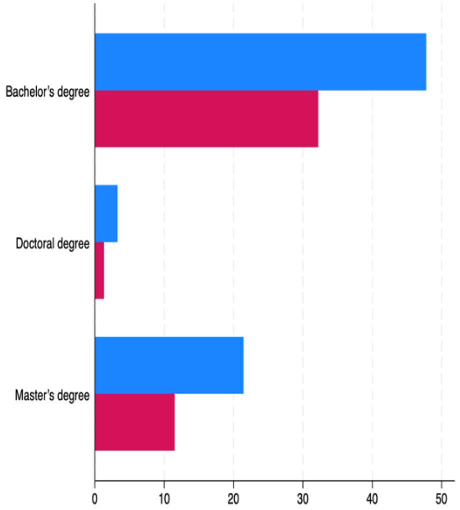
Regions



Age, gender



Education



Blue = Diffbot, Red = BPE

Left: NACE1 shares for UK-based organisations active in Diffbot (end-2021) and firms in ONS Business Population Estimates (start of 2022)

Right: NUTS 1 shares (R) for UK-based organisations active in Diffbot (end-2021) and firms in ONS Business Population Estimates (start of 2022)

Blue = Diffbot, Red = Census / APS

Left: age group and gender shares for individuals in Diffbot employed at the end of 2021 and individuals in the 2021 Census (England and Wales)

Right: Bachelor, masters and PhD shares for individuals in Diffbot employed at the end of 2019, and 2019 ILO data for the UK (taken from the EULFS).

Diffbot data is raw, no cleaning

[back]

Orbis

- We apply the definition of medium-large company in Orbis Historical for the period 2007-2020, using data from unconsolidated balance sheets only. We obtain a sample of **55,775** companies. We then match this sample of companies to Companies House (through OpenCorporates), which leaves us with a sample of **55,187** companies observed for 2007-2020.

- We follow the cleaning procedures for Orbis data documented in Kalemli-Ozcan et al. (2019) and De Loecker, Obermeier and Van Reenen (2022).
- We keep only firm-year observations for which financial variables are expressed in GBP pounds.
- We use the account closing date to determine the calendar year. If the closing date is after or on June 1st, we assign it to the current year. If it is before June 1st, we assign it to the previous year.
- At this stage, Orbis may contain multiple annual observations for some firms. We design a routine of sequential steps to remove firm-year observations duplicates, similar to De Loecker, Obermeier and Van Reenen (2022):
    1. We keep the annual report values when both the annual report and local registry filing are present, and the annual report values are non-missing.
    2. When annual report and local registry filing values are not the same (and both are non-missing), we prefer annual report values.
    3. When annual report values are missing, and local registry filings are non-missing, we keep local registry filings.
    4. After the selection above, we prefer consolidated accounts to unconsolidated accounts.
    5. We remove remaining duplicates by taking the observations with fewer missing values for the number of employees, EBITDA and costs of employees.

back

Patent variables

- We initially match patent data to firms using Orbis IP through the BvD identification number. In tests we find the global matching rate > PATSTAT through HAN id.
- In later versions of the build we will complement Orbis IP data with PATSTAT Global, for UKIPO and USP-JP-EPO offices.
- We follow standard practices in the literature (e.g. Arora et al., 2021, DISCERN; Bloom et al., 2020):
  - We use patents applied to both local (UK) and worldwide patent offices (USPTO, EPO…);
  - We reconstruct patent families using INPADOC in Patstat (to account for the fact that the same invention can be protected by multiple patents);
  - Patents are assigned to firms based on the fractional counting method (if a patent has two applicants, we assign ½ to each);
  - We assign patents to the priority year of application, considered the closest date to the original invention.

[back]

Sample quality

- Check *Pr*(CRN match). Not explained by observables [more]

- Look at workforce coverage by year. For company *i*, year *t*:

$$C_{it} = (\text{Workers in Diffbot})_{it} \,/\, (\text{workers in OC-Orbis})_{it} \qquad (1)$$

- Formally test for selection (adapting Fedyk & Hodson 2022). We find that observables (= controls) explain ~78% of coverage ratio [more]

$$C_{ijt} = F(\text{observables}_{it}, \text{sector}_{jt}) \qquad (2)$$

- **From the 33,081 CRN matches we select 11,233 firms** with coverage ratios ≥ 0.25 [more]. We observe ~800k workers between 2007-2023

back

CRN matching test

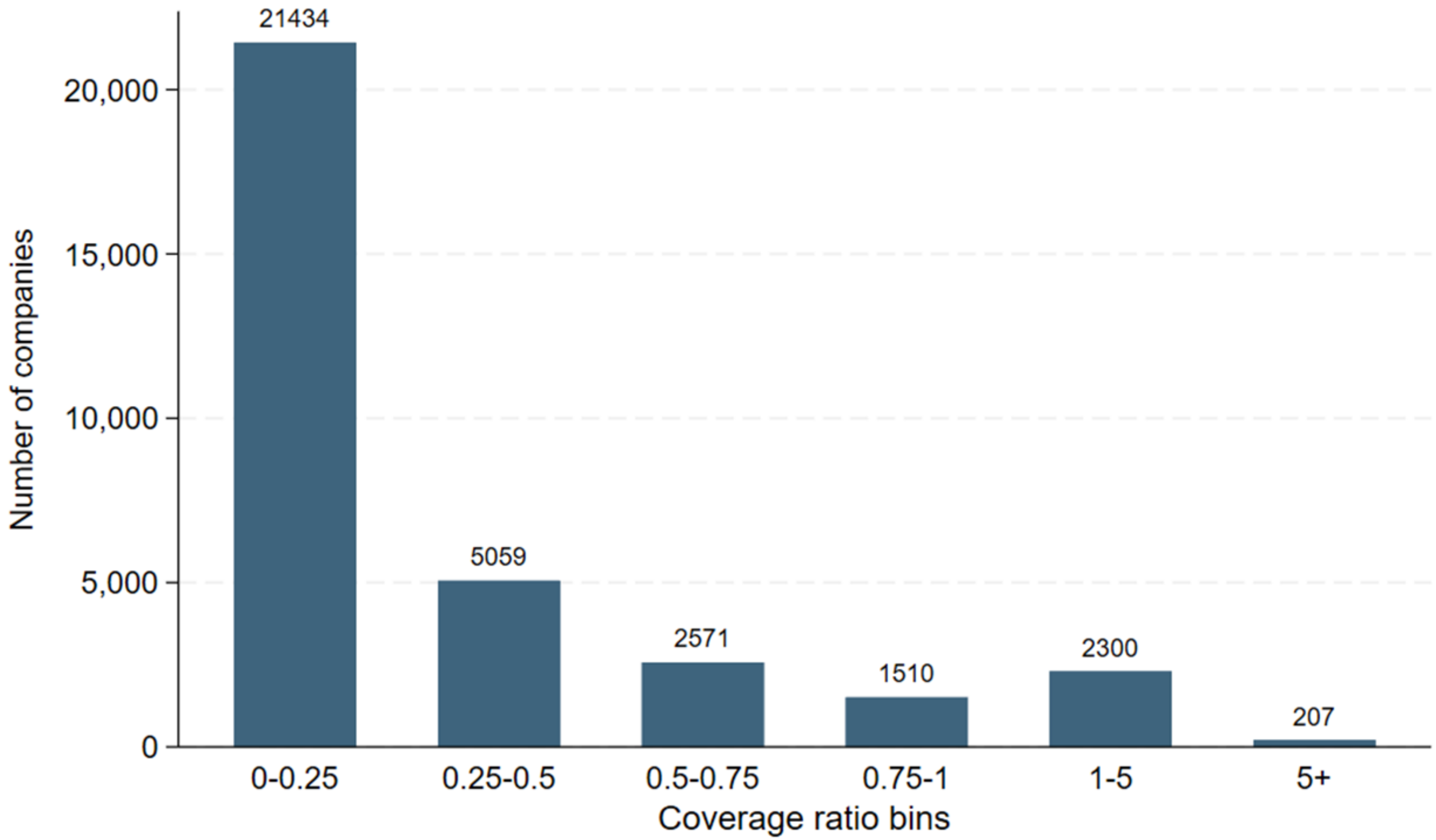
- **We run an LPM to determine predictors of CRN matches**. Predictors are:
  - **Dissolved**: A dummy == 1 for whether the company is dissolved [Source: CH]
  - **Average firm size:** Average firm size for each company's sample period (log of the number of the employees)
  - **CH Incorporation year:** Incorporation year
  - **Female board share:** Share of all-time female board members over the all-time total number of board members
  - **Non-UK board share:** Share of all-time non-UK board members
  - **London-based:** A dummy == 1 for whether the company is based in London (Orbis' address)
  - **SIC1 sector dummies** (agriculture omitted category)
- **Observables only explain 9% of the probability of a CRN match, suggesting that matching is essentially random.** We find that the probability to find a firm in Diffbot:
  - Decreases if the firm is dissolved
  - Increases as the firm size increases
  - Decreases as the firm gets older
  - Increases as the female board share increases
  - Decreases as the non-UK board share increases.

[back](#)

Coverage regressions

	(1)	(2)	(3)
Dependent variable:		Coverage ratio	
Firm age	-0.0687***	-0.0485***	-0.0558*
	(0.00834)	(0.0184)	(0.0286)
Foreign subsidiaries	0.423	-0.0575	-0.0524
	(0.264)	(0.0379)	(0.0400)
GUO	-0.0252***	0.0172	0.0271*
	(0.00932)	(0.0201)	(0.0150)
Female board share	0.156***		
	(0.0363)		
Non-UK board share	-0.00516		
	(0.0454)		
Year FE	Yes	Yes	-
Region FE	Yes	-	-
Industry FE	Yes	-	-
Firm FE	-	Yes	Yes
Industry-Year FE	-	-	Yes
Observations	154,125	154,579	148,707
R-squared	0.00995	0.775	0.782

Coverage by bins



Sample = 33,081 companies

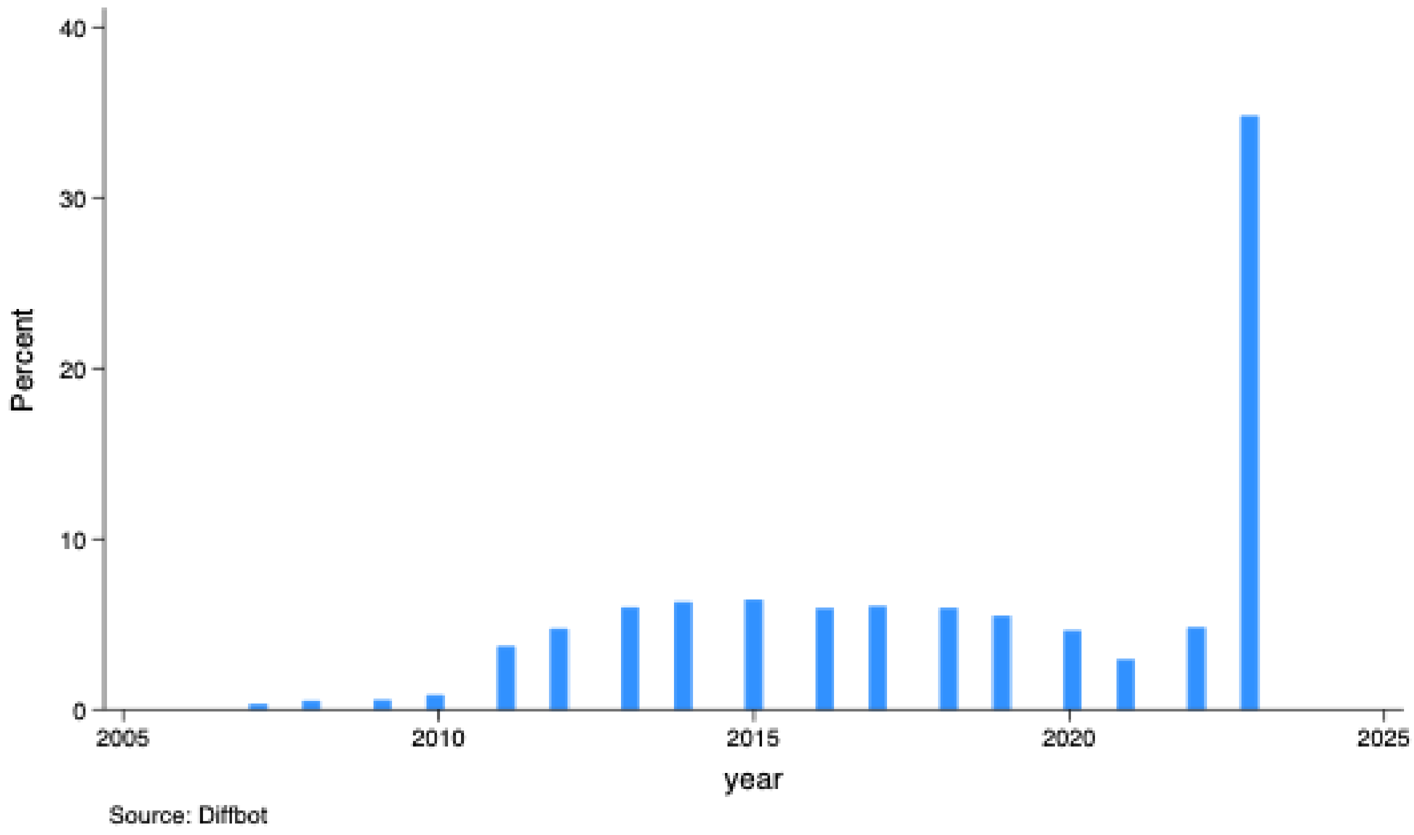
[back]

Skills data

- **TLDR;**
  - Diffbot observes skills for most workers in our data.
  - Raw info comes largely from online testimonials and job descriptions. Diffbot uses a typology of 32k skills to filter this (filters = mentions on wikidata, frequency of mentions).
  - Diffbot skills are observed once per worker – there's no individual-level time variation.
  - Workers move in and out of the firms in our sample.
  - This means that we observe worker skills in a range of years, depending on workers' last year of employment in our firms.
  - Worker moves mean that firms' stock of skills is time-varying, even though workers are only observed once in the data.
- **Coverage rate** – we observe skills for 82.8% of workers in our data.
- **Coverage by year** – around 1/3 of workers are observed in 2023; almost the rest are observed between 2011 – 2022

[back]

Diffbot coverage rate over time



Notes: observed year of skills for workers in our sample. There are 753,188 workers in our WP1-3 sample, of whom 655,504 have observed skills.

[back]

Selection into skills

- Even though we observe Diffbot skills for >80% of workers, there could be selection into having skills.
- Migrants, specifically, might need to demonstrate skills more than natives.
- We run two checks to compare workers in our sample with/out observed skills:
  - For workers with/out observed skills, we t-test observables
  - We run a worker-level LPM, regressing the probability of having observed skills on a set of dummies for worker-level observables
- Both show significant differences between workers with/out skills, but a) the effect of specific observables is always small and b) in the selection regression, overall model fit is very low
- Ranking betas by size from the LPM regression: foreign language > female > tech occupation > manager > humanities degree > STEM degree > migrant. This suggests that migrant selection is relatively unimportant as a predictor of having observed skills.
- Oster tests on foreign language, female and migrant give deltas of ~ 8, 5 and 2.

[back]

Depvar = has skills	(1)	(2)
migrant	0.0226***	0.0204***
	(0.00110)	(0.00120)
foreign_language	0.0826***	0.0725***
	(0.00123)	(0.00128)
age		-0.00197***
		(0.0000339)
female		0.00540***
		(0.000995)
stem_degree		0.0209***
		(0.00113)
economics_degree		0.0173***
		(0.00111)
humanities_and_arts_degree		0.0290***
		(0.00127)
oxbridge_degree		-0.0160***
		(0.00265)
russell_degree		0.000136
		(0.00129)
graduate		0.00326**
		(0.00142)
post_graduate		0.0135***
		(0.00177)
phd		0.0350***
		(0.00313)
tech_occupation		0.0367***
		(0.00157)
managerial_occupation		0.0299***
		(0.000994)
worker_experience		0.00936***
		(0.0000715)
Observations	538872	515771
R^2	0.0187	0.0688

Source: Diffbot. Sample = 753,188 workers. 655,504 have observed skills.

These regressions run a worker-level LPM, regressing the probability of having observed skills on a set of dummies for worker-level observables (plus age and years of experience). Specifications also include year dummies, to take account of year skills are observed.

Standard errors in parentheses are clustered at the individual level.

[back]

Benchmarking

Variables	(1)	(2)	(3)	(4)	(5)
(mean) top1_topic_gpt_4o	1.000				
(mean) graduate	-0.360***	1.000			
(mean) post_graduate	0.047	0.164***	1.000		
(mean) phd	0.210***	-0.136***	0.419***	1.000	
(mean) worker_experience	0.350***	-0.068	-0.008	-0.003	1.000

Variables	(1)	(2)	(3)	(4)	(5)
(mean) high_skill	1.000				
(mean) graduate	-0.236***	1.000			
(mean) post_graduate	0.066	0.164***	1.000		
(mean) phd	0.234***	-0.136***	0.419***	1.000	
(mean) worker_experience	0.243***	-0.068	-0.008	-0.003	1.000

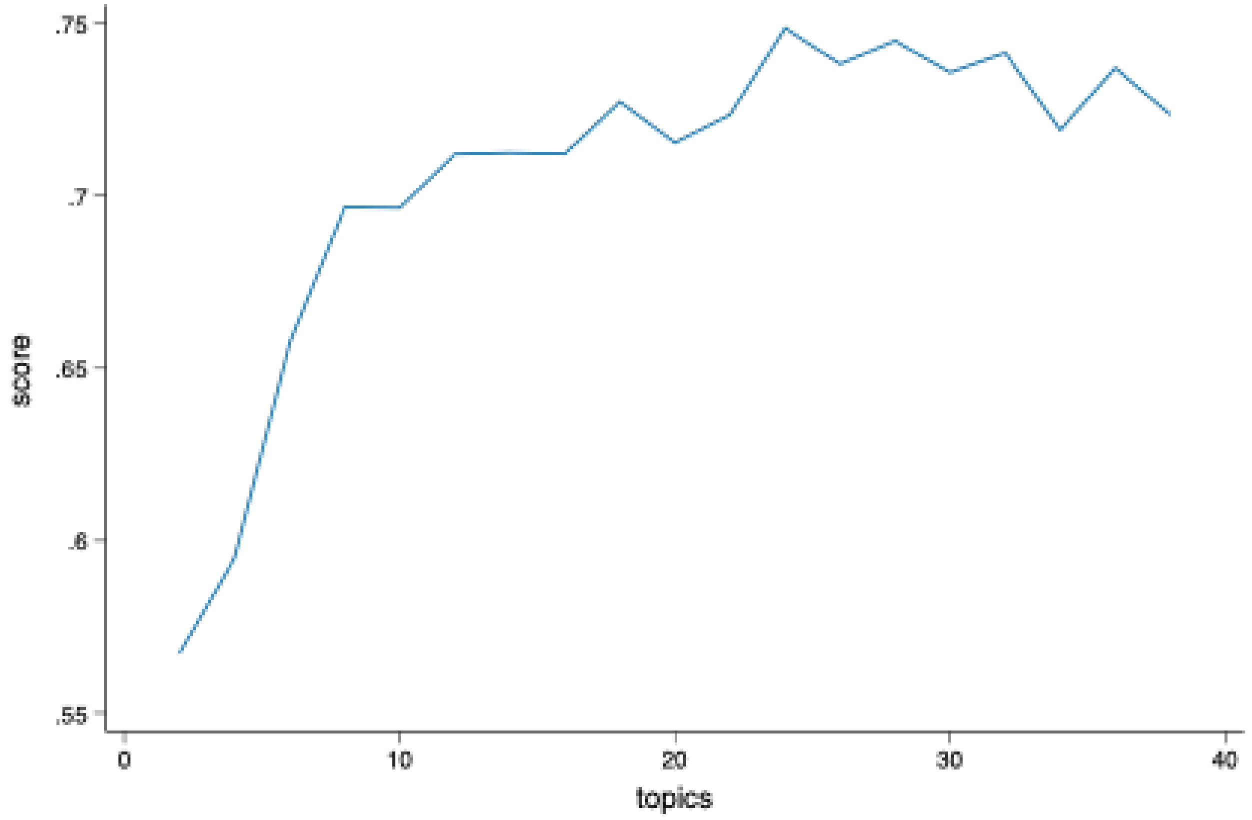
[back]

Skills processing workflow

- Input = 800,511 rows, 655,504 with skills
- Pre-processing: normalize common phrases, remove stopwords => tokenise, build bigrams, lemmatize => Build corpus with Gensim Dictionary (filtered: min 100 docs, max 95%)
- Explore model fit across range of #topics => optimised for 25
- Train a 25-topic LDA model using MALLET, assign topics to documents, and export topic word list
- Extract top 3 topics (by probabilities) for each individual
- Align topic IDs with manual labelings and ISCO complexity (GPT-4o, Claude, GPT-o3). Compare sensitivity of rankings

back

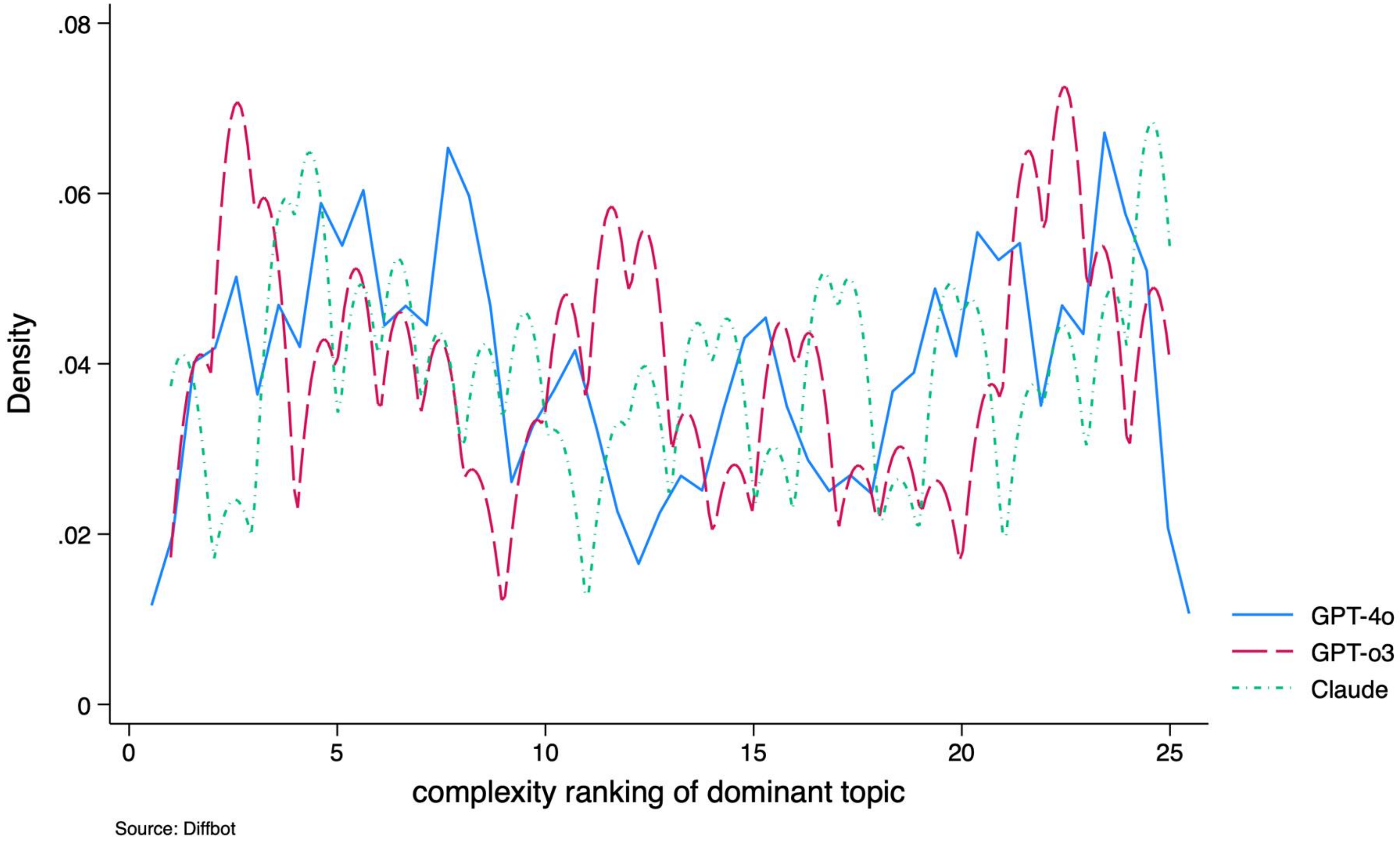
LDA topic optimisation



Source: Diffbot.
Y-axis indicates goodness of fit. X-axis measures number of topics. Sample is 454k workers for which we observe Diffbot skills.

[back]

LLM ISCO-complexity checks



Notes: There are 753,188 workers in our WP1-3 sample, of whom 655,504 have observed skills.

Coverage summary stats

	Median	Mean	25^{th} p	75^{th} p	Min	Max	Obs
Panel A: Coverage ratio: #Diffbot employees/#Orbis employees							
Time-vary coverage ratio	0.12	0.36	0.03	0.34	0	197	192,989
Average coverage ratio	0.14	0.44	0.03	0.39	0	174.5	33,081
Panel B: Coverage ratio: #Diffbot employees/#Orbis employees excluding obs. with coverage ratio > 1							
Time-var coverage ratio	0.11	0.20	0.03	0.28	0	1	181,875
Average coverage ratio	0.12	0.21	0.03	0.30	0	1	30,574

The table shows descriptive statistics for the time-varying coverage ratio and the average coverage ratio. Unique N = 33,081 companies (Panel A), 30,574 (Panel B)

[back]

Shift-share IV

- We apply a shift-share IV, building on the design in Balsmeier et al (2025)
- We use 1881 Census data for England and Wales, which contains surname information at individual level, geocoded to parishes.
- Our instrument combines the initial share of a specific surname in a firm with the evolution of that surname over years:

$$\widehat{MigShare}_{ft} = 1 - \left(\sum_{s=1}^{S} \frac{UK_{s,f,t_0}}{N_{f,t_0}} \times (UK_{s,t} - UK_{s,t_0}) \right) \quad (4)$$

where Uk_{s,f,t_0} is the number of workers with historic UK surnames *s* in firm *f* in year *t0*, while N_{f,t_0} is the number of unique surnames in firm *f* in year *t0*. The term $(UK_{s,t} - UK_{s,t_0})$ is the relative change of individuals with surname *s* between year *t* and year t_0

[back]

Measuring migrant status

- Birth country is not observed in Diffbot. Following Jin et al (2025), Lee and Glennon (2023) etc., we proxy using location of level of lowest observed education => where people went to school or university
  - Foreign students in UK > UK abroad => likely understates true migration. Secondary evidence supports this (HESA 2024)
  - We run a validation exercise on a UCL department: 35% response rate to online survey, 89.7% correct attribution, lower bound on true migrant status
- **Simple proxies:** share migrant in firm *i*, year *t*
- **Fractionalisation indices:** scaled 0 ~ 1. Diversity = more + more evenly-sized groups
  - Migrant diversity: 194 birth countries => 14 country blocs

back

Company summary statistics

Variable	A. All years			B. 2007			C. 2023		
	count	mean	sd	count	mean	sd	count	mean	sd
Log TFP (Olley-Pakes)	63851	3.866	1.610	232	3.557	1.511	4258	3.931	1.668
Number of patents	70911	0.094	1.429	262	0.075	0.799	4750	0.000	0.000
Number of citations (all-time)	70911	0.475	11.192	262	0.691	7.101	4750	0.000	0.000
Share migrant workers	70911	0.121	0.147	262	0.123	0.161	4750	0.127	0.143
Share of workers with a college or higher degree	70911	0.501	0.245	262	0.513	0.244	4750	0.510	0.241
Share migrants with degree or above	70911	0.105	0.140	262	0.104	0.146	4750	0.110	0.137
Share UK workers with degree or above	70911	0.389	0.208	262	0.401	0.215	4750	0.392	0.201
Share of workers in tech occupations	70911	0.074	0.095	262	0.081	0.108	4750	0.077	0.093
Share of workers in stem occupations	70911	0.082	0.110	262	0.089	0.123	4750	0.086	0.109
Share of workers in managerial occupations	70911	0.374	0.200	262	0.401	0.214	4750	0.366	0.194
Share of migrant workers in tech occupations	70911	0.010	0.032	262	0.012	0.044	4750	0.012	0.036
Share of UK workers in tech occupations	70911	0.038	0.061	262	0.038	0.057	4750	0.039	0.057
Share of migrant workers in stem occupations	70911	0.012	0.034	262	0.014	0.045	4750	0.014	0.038
Share of UK workers in stem occupations	70911	0.044	0.072	262	0.045	0.069	4750	0.045	0.067
Share of migrants in managerial occupations	70911	0.040	0.079	262	0.040	0.077	4750	0.041	0.077
Share of UK workers in managerial occupations	70911	0.185	0.140	262	0.205	0.145	4750	0.183	0.131
Workforce average age	69518	43.639	5.832	257	49.241	5.388	4668	41.950	5.582
Share of females	70911	0.334	0.208	262	0.323	0.238	4750	0.341	0.201
Number of employees	61519	89.664	116.229	200	79.855	75.386	3927	105.385	126.395
Firm age	70911	6.006	3.515	262	1.000	0.000	4750	11.335	2.450
Share of workers in non-executive occupations	70911	0.022	0.061	262	0.020	0.061	4750	0.022	0.058
Company has foreign subsidiaries	70911	0.002	0.042	262	0.000	0.000	4750	0.000	0.000
Number of subsidiaries	70911	0.807	2.946	262	1.115	3.610	4750	0.000	0.000
Log firm revenues	68857	16.700	1.186	245	16.659	1.325	4518	16.938	1.229
Observations	*70911*			*262*			*4750*		

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

[back]

Sensitivity checks

- **Results robust to the following checks** [more]
  - Alternative outcomes (log GVA, log revenue / worker)
  - Alternate fixed effects (region*year, sector*region, sector*region*year)
  - Alternate sample period (2010-2023)

  - Subsamples with higher coverage rates (cap at 1, ≥ 0.5, ≥ 0.75)
  - Penalising observations with lower coverage rates

  - Removing TFP and migrant share outliers (top 1%) [more]

  - Check whether firm characteristics vary by coverage rate bins [more]

back

Sensitivity checks

	(1) main	(2) gva	(3) revprod	(4) _kt	(5) _kj	(6) _jt	(7) _jkt	(8) t2010	(9) cr_1	(10) cr5_1	(11) cr75_1	(12) wgt
Migrant share	0.0234	0.0425	0.0430	0.0274	0.0234	0.00757	0.0415	0.0177	-0.0173	0.0358	-0.0352	-0.0420
	(0.0840)	(0.107)	(0.0841)	(0.0839)	(0.0858)	(0.0922)	(0.116)	(0.0848)	(0.0859)	(0.114)	(0.216)	(0.167)
N groups	6141	6140	6133	6140	6141	5991	4996	6089	5752	2848	1053	6070
Observations	44996	44992	44913	44989	44996	43649	35336	44708	40100	16349	4308	44183
R^2	0.942	0.788	0.907	0.942	0.942	0.947	0.951	0.942	0.951	0.943	0.944	0.926

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT. All regressions fit controls, area, region, year and firm FE.
Col 1 fits our main result. Cols 2 and 3 use log GVA and log revenue/worker as the outcome. Col 4 adds region*year FE. Col 5 adds 4-digit sector*region FE. Col 6 adds sector*year FE. Col 7 adds sector*region*year FE. Col 8 restricts the sample to 2010-2023, where obs are less sparse. Col 9 restricts to obs where the coverage ratio <=1. Col 10 conditions on coverage ratios between 0.5 and 1. Col 11 conditions on coverage ratios between 0.75 and 1. Column 12 weighs observations by coverage rate. Other notes per main table.

back

Outliers

Log TFP	(1) Main	(2) Excl. 99th perc. sh. migrant	(3) Excl. 99th perc. TFP	(4) Excl. 99th perc. TFP and mig.	(5) Excl. MVOs
Migrant share	0.0234	-0.0362	0.0234	-0.0262	0.0117
	(0.0840)	(0.0827)	(0.0840)	(0.0821)	(0.0850)
N groups	6141	6102	6141	6060	6123
Observations	44996	44681	44996	44356	44824
R^2	0.942	0.943	0.942	0.938	0.942

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT. All regressions fit controls, area, region, year and firm FE. Col 5 fits excludes multi-variate outliers as defined by Bacon Stata package. Other notes per main table.

back

Stats by coverage rate

	Coverage rate: 0.25-0.5				Coverage rate: 0.5-0.75				Coverage rate: > 0.75			
	Obs	Mean	s.d.	Median	Obs	Mean	s.d.	Median	Obs	Mean	s.d.	Median
TFP (Olley and Pakes)	24141	3.91	1.59	4.35	12243	3.72	1.45	3.59	9395	3.50	1.43	3.21
Share migrant (t-1)	24141	0.098	0.12	0.062	12243	0.13	0.13	0.088	9395	0.16	0.15	0.11
Average workforce age (t-1)	24141	44.1	5.59	43.6	12243	43.3	4.96	42.9	9395	42.6	4.98	42.3
Share of college (t-1)	24141	0.45	0.22	0.43	12243	0.54	0.22	0.55	9395	0.57	0.22	0.58
Share of female (t-1)	24141	0.32	0.19	0.29	12243	0.35	0.19	0.33	9395	0.36	0.19	0.34
Number of Orbis empl (t-1)	24141	108.0	123.1	71	12243	86.8	87.0	61	9395	64.4	76.5	47
Firm age (t-1)	24141	6.06	3.19	6	12243	5.84	3.14	6	9395	5.87	3.20	6
Share of non-exec workers (t-1)	24141	0.018	0.045	0	12243	0.019	0.044	0	9395	0.020	0.050	0
Share of foreign subs. (t-1)	24141	0.0013	0.036	0	12243	0.0015	0.038	0	9395	0.0027	0.052	0
Number of subs. (t-1)	24141	0.84	2.90	0	12243	0.87	2.49	0	9395	0.80	2.09	0
Number of patents (t-1)	24141	0.39	0.45	0.36	12243	0.63	0.44	0.60	9395	1.67	4.62	0.95
Number of citations (t-1)	24141	0.10	1.29	0	12243	0.13	1.95	0	9395	0.054	0.91	0
Log of firm revenues (t-1)	24141	0.37	5.73	0	12243	0.87	18.3	0	9395	0.23	3.75	0

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

back

Placebo check

Migrant share	Beta	Obs	R^2
L.TFP (Olley and Pakes method)	-0.000679	45028	0.908
	(0.00123)		
L2.TFP (Olley and Pakes method)	0.000615	38512	0.920
	(0.00118)		
L3.TFP (Olley and Pakes method)	-0.000438	33523	0.929
	(0.00109)		
L4.TFP (Olley and Pakes method)	-0.000270	28863	0.939
	(0.00109)		
L5.TFP (Olley and Pakes method)	-0.0000718	24492	0.947
	(0.00101)		
L6.TFP (Olley and Pakes method)	-0.000569	20328	0.953
	(0.00127)		
L7.TFP (Olley and Pakes method)	-0.00188*	16365	0.963
	(0.00107)		
L8.TFP (Olley and Pakes method)	-0.00153	12680	0.972
	(0.00140)		
L9.TFP (Olley and Pakes method)	-0.00116	9278	0.980
	(0.00114)		
L10.TFP (Olley and Pakes method)	0.00142	6322	0.983
	(0.00223)		

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT. Notes per main table.

[back]

Experience / human capital splits

Log(TFP)	(1)	(2)	(3)	(4)	(5)	(6)
Firm migrant share	0.0904					
	(0.0901)					
L.Average migrants' experience	-0.000253	0.0113***	0.0129***	0.00779***	0.00200*	-0.000377
	(0.00104)	(0.00352)	(0.00356)	(0.00281)	(0.00106)	(0.00104)
L.Average natives' experience	-0.00195	0.0330***	0.0368***	0.00911	-0.00648***	-0.00183
	(0.00275)	(0.00527)	(0.00728)	(0.00590)	(0.00248)	(0.00275)
L. Share migrants with degree +		0.698***	0.703***	0.812***	0.208***	-0.159
		(0.175)	(0.170)	(0.134)	(0.0643)	(0.100)
L. Share natives with degree +		-0.105	0.170	0.582***	0.187***	-0.161**
		(0.119)	(0.117)	(0.0925)	(0.0396)	(0.0684)
L.Share workers with degree +	-0.172**					
	(0.0667)					
Controls	Yes	No	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	-	Yes	Yes	Yes	Yes	-
Industry FE	-	No	No	1-digit	4-digit	-
Firm FE	Yes	No	No	No	No	Yes
N groups	5194	5431	5431	5431	5424	5194
Observations	35874	36111	36111	36111	36104	35874
R^2	0.943	0.0271	0.0693	0.417	0.876	0.943

back

Language measures

- **Language:** we use the first language workers say they speak
  - 71.7% of workers in our test data do not report a first language; in these cases we assume first language = English
  - We map given first languages to language families, removing cases where people declare a non-native level of competence
  - Where people give this information in other languages, we use a language detection algorithm

  - Share foreign language in firm I, year t
  - Fractionalisation index of linguistic diversity: 156 languages => 39 branches => 23 families

back

Proximity / diversity metrics

Depvar = log(TFP)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Firm migrant share	0.0234	0.104	0.0971											
	(0.0840)	(0.0943)	(0.0946)											
Birth country diversity		-0.133*			-0.393**									
		(0.0684)			(0.184)									
Birth country diversity sq					0.266*									
					(0.151)									
Birth country bloc diversity			-0.125*	-0.0905		-0.432**								
			(0.0698)	(0.0616)		(0.186)								
Birth country bloc div sq						0.305**								
						(0.152)								
Share foreign language							0.108				0.137			
							(0.113)				(0.117)			
Language diversity								0.0233				-0.207		
								(0.0682)				(0.128)		
Language diversity sq												0.736**		
												(0.348)		
Language branch diversity									0.0231				-0.316**	
									(0.0770)				(0.136)	
Language branch div sq													1.266***	
													(0.436)	
Language family diversity										-0.142	-0.216			-0.358
										(0.194)	(0.201)			(0.314)
Language family diversity sq														1.323
														(1.504)
N groups	6141	6141	6141	6141	6141	6141	6141	6141	6141	6141	6141	6141	6141	6141
Observations	44996	44996	44996	44996	44996	44996	44996	44996	44996	44996	44996	44996	44996	44996
R-squared	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT. All regressions fit controls, year and area-region-firm FE. Other notes per main table.

back

Worker summary statistics

	A. All workers			B. Workers with Diffbot skills		
	All workers	Natives	Migrants	All workers	Natives	Migrants
Has degree or higher qualifications	0.808	0.788	0.889	0.815	0.795	0.893
	(0.394)	(0.408)	(0.315)	(0.388)	(0.403)	(0.309)
Has degree	0.600	0.625	0.496	0.602	0.629	0.491
	(0.490)	(0.484)	(0.500)	(0.490)	(0.483)	(0.500)
Has postgraduate degree	0.187	0.147	0.354	0.191	0.149	0.362
	(0.390)	(0.354)	(0.478)	(0.393)	(0.356)	(0.480)
Has PhD	0.0209	0.0166	0.0389	0.0221	0.0175	0.0408
	(0.143)	(0.128)	(0.193)	(0.147)	(0.131)	(0.198)
Years of labour market experience	12.81	12.93	12.30	13.23	13.41	12.52
	(8.792)	(8.978)	(7.948)	(8.779)	(8.974)	(7.904)
Tech occupation	0.0815	0.0775	0.0979	0.0849	0.0805	0.102
	(0.274)	(0.267)	(0.297)	(0.279)	(0.272)	(0.303)
STEM occupation	0.0953	0.0893	0.120	0.0992	0.0928	0.125
	(0.294)	(0.285)	(0.325)	(0.299)	(0.290)	(0.331)
Managerial occupation	0.301	0.305	0.285	0.313	0.318	0.291
	(0.459)	(0.460)	(0.451)	(0.464)	(0.466)	(0.454)
Most probable topic (rank 1)				13.10	13.06	13.26
				(7.154)	(7.207)	(6.934)
GPT-4o-based ranking of top1_topic				13.07	12.87	13.87
				(7.502)	(7.479)	(7.542)
Observations	*538872*	*434463*	*104409*	*466715*	*373879*	*92836*

Source: Diffbot

[back]

Distinct skills

- **Do migrants carry distinctive skills compared to natives doing the same jobs?** We regress the following cross-sectional regression for individual *i* in SOC4 bin *o*:

$$\text{SKILL\_io} = a + b\text{MIGRANT\_io} + \mathbf{X}\text{c\_io} + \text{e\_io} \qquad (5)$$

- Where
  - SKILL is the dominant topic held by the individual {1,25}
  - MIGRANT is a dummy for migrant status
  - **X** is a set of individual-level controls, including highest qualification, years of experience and the year the worker is observed (to take account of coverage differences across years)
  - IV specifications use MIGRANT_hat_{io} = $1 - (\text{HISTORIC\_UK})_{io}$, where HISTORIC_UK is a dummy = 1 for workers with surnames present in the 1881 Census

- In this regression topics are numbered but have no inherent ordering, i.e. topic 20 is not ‘better’ or ‘worse’ than topic 10. This means we focus on the significance of *b* but not its sign. **Significant *b* implies that migrants have distinctive dominant topics relative to natives, controlling for individual characteristics.**

Migrants carry (slightly) distinctive skills

	(1) ols	(2) rf	(3) iv	(4) placebo
migrant	0.155***		1.416***	-0.0443
	(0.0258)		(0.0886)	(0.0276)
Predicted migrant based on historic name		0.378***		
		(0.0235)		
Observations	437630	437630	437630	402088
R^2	0.122	0.122	-0.00154	0.00115
Oster δ / migrant	11.16			
First stage, predicted migrant			0.267***	
			(0.002)	
K-P under-identification test			24053.8	
K-P weak instrument test			27344.7	

Source: Diffbot. Sample is workers with Diffbot skills and observables. We regress a worker's dominant skills topic (numbered 1-25) on migrant status and other observable characteristics. All specifications include year dummies (to control for year of observation), SOC4 dummies, and controls (female, foreign language speaker, graduate or above, degree type, years of experience). Robust standard errors in parentheses. Col (1) fits OLS. Col (2) fits reduced form. Col (3) uses IV specification: predicted migrant status based on historic name info. Col (4) runs a placebo check which shuffles the dominant topic count across workers.

[back]

Distinctive skills: sensitivity

	(1) ols	(2) ologit	(3) oprobit	(4) _2023	(5) _1123	(6) top2	(7) top3
migrant	0.155***	0.0378***	0.00964**	0.0844*	0.158***	0.0913***	-0.0678**
	(0.0258)	(0.00646)	(0.00388)	(0.0455)	(0.0261)	(0.0266)	(0.0271)
Observations	437636	437636	437636	145789	427953	437636	437636
R^2	0.122			0.116	0.121	0.0320	0.0107
Pseudo-R^2		0.0240	0.0216				
Oster δ / migrant	*11.16*						

Source: Diffbot. Sample is workers with Diffbot skills and observables. We regress a worker's dominant skills topic (numbered 1-25) on migrant status and other observable characteristics. All specifications include year dummies (to control for year of observation), SOC4 dummies, and controls (female, foreign language speaker, graduate or above, degree type, years of experience). Robust standard errors in parentheses. Cols (2) and (3) fit ordered logit and probit estimators. Cols (4) and (5) regress for 2023 and 2011-2023. Cols (6) and (7) use alternate dominant topics.

[Descriptives] [back]

Firms add migrants faster in tech & STEM roles than management roles

MNR in …	(1) mgt roles	(2) tech roles	(3) STEM roles	(4) mgt roles	(5) tech roles	(6) STEM roles
Firm migrant share	1.402***	1.493***	1.537***			
	(0.111)	(0.138)	(0.141)			
Share high skill migrants				1.431***	1.628***	1.644***
				(0.124)	(0.155)	(0.163)
N groups	6657	4316	4298	6657	4316	4298
Observations	49703	29484	29669	49703	29484	29669
R^2	0.784	0.706	0.715	0.782	0.706	0.715

Source: Diffbot, Orbis, Historical OpenCorporates, PATSTAT. All regressions fit controls, area, region, year and firm FE. Other notes per main table.

back

Role specialisation descriptives

Full sample	**count**	**mean**	**sd**	**p50**
Management roles (migrant share / native share) ratio	63483	0.25	0.54	0.071
Tech roles (migrant share / native share) ratio	36324	0.26	0.62	0
STEM roles (migrant share / native share) ratio	36528	0.27	0.63	0
Migrant tech role specialisation	21880	1.37	2.80	0
Migrant stem role specialisation	21883	1.41	2.87	0.31
Migrant management role specialisation	14220	0.73	1.09	0.43
Observations	64363			

Regression sample	**count**	**mean**	**sd**	**p50**
Management roles (migrant share / native share) ratio	42553	0.25	0.54	0.100
Tech roles (migrant share / native share) ratio	25993	0.27	0.61	0
STEM roles (migrant share / native share) ratio	26118	0.28	0.62	0
Migrant tech role specialisation	16474	1.41	2.76	0.30
Migrant stem role specialisation	16493	1.45	2.77	0.43
Migrant management role specialisation	10883	0.73	1.03	0.44
Observations	42969			

Source: Diffbot, Orbis, Historical OpenCorporates, PATSTAT.

[back]

No link from migrant task specialisation to TFP

Log(TFP)	(1)	(2)	(3)
Migrant tech specialisation	-0.00318		
	(0.00231)		
Migrant STEM specialisation		-0.00321	
		(0.00241)	
Migrant management specialisation			0.00529
			(0.00712)
N groups	2484	2483	1728
Observations	15080	15170	10183

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT.
All regressions fit controls, area, region, year and firm FE.. Other notes per main table.