Migration, Diversity and Firm Performance*

Tom Kemeny[1], Max Nathan[†2,3], Ceren Ozgen[4], Guido Pialli[2,5], Anna Rosso[6], Mateo Seré[2] and Anna Valero[3,7]

[1]University of Toronto, Canada
[2]University College London, United Kingdom
[3]Centre for Economic Performance, United Kingdom
[4]University of Birmingham, United Kingdom
[5]University of Turin, Italy
[6]University of Milan, Italy
[7]London School of Economics and Political Science, United Kingdom

April 24, 2026

Abstract

This paper documents the effects of migrant workers and migrant diversity on firms' innovation and productivity. We build a novel matched worker-firm dataset leveraging over 500,000 online worker profiles alongside rich company microdata. Individual education and career histories are combined with large firms in the UK company register, plus detailed financial and patents data. To identify causal effects and underlying mechanisms, we implement placebo tests and instrumental variable strategies exploiting historical name patterns. We document three main findings. First, using our granular worker-level data, we show that migrant workers possess distinctive, and more complex skills than natives in the same jobs, especially in technical and STEM roles. Second, migrant specialization in technical occupations is robustly associated with higher firm-level patenting. Third, however, we find a positive but non-significant relationship between firm productivity and its migrant share, with larger magnitudes for highly educated migrants and in technology- and STEM-intensive firms. Together, these results highlight the skill composition and occupational allocation of migrants as key channels linking migration to firm performance.

JEL classification: C81, J24, J61, L25, M14, O31
Keywords: immigration, human capital, innovation, productivity, data science

*Draft paper. Please do not cite without permission. Thanks to Sameera Siddiqui for outstanding research assistance. Audiences at GEOINNO 2026 and 2024, GCEG 2025, RSA Winter 2025, UKRI, UCL, Birmingham, Ca'Foscari, Hamburg and Reading provided helpful comments. Thanks to Rebecca Lee at OpenCorporates, Filipe Mesquita at Diffbot, Paul Longley and Justin van Dijk at UCL for help and advice on data/code, and to our advisory board for constructive feedback. For data, thanks to Diffbot, OpenCorporates, Orbis / Orbis Historical, Orbis IP and PATSTAT Global. This research is funded by UKRI Grant ES/W010232/1. The project was reviewed and approved by the UCL Research Ethics Board, application 22883/00. The usual disclaimers apply.

†Corresponding author. E-mail: max.nathan@ucl.ac.uk

Author contributions

Tom Kemeny: conceptualisation, methodology, formal analysis, writing - reviewing and editing; **Max Nathan:** conceptualisation, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, writing - original draft; **Ceren Ozgen:** conceptualisation, methodology, writing - reviewing and editing; **Guido Pialli:** data curation, methodology, formal analysis, software, writing - original draft; **Anna Rosso:** data curation, methodology, software, writing - reviewing and editing; **Mateo Seré:** data curation, formal analysis, methodology, software, validation; **Anna Valero:** conceptualisation, methodology, writing - reviewing and editing. Authors are listed alphabetically.

1 Introduction

This study examines whether and how migrant workers affect firm performance in the UK, focusing on their potential impacts on productivity and innovation. Understanding this relationship is important for a number of reasons. First, in the UK as in many countries, immigration and its economic and social impacts commands enormous political and policy attention. Recent permanent migration to OECD countries has reached record highs, while flows have become increasingly skill-biased: in 2010, 35 percent of OECD country migrants had degrees or above; by 2023, this had risen to 50 percent (OECD, 2024; Commission, 2023). These compositional shifts are often not reflected in debates about the economic impacts of migration, which focus instead on fiscal and labour market effects (Manning, 2025). Second, it is especially important to understand migration's impacts on productivity, the single biggest influence on long-term economic growth. Existing evidence on migration is mixed, and highlights differences across industry and types of migrant worker - see Hall and Manning (2024) and Ozgen (2021) for recent reviews. Understanding the mechanisms that drive these higher-level differences is an important challenge for researchers. A third, related challenge is the lack of granular, large-scale employer-employee data in many countries, including the UK: the large majority of studies cited have to use area-level migration data rather than directly observing migrant workers in firms.

In this paper we tackle these challenges using a new, large and detailed worker-firm dataset.Using the UK open companies register and a commercial knowledge graph of the entire public internet, we match UK companies in the graph using precise identifiers and extract their worker profiles. We link companies and workers to rich company-level financial and patent data from Orbis Historical and Orbis IP / PATSTAT. The data cover over 9,000 larger UK companies and over 530,000 workers employed in these firms between 2007 and 2023. Following Jin et al. (2025b) and Lee and Glennon (2023), we use workers' lowest observed education location to proxy for migrant/native status. Using secondary evidence and a validation exercise, we show that this should give us a lower bound on the true migrant share in our data. Similarly to Dorn et al. (2025) and Fedyk and Hodson (2022),

we use natural language processing on individuals' stated skills to model clusters of learned capabilities, showing that these correlate with worker years of experience but are distinct from formal qualifications. Section 3 summarises the data, build and validation; see Gray et al. (2025) for full details.

We use these novel resources to look at overall linkages between firm demography and firm performance, using both the company and worker-level layers of our data to explore mechanisms. To give a descriptive sense, Figure 1 shows a binned scatterplot of estimated TFP growth against firms' estimated migrant share, using our company-level panel 2007-2023 and controlling for 4-digit industry, region and year dummies. We find a weakly positive association between firms' migrant worker share and productivity growth.

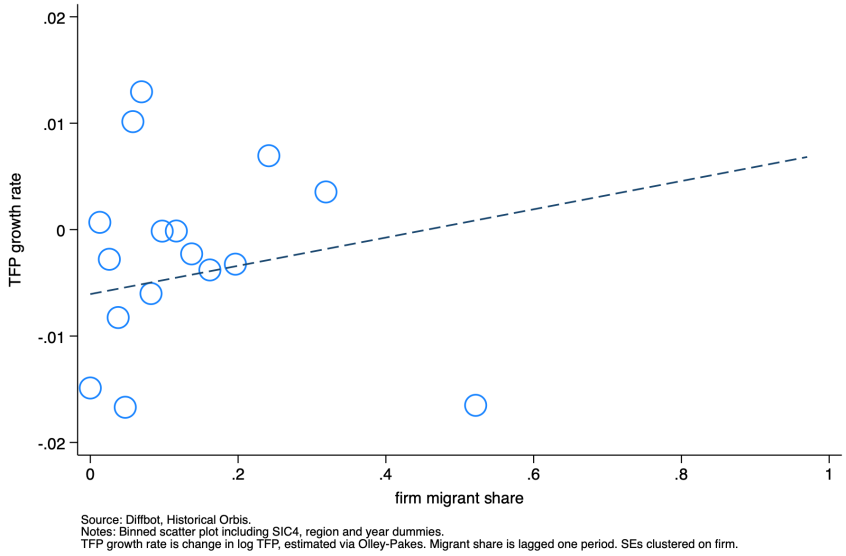

Figure 1. Binned scatter plot of firm TFP growth against migrant share, 2007-2023

We use instrumental variables to identify the causal effects of migration on firm performance, leveraging historic Census name information and the deep spatial persistence of names and name families over time (Balsmeier et al., 2025; Burchardi et al., 2018; Barone and Mocetti, 2021; Clark et al., 2015; Clark and Cummins, 2014). Our firm-level instrument is a shift-share design. The share is derived from surnames present in the UK in the 1881 Census, for example 'Jones', which we term 'UK-established' surnames. We argue that given name persistence, native workers in our sample are more likely to have UK-

established surnames. We use worker-level surname information to fuzzily predict initial firm-level shares of UK-born workers. The shift is the change in the overall prevalence of UK-established surnames in the whole sample, relative to the base year. We argue that, conditional on controls, there is no obvious pathway from having a relatively larger share of workers named 'Jones', or similar, to company-level performance. We also use an alternate instrument to directly proxy for UK-born status in the worker-level data. In both cases there are two constraints on precision: we will mis-ascribe both UK-born descendants of migrants arriving after 1881, and in-sample migrant workers who happen to share 'UK-established' names. These constraints work in opposing directions, meaning that instruments are noisy but not biased.

We have three main findings. First, in both OLS and IV regressions, we find small, positive, and non-significant associations between firm migrant shares and TFP growth rates and levels. Industry and firm heterogeneity explain these results. While coefficients of the migrant share are often larger, extensions by industry bloc, occupational mix, migrant and UK-born qualifications and experience also show positive but non-significant effects. Fitting workforce birth country / linguistic diversity measures (Alesina et al., 2016) also yields nonsignificant results. As with Hall and Manning (2024) these company-level results do not definitively rule migrant-productivity effects in or out: we turn to the worker data to explore potential mechanisms. Our second finding is that, in both OLS and IV estimations, migrants and UK-born workers doing the same types of jobs carry distinctive capabilities, controlling for qualifications, experience and other observables. Similarly, we show that migrants in 'tech' and STEM roles have more complex capabilities than UK-born counterparts in the same roles. Third, building on these findings and frameworks in Mayda et al. (2022), Peri and Sparber (2011) and others, we construct firm-level measures of migrant and native specialisation and deploy these in our company panel. Consistent with Mayda et al. (2022), we find that migrant tech specialisation is associated with higher patenting in firms, even when we control for alternate channels such as migrant human capital and workforce diversity. The effect is most pronounced for firms that already patent. We find no effect of specialisation on firm TFP.

The paper makes two main contributions to the field. First, we generate new evidence on migrant workers' role in shaping firm outcomes. In particular, we shed new light on worker-level differences and relate these to firm outcomes. Second, we introduce novel, at-scale worker-firm data which complement conventional administrative sources, and which can be reproduced in other countries with open company registers (Gray et al., 2025). As far as we are aware, this is the first paper to leverage these kinds of data to explore these questions. A version of the data will be publicly available for researchers.

1.1 Related literatures

This paper relates to four bodies of research. The first is on migration and firm productivity (Campo et al., 2024; Exadaktylos et al., 2024; Nam and Portes, 2023; Fabling et al., 2022; Dale-Olsen and Finseraas, 2020; Ottaviano et al., 2018; Mitaritonna et al., 2017; Trax et al., 2015; Parrotta et al., 2014a; Paserman, 2013; Peri, 2012). The three closest papers to ours are Exadaktylos et al. (2024), Hall and Manning (2024) and Ottaviano et al. (2018), all of which look at the UK case. Exadaktylos et al. (2024) show that hiring foreign managers increases TFP in UK manufacturing firms. Hall and Manning (2024) combine rich firm-level productivity data and detailed neighbourhood-level migration measures. Ottaviano et al. (2018) combine firm-level export and output/worker and local-authority level migration measures. Compared with these papers, we work with a much larger, cross-industry company-level sample, we can observe detailed workforce characteristics, and thus we can test relationships and mechanisms in more detail.

Second, our paper also draws on a narrower set of studies that explore migrant workers and task specialisation in firms (Mayda et al., 2022; Lin, 2019; Foged and Peri, 2016; Peri and Sparber, 2011,0). The closest papers here are Mayda et al. (2022), Lin (2019) and Peri and Sparber (2011) which focus on highly educated migrants and complex work settings.

Third, we also contribute to a large and growing literature on skilled migrants and innovation, which encompasses work on academic researchers and inventors (Pellegrino et al., 2023; Widmann, 2023; Hofstra et al., 2020; Ferrucci and Lissoni, 2019), entrepreneurs (Lee et al., 2025) as well as firm-level (Mack et al., 2025; Mayda et al., 2022; Parrotta et al.,

2014b; Liu et al., 2023; Doran et al., 2022), area-level (Boberg-Fazlić and Sharp, 2023; Akcigit et al., 2017; Cooke and Kemeny, 2017; Nathan, 2015; Hunt and Gauthier-Loiselle, 2010) and cross-country studies (Jin et al., 2025a; Wigger, 2021; Akcigit et al., 2018; Kerr, 2008).

Finally, our paper forms part of an emerging literature that combine web data and firm-level data to explore a range of research questions (see (Dahlke et al., 2025) for a review). Studies using similar dataframes to ours include (Jeffers, 2024; Gagliardi et al., 2024; Lee and Glennon, 2023; Babina et al., 2022; Fedyk and Hodson, 2022; Tambe et al., 2020; Rock, 2019).

2 Framework

The overall effect of migrant workers on firm productivity is ambiguous: theory suggests positive channels, negative channels and confounders (Ozgen, 2021).

First, migrant workers may be positively selected on qualifications and/or capabilities which make them directly more innovative than native workers, more productive, or both (Hunt and Gauthier-Loiselle, 2010; Azoulay et al., 2022). Second, these migrant-native differences may allow firms to improve worker-role matching, driving up firm performance through migrant / native specialisation (Peri and Sparber, 2009,0; Foged and Peri, 2016; Lin, 2019; Mayda et al., 2022). Third, to the extent migrants bring different experiences and types of knowledge to a firm, the resulting 'cognitive diversity' can help innovation (Page, 2007; Kerr and Lincoln, 2010; Nathan, 2015; Kemeny, 2017; Glover and Kim, 2021) and improve scrutiny (Levine et al., 2014), both feeding through to productivity improvements.

These positive channels are potentially countered by mechanisms running in the other direction. First, migrants may be negatively, not positively selected into firms (for example, if their lower wage costs outweigh performance differences). Second, discrimination by a firm's managers may lead to suboptimal worker/task matching. Third, diverse teams may find it harder to communicate and trust each other, at least in the short term, affecting innovation (Alesina and Ferrara, 2005; Cheng and Weinberg, 2021).

In practice, other forces may dominate these channels: overall effect sizes may not be very large; migrants and natives may be highly or completely substitutable; and birth country diversity may not correlate with cognitive diversity, if workers are identical on other dimensions - so-called 'McKinsey multiculturalism' - or if foreign experience per se may be more important than country of birth (Exadaktylos et al., 2024).

The size of these channels is also likely to be heterogenous across industries, business processes and (potentially) locations. In particular, skilled migrants may sort into 'knowledge-intensive' settings such as tech, finance or business services, characterised by high levels of task complexity (Kerr, 2008; Berliant and Fujita, 2012). Impacts may also be amplified for firms in big cities, due to workforce composition effects, interactions with agglomeration effects, or both (Ottaviano and Peri, 2006).

This range of channels and moderators in play has implications for empirical work. Consistent with our firm-level results, observing the overall migrant TFP linkage in a pool of firms will net out positive, negative and zero mechanisms. Similarly, aggregate relationships may be zero or non-significant, if positive channels dominate in some sectors, and negative / zero effects in others.

3 Data

We construct a novel worker-firm dataset leveraging over 500,000 online worker profiles alongside rich company microdata. Here we summarise data sources and build.

3.1 Data sources

Diffbot is the world's largest commercial knowledge graph, built from the public web (Mesquita et al., 2019). At the start of 2025, the graph included 278.9m active companies and 231.3m individuals in employment worldwide; in the UK the graph covered 4.7m active companies and 10.6m workers.[1]

Diffbot extracts individual and company-level profiles from sources across the public

[1] See www.diffbot.com for more detail.

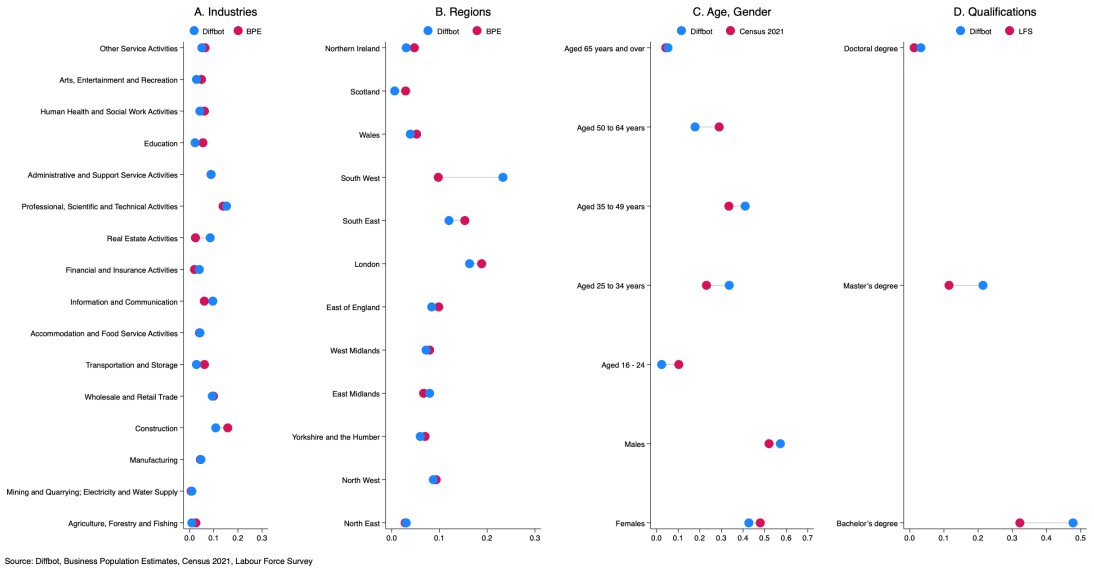

Figure 2. Benchmarking UK firms and workers in Diffbot

Panel A compares the distribution of active Diffbot companies across industries at the start of 2022, against those in the 2022 BPE. Panel B repeats the exercise for regions. Panel C compares UK worker age and gender distribution in Diffbot at the start of 2022, against 2021 England & Wales Census data. Panel D compares highest reported qualifications for workers in Diffbot against the Labour Force Survey.

web and organises them into a searchable knowledge graph. Individual profiles include gender, year of birth, languages spoken, as well as detailed education and employment histories (education dates, subject, institution and qualification; employer identifiers, start and end dates, job titles, and job descriptions). The data also includes a taxonomy of worker skills (see below). Diffbot uses supervised learning techniques to assign individual workers to companies based on their career histories. Workers typically keep online professional profiles updated at least once a year, mitigating concerns about measurement error from 'laggy' data (Jeffers, 2024).

For firms in the United Kingdom, Diffbot profiles include company names and unique identifiers from Companies House, the UK's Open Companies Register. We use these identifiers to precisely match individuals and companies in Diffbot with company-level information from Bureau van Dijk's Orbis and Orbis Historical, which gives us detailed, long-running productivity and financial data. We supplement this with information on patenting from

Orbis IP and PATSTAT Global.[2] Appendix A gives further details of these data.

Figure 2 summarises tests comparing coverage of UK-based firms and workers in Diffbot against UK administrative datasets. Overall, firm counts are in line with those in administrative data, and coverage across sectors and regions is also broadly representative. Reflecting its sources, Diffbot slightly over-samples on services and on sectors with a strong online presence. Similarly, Diffbot skews slightly male, and more strongly towards younger workers, those with degrees and those in professional / managerial / technical occupations. This is consistent with other studies using similar data (Dorn et al., 2025; Fedyk and Hodson, 2022; Rajkumar et al., 2022).

3.2 Company dataframe

Our sampling frame consists of 'medium' and 'large' UK companies which provide complete audited annual accounts, and are defined using turnover, balance sheet and workforce thresholds. This frame includes 55,187 firms active between 2007 and 2023.[3] We search for these companies in Diffbot, keeping only matches on the company record number (CRN) and where there is at least one worker in Diffbot. The result is a subsample of 33,081 firms, just under 60 percent of the starting sample. Validation tests suggest matches are a random subset of our starting sample. See Appendix B for details.

As Diffbot does not capture the population of workers, we measure the 'coverage ratio'—the count of workers in Diffbot divided by the reported headcount in Orbis/Orbis Historical—and explore its predictors. Tests suggest selection on both observables and unobservables is trivial, with controls and company fixed effects explaining almost 80 percent of coverage ratio variation (See Appendix B). We keep 11,233 companies with coverage ratios of 0.25 or above, allowing us to restrict to high-coverage companies in robustness tests. We extract the all-time workforce for CRN-matched companies from Diffbot, building a full matrix of worker characteristics and employment spells. We use LinkTransformer (Arora and Dell, 2023) to match job descriptions to 2020 UK Standard Occupational Codes (SOCs) at 4-digit

[2] Revelio and Cognism, which are increasingly used in social science research, do not contain firm-level identifiers, and linking to firm data requires fuzzy matching which can deal with the complexities of corporate structures. See Fedyk and Hodson (2022) for an example of the workflow required.

[3] Our panel starts in 2007 because Historical Orbis is not available before this date.

level.

For the company panel, we collapse worker-level data to company-year level. We merge in productivity metrics and financial information from Orbis and Orbis Historical. We also match in company-level patent counts from Orbis IP, using BvD identification numbers. We trim for missing data, as well as revised company coverage. This reduces our sample to an unbalanced panel of 9,007 companies, which draws on a matrix for 542,534 workers.

3.3 Worker dataframe

To explore worker characteristics in detail, we merge profiles from the worker matrix with variables built from Diffbot skills. We use topic modelling and large language models (LLMs) to build skills variables. Diffbot's skills ontology includes roughly 32,000 professional skills, drawn from individual profile content (e.g. qualifications, job titles and descriptions, self-described skills and endorsements). Diffbot validates these based on prevalence in Wikidata and online professional platforms. Skills terms cover general capabilities (e.g. team working, office software), management skills (e.g. team leadership) and specialised capabilities (e.g. Python, SQL, econometrics). Intuitively, and in line with human capital models (Dorn et al., 2025), we can think of Diffbot skills as representing learned capabilities that help workers complete tasks, and which will be correlated with experience, but distinct from formal qualifications. We directly test these assumptions in Section 4.

Diffbot skills are observed once per worker, so we treat them as individual time-invariant characteristics. We observe skills for over 82 percent of workers in our data, finding no evidence of selection. As with other work in this area (Dorn et al., 2025; Fedyk and Hodson, 2022) we use topic modelling to cluster the original 32,000 skill types. We optimise for a 25-topic model. We use LLMs to label topics and to score topics by the complexity of the activities in each topic, using the ISCO typology as a benchmark. Appendix C gives more details of the build, diagnostics and labelling. This gives us a dataframe of over 530k workers, of whom 466,715 individuals have observed skills between 2007 and 2023. Of these, 35 percent are observed in 2023, the rest in small shares across other years. In our main analysis we use the whole sample; we show that results do not vary when we change the

time window.

3.4 Company outcome measures

Our first company-level outcome is estimated Total Factor Productivity. Firm-level studies of migrant–productivity links take one of two approaches in modelling firm productivity. Studies like Hall and Manning (2024) and Dale-Olsen and Finseraas (2020) directly estimate total factor productivity using a control function setting (Olley and Pakes, 1996) including migrant share as an additional covariate.[4] The second, larger group of studies (i.e., Exadaktylos et al., 2024; Ottaviano et al., 2018; Mitaritonna et al., 2017; Trax et al., 2015; Parrotta et al., 2014a; Paserman, 2013) uses a two-step design. First, firm productivity is derived as a residual using a control function or other estimation.[5] Next, productivity estimates are regressed on firm migrant share, covariates and fixed effects.

When run together, these two types of designs deliver similar coefficients of the migrant share (Hall and Manning, 2024; Dale-Olsen and Finseraas, 2020). We therefore follow the two-step approach, which has the advantage of being less computationally demanding.[6] In the first step, we derive TFP estimates using a control function design including capital and labour. In the second step, we regress these estimates on firm-level migrant shares (or instrumented shares), covariates and fixed effects (see Section 5 for details.) Specifically, we estimate TFP using a production function with log-transformed labour l and capital k as inputs:

$$y_{it} = \alpha + \beta_1 l_{it} + \beta_2 k_{it} + e_{it} \tag{1}$$

TFP is then defined as the residual:

$$\hat{TFP}_{ijt} = y_{ijt} - \beta_1 \hat{l}_{ijt} - \beta_2 \hat{k}_{ijt} \tag{2}$$

[4] Control function designs have the advantage of controlling for firms' endogenous capital and labour choices, and instrument for unobserved firm-level productivity shocks using investment choices.

[5] Ottaviano et al. (2018) estimate labour productivity; Trax et al. (2015) use GVA/per worker.

[6] We ran runtime tests on a UCL server, comparing OLS vs. control function estimates with and without firm fixed effects. Results available on request.

where $\hat{l}_{ijt}$ and $\hat{k}_{ijt}$ are the industry-specific elasticities obtained from estimating eq. 2 in each industry j. Our preferred estimation for eq. 1 uses the method defined in Olley and Pakes (1996). In sensitivity tests, we also apply an ACF correction (Ackerberg et al., 2015), yielding almost identical results.

We also build measures of firms' patent counts and all-time patent citations. 678 companies in our sample file patents, just over 7.5 percent of the sample. We build time-varying patent counts using the fractional method, so that patents with two applicants are each assigned a half patent (Autor et al., 2020; Arora et al., 2021). Per Appendix A, patent counts cover the three major patent offices worldwide (USPTO, EPO and JPO) as well as the UK Intellectual Property Office.

3.5 Measuring migrant status

Country of birth is not directly observed for individuals in Diffbot.[7] Instead, following Jin et al. (2025b) and Lee and Glennon (2023) we use the country of an individual's earliest available educational qualification to proxy for their migrant/native status. In our data, the lowest observed level of education for workers is typically university. As UK universities attract a lot of foreign students, our measure is likely to generate a lower bound on the true share of migrants. Appendix D summarises secondary data and validation exercises, both of which provide strong supporting evidence for the validity of the measure and the lower bound assumption.

4 Descriptives

4.1 Company descriptives

Table 1 gives summary statistics for the company panel. Panel A shows all years, Panels B and C 2007 and 2023 respectively. Appendix Table E1 decomposes the table by coverage rate bins, showing little variation across firms with different levels of coverage.

Firms' estimated TFP growth varies widely across firms but with little time variation in the average firm. Over time, TFP growth rates shift very slightly rightwards (Figure 3,

[7] Specifically, it is a field in the graph but almost always empty for workers.

Table 1. Descriptive statistics

	A. All years			B. 2007			C. 2023		
Variable	N. obs	Mean	Sd	N. obs	Mean	Sd	N. obs	Mean	Sd
Log TFP (Olley-Pakes)	63,851	3.866	1.610	232	3.557	1.511	4,258	3.931	1.668
Number of patents	70,911	0.094	1.429	262	0.075	0.799	4,750	0.000	0.000
Number of citations (all-time)	70,911	0.475	11.192	262	0.691	7.101	4,750	0.000	0.000
Share migrant workers	70,911	0.121	0.147	262	0.123	0.161	4,750	0.127	0.143
Share of workers with a college or higher degree	70,911	0.501	0.245	262	0.513	0.244	4,750	0.510	0.241
Share migrants with degree or above	70,911	0.105	0.140	262	0.104	0.146	4,750	0.110	0.137
Share UK workers with degree or above	70,911	0.389	0.208	262	0.401	0.215	4,750	0.392	0.201
Share of workers in tech occupations	70,911	0.074	0.095	262	0.081	0.108	4,750	0.077	0.093
Share of workers in stem occupations	70,911	0.082	0.110	262	0.089	0.123	4,750	0.086	0.109
Share of workers in managerial occupations	70,911	0.374	0.200	262	0.401	0.214	4,750	0.366	0.194
Share of migrant workers in tech occupations	70,911	0.010	0.032	262	0.012	0.044	4,750	0.012	0.036
Share of UK workers in tech occupations	70,911	0.038	0.061	262	0.038	0.057	4,750	0.039	0.057
Share of migrant workers in stem occupations	70,911	0.012	0.034	262	0.014	0.045	4,750	0.014	0.038
Share of UK workers in stem occupations	70,911	0.044	0.072	262	0.045	0.069	4,750	0.045	0.067
Share of migrants in managerial occupations	70,911	0.040	0.079	262	0.040	0.077	4,750	0.041	0.077
Share of UK workers in managerial occupations	70,911	0.185	0.140	262	0.205	0.145	4,750	0.183	0.131
Workforce average age	69,518	43.639	5.832	257	49.241	5.388	4,668	41.950	5.582
Share of females	70,911	0.334	0.208	262	0.323	0.238	4,750	0.341	0.201
Number of employees	61,519	89.664	116.229	200	79.855	75.386	3,927	105.385	126.395
Firm age	70,911	6.006	3.515	262	1.000	0.000	4,750	11.335	2.450
Share of workers in non-executive occupations	70,911	0.022	0.061	262	0.020	0.061	4,750	0.022	0.058
Company has foreign subsidiaries	70,911	0.002	0.042	262	0.000	0.000	4,750	0.000	0.000
Number of subsidiaries	70,911	0.807	2.946	262	1.115	3.610	4,750	0.000	0.000
Log firm revenues	68,857	16.700	1.186	245	16.659	1.325	4,518	16.938	1.229

Notes: This table reports descriptive statistics (number of obs., mean and standard deviation) for the main variables used in the regression model.

Panel A). Patenting is very unevenly distributed across firms, as only a minority patent at all. Consistent with the sampling frame and earlier diagnostics, in the average firm the share of workers with a degree or higher degree is 50 percent; shares of tech / STEM workers are just under 10 percent, and managers 38 percent.

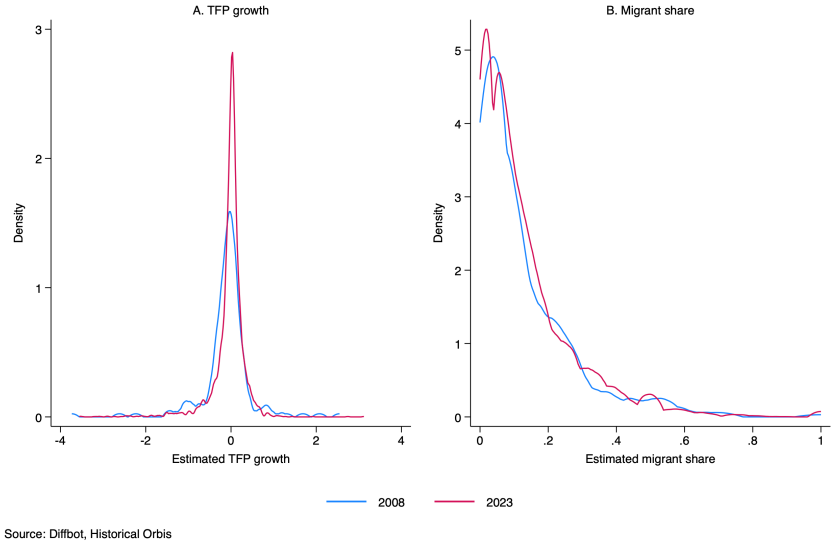

Figure 3. Kernel density plots of TFP growth and firm migrant share

The mean share of migrants in the average firm's workforce is 12 percent, slightly lower than the share of migrants in the UK workforce. Somewhat surprisingly, the workforce shares of migrants who are graduates or above, tech/STEM workers and managers are lower than UK-born, although the former rises slightly over time.

The share of migrant workers among large firms is relatively time-invariant (Figure 3, Panel B), despite the major changes to overall migrant flows and sending countries during the sample period. Figure 4 decomposes migrant worker types by sending country blocs. There is little time variation in workforce composition, with the exceptions of three sending blocs where shares increase over time: Central America and the Caribbean, Eastern Europe, and Western Europe.

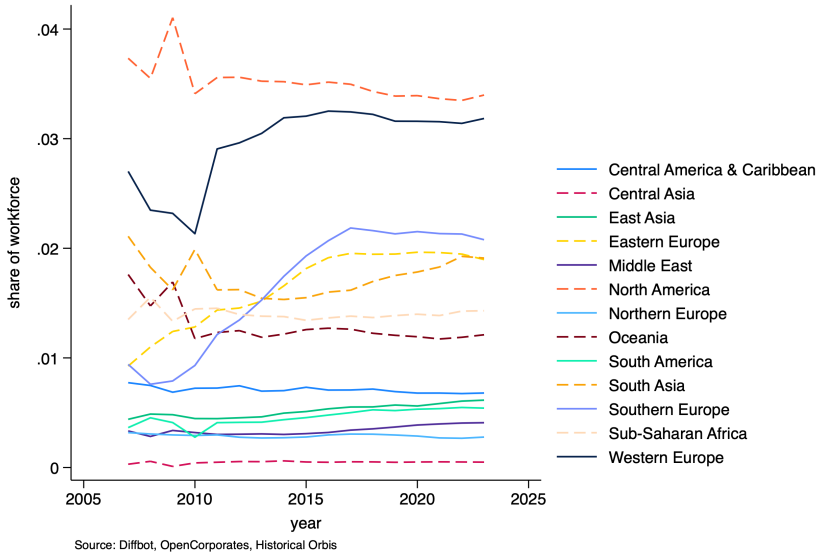

Figure 4. Change in migrant workforce composition by sending country bloc

4.2 Worker descriptives

Table 2 shows summary statistics for the merged worker dataframe, providing more detail on qualifications, experience and skills. For each characteristic, we show means and standard deviations for all workers, then decompose these for migrants and for natives. Panel A describes the full sample, while in Panel B we restrict to cover only workers for whom we observe skills. As a cross-check, Appendix Table E2 shows summary statistics as workforce-level shares and standard deviations. These are close to the company-level results in Table 1, but diverge slightly because workers are unevenly distributed across companies.

In-sample UK-born workers have a higher graduate share than migrants, and have on average a year more work experience. However, the share of migrants with postgraduate education is substantially higher than for natives. Migrants are also more likely to work in tech and STEM occupations, and less likely to work in managerial occupations. The average migrant and native have very similar dominant skills, while the average migrant exhibits a higher level of skill complexity. All reported differences distinguishing migrants and natives are statistically significant.

Table 3 shows the top 25 occupations, crosswalked from job titles to SOC2020 4-digit codes as above. Panel A gives counts for the full sample. Consistent with the diagnostics

Table 2. 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	538,872	434,463	104,409	466,715	373,879	92,836

Notes: The table shows means (with standard deviations in parentheses) for each worker characteristic, for the whole sample, natives and migrants, respectively. *Source:* Diffbot.

in Section 3, in the full sample, workers in managerial, professional and technical roles predominate. Panel B gives the top 25 occupations by shares of migrant workers. Notably, the most migrant-intensive occupations are split between managerial, technical and professional roles; roles in retail/leisure; foreign-language related activities (such as teaching and translation); and more routine occupations.

Table 3. Top SOC4 titles (all workers vs. migrant workers)

A. All workers		B. **Migrant workers**		
SOC4 title (all workers)	Count (all)	SOC4 title (migrant workers)	Count (total)	Migrant share
financial accounts managers	27411	biochemists and biomedical scientists	124	39.5%
management consultants and business analysts	17182	restaurant and catering establishment managers and proprietors	152	37.5%
customer service managers	14364	leisure and travel service occupations nec	32	37.5%
marketing and commercial managers	14248	functional managers and directors nec	102	37.3%
office managers	13620	authors writers and translators	583	36.7%
programmers and software development professionals	11232	dental nurses	36	36.1%
sales administrators	11224	generalist medical practitioners	127	35.4%
it project managers	10969	assemblers and routine operatives nec	48	35.4%
finance officers	10686	postal workers mail sorters and messengers	33	33.3%
business sales executives	10036	fishing and other elementary agriculture occupations nec	3	33.3%
it managers	9891	bar and catering supervisors	251	33.1%
information technology directors	9610	architects	4780	33.0%
human resource managers and directors	9350	housekeepers and related occupations	73	32.9%
mechanical engineers	8932	education advisers and school inspectors	46	32.6%
directors in consultancy services	8615	butchers	128	32.0%
data analysts	8586	market and street traders and assistants	931	31.5%
senior care workers	7659	refuse and salvage occupations	32	31.3%
sales accounts and business development managers	7418	special and additional needs education teaching professionals	61	31.1%
customer service supervisors	7069	business and related research professionals	1231	31.1%
sales and retail assistants	6917	programmers and software development professionals	11232	31.0%
solicitors and lawyers	6867	launderers dry cleaners and pressers	13	30.8%
sales supervisors	6344	teachers of english as a foreign language	449	30.5%
it operations technicians	6135	importers and exporters	648	30.2%
engineering technicians	5819	betting shop and gambling establishment managers	80	30.0%
project support officers	5471	residential day and domiciliary care managers and proprietors	40	30.0%

Notes: This table reports counts of workers in the top-25 occupations crosswalked to SOC4. Panel A provides data for all workers, while Panel B for migrant workers. *Source:* Diffbot.

Table 4 shows the results of Diffbot skills benchmarking. Skills in online resumes likely reflect multidimensional skills that a) derive from experience and b) are distinct from formal qualifications, since people with qualifications in (say) humanities or STEM often then find work in unrelated occupations (Dorn et al., 2025; Deming and Noray, 2020). This implies that for any given type of job, skills should be positively correlated with workers' years in the labour market, and weakly correlated with levels of qualifications. Table 4 confirms both predictions. We build a matrix of 4-digit SOC bins and show pairwise correlations of SOC-level average job complexity (Panel A) or shares of 'high-skill' jobs (Panel B) against that SOC bin's share of graduates, share of post-graduates, share of PHDs and mean workforce experience in years. Both panels show a negative link between workers' Diffbot skills and the share of graduates in a given job, weaker positive correlations with postgraduate and

PHD shares, and strong positive correlations with workforce experience.

Table 4. Diffbot skills benchmarking

Variables	(1)	(2)	(3)	(4)	(5)
Panel A. Skill complexity					
(1) Average Skill complexity	1.000				
(2) Share of graduates	-0.360***	1.000			
(3) Share of postgraduates	0.047	0.164***	1.000		
(4) Shares of PhD	0.210***	-0.136***	0.419***	1.000	
(5) Average worker experience	0.350***	-0.068	-0.008	-0.003	1.000
Panel B. Skill jobs					
(1) Share of high-skilled jobs	1.000				
(2) Share graduates	-0.236***	1.000			
(3) Share of postgraduates	0.066	0.164***	1.000		
(4) Shares of PhD	0.234***	-0.136***	0.419***	1.000	
(5) Average worker experience	0.243***	-0.068	-0.008	-0.003	1.000

Notes: This table reports correlations between variables constructed at the SOC4 bin level. *Source:* Diffbot.

5 Design

Our main analysis has three steps. Using the company panel, we first explore the overall association between the presence of migrant workers and firm productivity. Second, using the individual cross-sectional data, we explore skill differences between migrants and natives. Third, building on those results, we construct measures of job specialisation by nativity, and use the company panel to test whether the presence of specialized skilled migrants is associated with innovation.

5.1 Firm productivity and migrant workers

To measure the relationship between migrant presence and firm productivity, we regress firm TFP growth on workforce migrant share, controls and fixed effects. Our baseline specification for company i, sector j, area a and year t is as follows:

$$Y_{ijat} = b_1 + b_2\, MIG_{it-1} + \mathbf{X}_{ijat-1} + I_{ija} + T_t + e_{ijat} \tag{3}$$

where Y_{ijat} is the annual change in log total factor productivity and MIG_{it-1} is the es-

timated share migrants in the firm in the previous year. $\mathbf{X}_{ijat-1}$ is an array of lagged company-level controls, including size, financials, corporate structure and workforce characteristics, notably share of graduates and workforce age distribution, as well as the firm's Diffbot coverage ratio. In extensions, we estimate separate parameters for skilled migrants and natives, and for migrant and native years of experience. No companies switch industry or location in our panel, consistent with the focus on medium / large businesses, so I_{ija} denotes a firm-industry-area fixed effect. When including firm fixed effects, industry and region dummies are absorbed. This parameter absorbs a great deal of variation in our firm panel.

Our parameter of interest is b_2. This combines any direct effect of migrant quality with indirect effects from spillovers on the production function. In eq. 3 this parameter is not causally identified. Unobservables, predominantly via imperfect workforce coverage, may bias our estimates. We aim to minimize bias from this source primarily through robustness tests where we restrict our sample to high-coverage observations, as well as through weighted regressions that penalise low-coverage companies.

In addition, more productive firms may hire more migrants, while a third, unobserved factor may determine both firm productivity and the migrant share. Typically such issues may produce upward bias in OLS estimates. To help mitigate these concerns we first run placebo checks: shuffling values of MIG_{it-1}, and reversing eq. 3 to estimate MIG_{it} on deep lags of firm TFP. To further address this issue, building on Balsmeier et al. (2025) and Burchardi et al. (2018), we instrument for migrant share using historical name settlement patterns. The basic structure of the firm-level instrument is a Bartik shift-share. In our setting, the share is based on the firm's initial distribution of what we term 'UK-established' surnames, and the shift is the change over time in the prevalence of those UK-established surnames in the whole sample relative to a base year. The intuition behind the instrument is that names and name families exhibit deep historical persistence (Barone and Mocetti, 2021; Clark and Cummins, 2014; Clark et al., 2015). Specifically, networks of names and surnames have clear geographies that persist over multiple generations (Balsmeier et al., 2025; Kandt et al., 2020; Mateos et al., 2011). Prior papers also show that surname information can

be used as an alternative proxy of ethnicity when direct measures are unavailable (Mateos et al., 2011).

This implies that workers with established UK surnames are, on average, more likely to be native, whereas workers without a UK-established surname are more likely to be migrants. On the other hand, the shift component relies on the premise that firms with a higher initial share of migrant workers tend to hire more migrants in subsequent periods when migrant labour supply increases; analogously, firms with an initially higher share of natives tend to hire more UK natives. We argue that our instrument satisfies the exclusion restriction because the initial allocation of workers to firms by surname is orthogonal to firm's economic conditions. Put differently, conditional on the overall distribution of UK-established surnames, having a relatively larger share of workers named, for example, "Jones" rather than other UK-established surnames, should be unrelated to firm characteristics correlated with productivity.

To classify surnames as UK-established, we rely on historical surnames from the 1881 UK Census data, extracted from the UK Data Archive (Woollard and Schurer, 2000). Specifically, we merge our sample of individuals' surnames in each firm's initial year with the surnames identified in 1881 UK Census, which gives us information on every household living in England and Wales on the night of Sunday 3 April, 1881. We focus on the 1881 census, because this is the English census that is most carefully digitized (Clark et al., 2015).

We obtain 62,889 distinct surnames across firms in their initial year, out of which we match 45.6% to the 1881 UK Census. We identify the matched surnames as 'UK-established names', while the other surnames are considered 'non-UK established'. We use this terminology because surnames in the 1881 Census will also include migrants moving to the UK at some point. Given the focus of our study, this is not a limitation, since it helps us to distinguish those surnames rooted in the UK from those individuals who are more likely to migrate to the UK in more recent migration waves. Nonetheless, the instrument will be noisy, for at least two reasons. First, because it assigns migrant status to UK-born descendants of post-1881, pre-sample arrivals, giving an upper bound on the true migrant

share. For example, UK-born descendants of Jewish arrivals from the 1930s, Polish settlement post-WWII, and Commonwealth migrants from the 1950s onwards will be deemed migrants. Second, it will mis-ascribe in-sample Anglophone migrants who happen to have historic UK surnames: for example, those from the US, Canada and Australasia, as well as Caribbean countries. This will give a lower bound on the true migrant share. Despite these challenges, the raw correlation between the share of non-UK established names and the share of migrants at the firm level is about 0.59.

For the company panel, we then build the following shift-share instrument:

$$\widehat{\text{MIG}}_{ft} = 1 - \left(\sum_{s=1}^{S} \frac{\text{UK}_{s,f,t_0}}{N_{f,t_0}} \times \left(\text{UK}_{s,t} - \text{UK}_{s,t_0} \right) \right) \tag{4}$$

where UK_{s,f,t_0} is the number of workers with historic UK surnames s in firm i in year t_0, while N_{f,t_0} is the number of unique surnames in firm i in year t_0. The term $\text{UK}_{s,t} - \text{UK}_{s,t_0}$ is the relative change of individuals with surname s between year t and year t_0.

5.2 Migrant / native differences

The framework in Section 2 suggests that migrants may be positively selected on human capital (Hunt and Gauthier-Loiselle, 2010) and/or that migrants may improve workforce cognitive diversity by enlarging the pool of backgrounds and experiences (Page, 2007). We use the company panel to explore these channels at workforce level. We shift to cross-sectional individual level data to work in a richer setting, and use a variant of the the historic surnames instrument for identification.

To test whether migrants and natives carry distinct skills, we run the following cross-sectional regression for individual i in SOC4 job bin o:

$$\text{SKILL}_{io} = a + b\,\text{MIGRANT}_{io} + \mathbf{X}_{io}\,c + e_{io} \tag{5}$$

Where SKILL is the dominant topic $t = 1, 25$ held by the individual, MIGRANT_{io} is a dummy for migrant status and $\mathbf{X}_{io}$ is a set of individual-level controls, including the year the worker is observed (to take account of coverage differences across years). Note that

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. A significant b implies that, relative to natives, migrants contribute different dominant topics.

Section 4 also shows a mixed picture on migrant / native human capital selection. To formalise this, we compare ISCO-complexity of Diffbot skills for migrants vs. natives doing the same types of jobs. We define role types as bundles of SOC4 occupations. We estimate for worker i, role type k, SOC4 bin o:

$$\text{COMPLEXITY}_{iok} = a + b\,\text{MIGRANT}^{\text{k}}{}_{io} + \mathbf{X}_{iok}\,c + e_{iok} \tag{6}$$

Here COMPLEXITY_{iok} is the ISCO-complexity of worker i's dominant topic (ranked 1-25, where 25 is the most complex); $\text{MIGRANT}^{\text{k}}{}_{io}$ is a dummy for migrants in k roles, and the omitted category is natives in k roles. $\mathbf{X}_{iok}$ is a set of individual-level controls. Significant b implies that migrants in o carry more complex skills than natives in k. ==As before, we interpret effect sizes using standard deviations as units.== In line with our framework, we run regressions for all migrants vs natives, and for those in tech, STEM and managerial roles.

The main identification challenge in equations 4 and 5 stems from worker unobservables. As a first pass we run Oster tests and a placebo check that randomizes topics across workers. We also fit a version of the historic names instrument to proxy for migrant status. The individual IV for worker i is given by:

$$\widehat{\text{MIGRANT}}_i = 1 - \text{UK}_i \tag{7}$$

Where UK_i is a dummy for workers with UK-established surnames. The raw correlation of these two variables is around 0.3, significant at 1 percent. We argue that for the workers in our data, the exclusion restriction is met because, conditional on observables, not having a UK-established surname is unlikely to be correlated with worker skills except through migrant status.

5.3 Migrant specialisation and firm performance

Consistent with our framework, in our data migrants and natives differ in levels and types of qualifications, experience and skills. This may allow them to specialise in different tasks and jobs, and in turn, this may influence firm-level outcomes (Peri and Sparber, 2009; Foged and Peri, 2016; Peri and Sparber, 2011; Lin, 2019; Mayda et al., 2022).

Skilled worker models in Mayda et al. (2022) and Peri and Sparber (2011) suggest skilled migrants will tend to specialise into technical roles (reflecting technical skills) and skilled natives into managerial roles (reflecting language skills and social capital). In our data, we should observe migrants' relative specialisation in technical roles, and potentially on relevant firm outcomes - notably innovation.

Building on Mayda et al. (2022), we first define the migrant/native ratio (MNR) as the share of migrants over the share of natives in three types of job: technical, STEM or managerial roles defined per Section 4.2.

The MNR in technical roles for firm i, year t is given as:

$$\text{MNRtech}_{it} = \frac{\text{Mtech}_{it}}{\text{Ntech}_{it}} \tag{8}$$

Where Mtech_{it} is the count of migrants doing technical jobs, and Ntech_{it} is the native equivalent. MNRs for STEM, which include technical and related management tasks, and management alone, are defined analogously. To explore whether firms with higher migrant workforce shares exhibit more job specialisation, we estimate for company i, sector j, area a and year t:

$$\text{MNR}^{\text{t}}{}_{it} = b_1 + b_2\, MIG_{it-1} + \mathbf{X}_{ijat-1} + I_{ija} + T_t + e_{ijat} \tag{9}$$

Where t denotes either the firm's technical, STEM or managerial MNR. Comparing estimates of b_2 across MNRs tells us the relative extent of migrant-native specialisation across technical / managerial domains.

Firms vary in their underlying occupational structures. To normalise for this, per Mayda et al. (2022) we then define a firm's level of *relative* migrant specialisation in, say, technical

over managerial roles, as the ratio of each MNR. Specifically, tech specialisation is given as:

$$\text{MStech}_{it} = \frac{\text{MNRtech}_{it}}{\text{MNRman}_{it}} \tag{10}$$

Where MNRtech_{it} is migrant-native specialisation in tech roles, and MNRman_{it} is the migrant-native ratio in managerial roles. To look at links from migrant specialisation to firm outcomes, we plug these specialisation measures into equation 3. We include TFP as a dependent variable, alongside patent counts and total citations for patenting firms.

6 Productivity results

6.1 Main results

Table 5 shows OLS and IV results from equation (3), which estimates the TFP-migrant share link for the company panel. Column 1 runs a simple specification with year and area dummies, as well as their interactions. Column 2 adds workforce and firm controls. Columns 3-6 add industry fixed effects at progressively higher levels of detail, from 1-digit to 4-digit levels. Column 7 includes the firm-industry-area fixed effect. Column 8 shows IV results.

Overall, we find a small positive association between firm migrant share and TFP growth, which becomes very small once we control for industry and firm heterogeneity.[8] Notably, the coefficient of b_2 more or less triples from the naive OLS specification in Column 1 to the full OLS specification in Column 7. IV estimates are larger and much less precisely estimated, consistent with the noisy character of the instrument. In column 8, our preferred specification, a 10 percent increase in firm migrant share is associated with an 0.00345 percent change in firm TFP growth.

Appendix Table E7 fits an alternative specification in levels. Here we find a positive significant association of migrant share that becomes very small and non-significant in the

[8] Appendix Table E5 gives detailed results including coefficients of controls; Table E6 shows a correlation matrix of controls.

Table 5. Main results

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS	(7) OLS	(8) IV
Firm migrant share	0.0178	0.0236	0.0225	0.0179	0.0149	0.0126	0.0687	0.345
	(0.0176)	(0.0182)	(0.0183)	(0.0183)	(0.0186)	(0.0187)	(0.0575)	(1.098)
Region-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	No	No	1-digit	2-digit	3-digit	4-digit	Absorbed	Absorbed
Firm FE	No	No	No	No	No	No	Yes	Yes
N firms	6,373	6,373	6,373	6,373	6,373	6,373	6,373	6,373
Observations	46,906	46,906	46,906	46,906	46,906	46,906	46,906	46,501
R-squared	0.00786	0.0105	0.0113	0.0145	0.0178	0.0211	0.177	0.0256
First stage coefficient	-	-	-	-	-	-	-	-0.002
K-P F-stat	-	-	-	-	-	-	-	10.71

Notes: The dependent variable is TFP growth. The estimation model is indicated in the column header. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

firm fixed effects specification. IV estimates are much larger than OLS estimates, and again are non-significant.

Similarly to Hall and Manning (2024) we cannot identify a significant causal link from migrant worker share to productivity, but neither can we rule this out. Overall, these results are consistent with our framework, in which the migrant share-TFP link is a priori ambiguous, and regressions on pooled data estimate the net effect of positive and negative channels, sectoral, occupational and workforce heterogeneity.

6.2 Robustness

Appendix Tables E8-E10 run our main result through an array of robustness tests. Our main story survives all of these checks.

Table E8 includes a series of sensitivity checks. Columns 1-4 fit alternative fixed effects specifications and sample periods. Columns 5-8 use subsamples with higher coverage ratios, or penalise observations with lower coverage rates. Columns 9-10 remove outliers.

Table E9 runs more structural checks for the influence of unobserved workers in our data. Panel A runs a Lee Bounds - style exercise, where for each firm-year cell, we assign

all workers unobserved by Diffbot as either migrants or natives.[9] Panel B estimates an alternative sample keeping only workers where we observe migrant / native status. Neither test changes our main results.

Table E10 runs a placebo test where we reverse equation (3), and regress migrant share on deep lags of firm TFP. Coefficients of TFP are non-significant, implying there are no common trends driving both firm productivity and the migrant share. Table E11 runs a second placebo test where we randomise values of the migrant share. Coefficients are non-significant, again consistent with the absence of omitted variables driving our main relationship.

6.3 Extensions

So far, our results pool all workers and firm types. Below we run a series of extensions on equation (3) where we leverage variation across industries, workforce occupational mix, workforce qualifications and experience, and workforce birth country and linguistic mix. Again, we find so significant associations in these regressions, so true effects may be zero. However, we cannot definitively rule productivity links in or out using our firm-level data.

Our framework implies that migrant productivity links may be stronger in more 'knowledge-intensive' sectors. Table E12 reproduces our analysis, interacting the migrant share with dummies for manufacturing, services, ONS-defined 'science and technology' sectors and 'high-tech' sectors as defined by NESTA (Bakhshi et al., 2015; Harris, 2015). Although some interaction terms are positive, none of these results is significant.

Our framework also implies that migrant - productivity links may be stronger in knowledge-intensive occupations and settings. In the pooled worker data, migrant shares of tech and STEM workers are higher than native shares; however, in the average firm, native shares of tech and STEM workers are higher than their migrant counterparts. Table E13 explores these channels at firm level. Column 1 runs our main result. Columns 2-3 plug in shares of migrants in tech or STEM occupations. Columns 4-5 interact firm migrant shares with a dummy for firms with above-mean share of STEM or tech workers; columns 6-7 interact

[9] Specifically, we calculate the number of unobserved workers per firm-year as the difference between employment counts in Orbis and workers in Diffbot.

firm migrant share with the continuous firm-level share of STEM or tech workers. OLS coefficients for cols 4-5 are large and significant at 1 percent, although variants with continuous STEM / tech shares are not. IV regressions, not shown here, are all non-significant.

Next we look at workforce-level migrant/native differences in qualifications and experience. Table **??** shows results. We fit the share of high-skilled migrants (defined as those with degrees or above), interactions of the migrant share with firms' high-skill intensity, and regressions including both migrant share and average migrant experience. All results are non-significant.

7 Worker skills results

Company-level regressions find small, non-significant effects of migrant share on productivity, which are explained by worker sorting across industries and firms. Consistent with our framework, this link varies by broad industry type, occupation mix, and workforce qualifications and experience, but industry and firm heterogeneity again explains the results. Our company-level data cannot precisely pin down linkages. Here we turn to the individual-level data to explore linkages in a richer cross-sectional setting.

7.1 Distinct skills

Our raw data shows differences in migrant and native qualifications, years of experience and Diffbot skills, which we use as proxies for capabilities developed during careers. We first explore whether migrants and natives have distinct Diffbot skills. Table 6 shows results for equation (4), which compares dominant skills topics for migrants and UK-born workers doing the same job (defined by detailed 4-digit SOC bin), and controlling for a range of observable characteristics, including qualifications and years of experience. Recall that in our raw data (Table 2) the average migrant and native have close but distinct dominant skills. Column 1 shows the OLS result, and finds a robust but very small difference. $\hat{b2}$ is 0.155, around 2 percent of a standard deviation. Oster tests suggest unobservables are very unlikely to explain our main result. Column 2 shows reduced form estimates using the individual historic names instrument. Coefficients are larger and significant, over 4 percent

of a standard deviation. IV estimates in column 3 are much larger, around 20 percent of a standard deviation. The R^2 for IV estimates is negative, however, suggesting very poor fit, and so we prefer the reduced form results here. Column 4 shows the placebo test: the coefficient is close to zero and non-significant.

Table 6. Diffbot distinctive skills test

	(1) OLS	(2) Reduced form	(3) IV	(4) Placebo
Migrant	0.155***		1.416***	-0.00197
	(0.0258)		(0.0886)	(0.0277)
Predicted migrant based on historic name		0.378***		
		(0.0235)		
Observations	437,630	437,630	437,630	402,096
R^2	0.122	0.122	-0.00154	0.00112
Oster delta / migrant	8.849			
Underidentification test			24053.8	
Under-identification test p-value			0	
K-P F-stat			27344.7	
First stage coefficient			0.267***	

Notes: 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). Column (1) fits OLS. Column (2) fits reduced form, using predicted migrant status based on historic name info. Column (3) uses predicted migrant status as an instrument. Column (4) runs a placebo check which shuffles the dominant topic count across workers. "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot.

Table E14 shows sensitivity and robustness checks: our result is robust to alternative estimators, time windows and alaternate specifications of workers' dominant skills.

7.2 Complex skills

We next turn to differences in skill levels. Our metric proxies this by relating key terms in each topic to descriptors in the ISCO hierarchy producing a ranking by ISCO-complexity. Table 7 gives results for equation (5) where we compare all migrants versus natives. We interpret the relative size and sign of these coefficients as telling us the correlates of migrant

versus UK-born workers in these jobs. Migrant status significantly predicts carrying more complex Diffbot skills, conditional on qualifications, experience, other observables, SOC4 dummies and year of observation.

Table 7. Diffbot complex skills test: migrants vs natives

	(1) OLS	(2) Reduced form	(3) IV	(4) Placebo
Migrant	0.623***		2.019***	0.0113
	(0.0256)		(0.0893)	(0.0276)
Predicted migrant based on historic name		0.539***		
		(0.0238)		
Observations	437,630	437,630	437,630	402,243
R^2	0.241	0.241	0.0562	0.00105
Oster delta / migrant	5.230			
Underidentification test			24069.2	
Under-identification test p-value			0	
K-P F-stat			27377.2	
First stage coefficient			0.267***	

Notes: Sample is workers with Diffbot 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 per Table 6. "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates.

Table 8 tightens focus, comparing ISCO-complexity for migrants in tech, STEM or managerial roles, versus natives in those roles. In all three cases, we find evidence of migrant selection on skills. Effects are most robust for migrants in tech and STEM roles: while IV estimates for migrants in managerial roles are very large, the model fit is essentially zero.

Table 8. Diffbot complex skills test: migrants vs natives in tech, STEM and managerial roles

	(1) OLS	(2) Reduced form	(3) IV	(4) OLS	(5) Reduced form	(6) IV	(7) OLS	(8) Reduced form	(9) IV
Migrant in tech occupation	0.895***		1.483***						
	(0.0812)		(0.221)						
Pr(migrant) in tech occupation		0.525***							
		(0.0783)							
Migrant in STEM occupation				0.795***		1.201***			
				(0.0737)		(0.199)			
Pr(migrant) in stem occupation					0.427***				
					(0.0709)				
Migrant in managerial occupation							0.625***		2.634
							(0.0466)		(0.17
Pr(migrant) in managerial occupation								0.669***	
								(0.0443)	
Observations	37224	37224	37224	43532	43532	43532	138052	138052	1380
R^2	0.215	0.214	0.00749	0.278	0.277	0.0102	0.190	0.191	-0.007
Oster delta / migrant	3.396			3.125			2.006		
Underidentification test			3513.7			4244.6			6091
Under-identification test p-value			0			0			0
K-P F-stat			4251.6			5134.9			6919
First stage coefficient			0.354***			0.356***			0.254

Notes: Sample is workers with Diffbot skills and observables. Regression compares the ISCO-complexity of skills for migrants versus migrants and natives doing the same role types. All specifications include year dummies, SOC4 dummies and controls per Table 6. "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates.

Table E15 gives sensitivity and robustness checks for migrants in tech roles. Columns 1 and 2 give OLS and IV specifications. Our IV result is robust to alternative LLM classifiers (columns 3-4), time windows (columns 5-10) and a placebo test as before (column 11).

Overall, we find that controlling for qualifications and years of experience, migrants and natives in the same roles bring slightly different Diffbot skills, and that migrants in tech and STEM roles, in particular, are selected on skills compared to natives in those roles. In the final part of the paper we build on these results by exploring the role of migrant tech and STEM specialisation on company innovation and TFP.

8 Migrant specialisation and firm performance

Our raw worker data, and worker-level regressions, show that migrants and natives differ on qualifications, experience and learned skills; and that migrants in tech and STEM roles are selected on Diffbot skills, relative to those in management roles. This suggests imperfect substitution between migrants and UK-born, UK-born comparative advantage in managerial roles, and the relative specialisation of migrants into technical roles.

8.1 Specialisation rates

Table 8 formally tests the extent of migrant specialisation using our company panel and firm fixed effects. We build company-year migrant-native ratios (MNRs) for tech, STEM and managerial roles per equation (8).

We condition our existing regression sample on companies where we observe MNRs for all three types, giving us around 24,600 company-year cells (and around 900 cells for the much smaller set of patenting companies). Table E16 gives summary statistics for these sub-samples. Per equation (9) first test whether the variation in firms' migrant share is associated with higher migrant-native ratios in these role types. (In this version of the paper we run OLS regressions with firm fixed effects; in future versions we will include IV specifications.)

Table 9 gives results. Columns 1-3 show coefficients of the migrant share on MNRs for tech roles, STEM roles and managerial roles respectively. Columns 4-6 repeat the analysis

for the share of migrants with degrees or above. Columns 7-9 restrict to the subsample of patenting firms. In all cases, a higher migrant share robustly links to a higher MNR, but coefficients are bigger for tech and STEM roles. This implies that firms add migrants faster into tech and STEM than managerial roles, especially for high-skilled migrant workers. Sensitivity checks for the subsample of patenting companies, available on request, give essentially identical results.

Table 9. Role specialisation and firm migrant share

	(1) Tech	(2) STEM	(3) Mgt	(4) Tech	(5) STEM	(6) Mgt
Firm migrant share	1.501***	1.535***	1.430***			
	(0.166)	(0.183)	(0.144)			
Firm high-skilled migrant share				1.606***	1.610***	1.456***
				(0.184)	(0.209)	(0.157)
N companies	3455	3445	5235	3455	3445	5235
Observations	21521	21623	35094	21521	21623	35094
R^2	0.759	0.778	0.837	0.759	0.778	0.836

Notes: Regressions estimate firms' share of migrants/natives in tech, STEM and management roles on lagged firm migrant share. All regressions fit controls, area, region, year and firm FE. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). All explanatory variables are 1-period lagged. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates.

8.2 Specialisation effects

Next we turn to the link between migrant specialisation and firm outcomes. Here, we run equation (3) fitting lagged migrant specialisation as well as the lagged migrant share, plus controls and firm fixed effects.[10] Table 10 summarises our results. Panel A gives results for firm TFP. Panel B gives results for patent counts, which we estimate using a PPML estimator. Panel C fits an LPM to uncover the extensive margin (comparing non-patenting and patenting firms). Panel D fits a count model on the intensive margin (the subsample of firms with at least one patent).

[10]Results excluding the migrant share are identical.

We find marginally significant links from migrant specialisation to firm productivity, with coefficients close to zero (Panel A). Conversely, we find a robust positive association from migrant tech specialisation to subsequent patent counts (Panel B). This effect is driven by higher patenting in the subset of actively patenting larger firms (Panels C-D). Tables E18-E19 report placebo tests for the migrant tech specialisation result. In Table E18 we regress specialisation on deep lags of firm patenting. In Table E19 we randomise migrant tech specialisation across firms. Results survive both tests.

These significant links from migrant specialisation to patent counts do not appear to feed through to patent quality. Table E20 reports results for all-time citations. All coefficients are non-significant.

Finally, we look at the relative roles of migrant tech specialisation, migrant qualifications and experience, and company-level birthplace diversity for this subsample of firms. Results are given in E21. As we progressively include measures of other channels, the coefficient of migrant tech specialisation reduces slightly, from 0.115 to 0.110, but remains significant at 5 percent. All other channel variables are non-significant (Columns 1-5). Results for citations remain unchanged (Column 6).

9 Conclusion

In this paper we explore the effects of migrant workers on firms' innovation and productivity, using novel, highly granular data to shed new light on impacts and mechanisms. We develop a large, novel worker-firm dataset for the UK 2007-2023 that links web and administrative sources. For identification we use instruments based on historic name settlement patterns. We have three main findings. We find a positive but non-significant link from firms' overall migrant share to TFP. Since this does not rule out an effect, we use our granular worker-level data to explore potential migrant-productivity channels in more detail. We find that migrant workers carry distinctive, and more complex learned capabilities than natives in the same jobs, particularly in tech and STEM roles. We also find that migrant worker specialisation into technical roles is linked to higher firm patent counts.

Table 10. Migrant specialisation and firm outcomes

	(1)	(2)	(3)
A: TFP			
Migrant tech role specialisation	-0.00416*		
	(0.00247)		
Migrant stem role specialisation		-0.00369	
		(0.00250)	
Migrant management role specialisation			0.0117*
			(0.00704)
N companies	2,243	2,232	1,574
Observations	12,912	12,968	8,809
R^2	0.952	0.951	0.943
B: Patenting			
Migrant tech role specialisation	0.115**		
	(0.0468)		
Migrant stem role specialisation		-0.00396	
		(0.0230)	
Migrant management role specialisation			-0.0617
			(0.0609)
N companies	189	188	148
Observations	987	990	762
Pseudo R^2	0.618	0.615	0.647
C. Patenting extensive margin			
Migrant tech role specialisation	0.000122		
	(0.000519)		
Migrant stem role specialisation		-0.000159	
		(0.000535)	
Migrant management role specialisation			-0.00184
			(0.00404)
N companies	2,243	2,192	1,506
Observations	12,912	12,619	8,384
R^2	0.499	0.497	0.510
D. Patenting intensive margin			
Migrant tech role specialisation	0.125***		
	(0.0461)		
Migrant stem role specialisation		-0.00150	
		(0.0243)	
Migrant management role specialisation			-0.0783
			(0.0500)
N companies	105	106	87
Observations	360	362	307
Pseudo R^2	0.619	0.615	0.654

Notes: Regressions estimate firm outcomes on migrant specialisation in tech, STEM or management roles. All regressions fit controls, area, region, year and firm FE. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). All explanatory variables are 1-period lagged. Robust standard errors in parentheses.

References

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Akcigit, U., Caicedo, S., Miguelez, E., Stantcheva, S., and Sterzi, V. (2018). Dancing with the stars: Innovation through interactions. *National Bureau of Economic Research Working Paper Series*, No. 24466.

Akcigit, U., Grigsby, J., and Nicholas, T. (2017). Immigration and the rise of american ingenuity. *American Economic Review*, 107(5):327–31.

Alesina, A. and Ferrara, E. L. (2005). Ethnic diversity and economic performance. *Journal of Economic Literature*, 43(3):762–800.

Alesina, A., Harnoss, J., and Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, 21(2):101–138.

Arora, A., Belenzon, S., and Sheer, L. (2021). Matching patents to compustat firms, 1980–2015: Dynamic reassignment, name changes, and ownership structures. *Research Policy*, 50(5):104217.

Arora, A. and Dell, M. (2023). Linktransformer: A unified package for record linkage with transformer language models. Report, arxiv.

Autor, D., Dorn, D., Hanson, G. H., Pisano, G., and Shu, P. (2020). Foreign competition and domestic innovation: Evidence from us patents. *American Economic Review: Insights*, 2(3):357–374.

Azoulay, P., Jones, B., Kim, J. D., and Miranda, J. (2022). Immigration and entrepreneurship in the united states. *American Economic Review: Insights*, 4(1):71–88.

Babina, T., Fedyk, A., He, A. X., and Hodson, J. (2022). Artificial intelligence, firm growth and product innovation. *Journal of Financial Economics*.

Bakhshi, H., Davies, J., Freeman, A., and Higgs, P. (2015). The geography of the uk's creative and high-tech economies. Report, NESTA.

Balsmeier, B., Fleming, L., Marx, M., and Shin, S. R. (2025). Startups, unicorns, and the local influx of inventors. *The Review of Economics and Statistics*, pages 1–44.

Barone, G. and Mocetti, S. (2021). Intergenerational mobility in the very long run: Florence 1427–2011. *The Review of Economic Studies*, 88(4):1863–1891.

Berliant, M. and Fujita, M. (2012). Culture and diversity in knowledge creation. *Regional Science and Urban Economics*, 42(4):648–662.

Boberg-Fazlić, N. and Sharp, P. (2023). Immigrant communities and knowledge spillovers: Danish-americans and the development of the dairy industry in the united states. *American Economic Journal: Macroeconomics*, Forthcoming.

Burchardi, K. B., Chaney, T., and Hassan, T. A. (2018). Migrants, ancestors, and foreign investments. *The Review of Economic Studies*, 86(4):1448–1486.

Campo, F., Forte, G., and Portes, J. (2024). The impact of migration on productivity: Evidence from the united kingdom. *The B.E. Journal of Economic Analysis Policy*, 24(2):537–564.

Cheng, W. and Weinberg, B. A. (2021). Marginalized and overlooked? minoritized groups and the adoption of new scientific ideas. Report, NBER.

Clark, G. and Cummins, N. (2014). Surnames and social mobility in England, 1170–2012. *Human Nature*, 25(4):517–537.

Clark, G., Cummins, N., Hao, Y., and Vidal, D. D. (2015). Surnames: A new source for the history of social mobility. *Explorations in Economic History*, 55:3–24.

Commission, O. . E. (2023). ndicators of immigrant integration 2023: Settling in. Report, OECD Publishing.

Cooke, A. and Kemeny, T. (2017). Cities, immigrant diversity, and complex problem solving. *Research Policy*, 46(6):1175–1185.

Cuibus, M., Walsh, P. W., and Němeček, F. (2025). Student migration to the uk. Report, Migration Observatory.

Dahlke, J., Schmidt, S., Lenz, D., Kinne, J., Dehghan, R., Abbasiharofteh, M., Schütz, M., Kriesch, L., Hottenrott, H., Kanilmaz, U. N., Grashof, N., Hajikhani, A., Liu, L., Riccaboni, M., Balland, P.-A., Wörter, M., and Rammer, C. (2025). The webai paradigm of innovation research: Extracting insight from organizational web data through ai. Report, ZEW.

Dale-Olsen, H. and Finseraas, H. (2020). Linguistic diversity and workplace productivity. *Labour Economics*, 64:101813.

De Loecker, J., Obermeier, T., and Van Reenen, J. (2024). Firms and inequality. *Oxford Open Economics*, 3(Supplement$_1$) : $i962 - -i982$.

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

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: a web-scale approach to probabilistic knowledge fusion. Report, Association for Computing Machinery.

Doran, K., Gelber, A., and Isen, A. (2022). The effects of high-skilled immigration policy on firms: Evidence from visa lotteries. Report, IZA.

Dorn, D., Schoner, F., Seebacher, M., Simon, L., and Woessmann, L. (2025). Multidimensional skills on linkedin profiles: Measuring human capital and the gender wage gap. Report, Rockwool Foundation.

Exadaktylos, D., Riccaboni, M., and Rungi, A. (2024). Talents from abroad. foreign managers and productivity in the united kingdom. *International Economics*, 177:100474.

Fabling, R., Maré, D. C., and Stevens, P. (2022). Migration and firm-level productivity. Report, IZA.

Fedyk, A. and Hodson, J. (2022). Trading on talent: Human capital and firm performance*. *Review of Finance*.

Ferrucci, E. and Lissoni, F. (2019). Foreign inventors in europe and the united states: Diversity and patent quality. *Research Policy*, 48(9):103774.

Foged, M. and Peri, G. (2016). Immigrants' effect on native workers: New analysis on longitudinal data. *American Economic Journal: Applied Economics*, 8(2):1–34.

Gagliardi, L., Mariani, M., and Breschi, S. (2024). Temporal availability and women career progression: Evidence from cross-time-zone acquisitions. *Organization Science*, 35(6):2178–2197.

Glover, J. and Kim, E. (2021). Optimal team composition: Diversity to foster implicit team incentives. *Management Science*, 67(9):5800–5820.

Gray, S., Kemeny, T., Nathan, M., Ozgen, C., Pialli, G., Rosso, A. C., Reades, Jon, S. M., and Valero, A. (2025). Graph-ee: Building employer - employee panels from a knowledge graph and company microdata. *Working Paper*.

Hall, T. and Manning, A. (2024). Only human? immigration and firm productivity in britain. Report, LSE.

Harris, J. (2015). Identifying science and technology businesses in official statistics. Report, ONS.

HESA (2024). Higher education student statistics: Uk, 2022/23 - where students come from and go to study. Report, HESA.

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., and McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291.

Hunt, J. and Gauthier-Loiselle, M. (2010). How much does immigration boost innovation? *American Economic Journal: Macroeconomics*, 2(2):31–56.

Jeffers, J. (2024). The impact of restricting labor mobility on corporate investment and entrepreneurship. *Review of Financial Studies*, 37(1):1–44.

Jin, W., Huang, Y., and Zhu, S. (2025a). Birthplace diversity of immigrants and technological novelty: Unpacking the internal structure of diversity. *Technological Forecasting and Social Change*, 221:124337.

Jin, Z., Kermani, A., and McQuade, T. (2025b). Native-immigrant entrepreneurial synergies. Report, NBER.

Kalemli-Özcan, Ş., Sørensen, B. E., Villegas-Sanchez, C., Volosovych, V., and Yeşiltaş, S. (2024). How to construct nationally representative firm-level data from the orbis global database: New facts on smes and aggregate implications for industry concentration. *American Economic Journal: Macroeconomics*, 16(2):353–374.

Kandt, J., van Dijk, J., and Longley, P. A. (2020). Family name origins and intergenerational demographic change in great britain. *Annals of the American Association of Geographers*, 110(6):1726–1742. doi: 10.1080/24694452.2020.1717328.

Kemeny, T. (2017). Immigrant diversity and economic performance in cities. *International Regional Science Review*, 40(2):164–208.

Kerr, W. (2008). Ethnic scientific communities and international technology diffusion. *Review of Economics and Statistics*, 90(3):518–537.

Kerr, W. and Lincoln, W. (2010). The supply side of innovation: H-1b visa reforms and u.s. ethnic invention. *Journal of Labor Economics*, 28(3):473–508.

Lee, K. M., Kim, M. J., Brown, J. D., Earle, J. S., and Liu, Z. (2025). Are immigrants more innovative? evidence from entrepreneurs. *Journal of Economics Management Strategy*, n/a(n/a).

Lee, S. and Glennon, B. (2023). The effect of immigration policy on founding location choice: Evidence from canada's start-up visa program. Report, NBER.

Levine, S. S., Apfelbaum, E. P., Bernard, M., Bartelt, V. L., Zajac, E. J., and Stark, D. (2014). Ethnic diversity deflates price bubbles. *Proceedings of the National Academy of Sciences*, 111(52):18524–18529.

Lin, G. C. (2019). High-skilled immigration and native task specialization in u.s. cities. *Regional Science and Urban Economics*, 77:289–305.

Liu, T., Mao, Y., and Tian, X. (2023). The role of human capital: Evidence from corporate innovation. *Journal of Empirical Finance*, 74:101435.

Mack, D. Z., Chen, G., Hsu, P.-H., Lee, Y. T., and George, G. (2025). Interfaces, social information processing, and diversity cascades: How board diversity influences invention output. *Research Policy*, 54(1):105148.

Manning, A. (2025). *Why Immigration Policy is Hard: And How To Make It Better*. Polity, Cambridge.

Mateos, P., Longley, P., and O'Sullivan, D. (2011). Ethnicity and population structure in personal naming networks. *PLoS ONE*, 6(9):doi:10.1371/journal.pone.0022943.

Mayda, A. M., Orefice, G., and Santoni, G. (2022). Skilled immigration, task allocation and the innovation of firms. Report, Ifo.

Mesquita, F., Cannaviccio, M., Schmidek, J., Mirza, P., and Barbosa, D. (2019). Knowledgenet: A benchmark dataset for knowledge base population. Report, Association for Computational Linguistics.

Mitaritonna, C., Orefice, G., and Peri, G. (2017). Immigrants and firms' outcomes: Evidence from france. *European Economic Review*, 96:62–82.

Nam, H. and Portes, J. (2023). Migration and productivity in the uk: An analysis of employee payroll data. Report, IZA.

Nathan, M. (2015). Same difference? minority ethnic inventors, diversity and innovation in the uk. *Journal of Economic Geography*, 15(1):129–168.

Nathan, M. and Rosso, A. (2022). Innovative events: product launches, innovation and firm performance. *Research Policy*, 51(1):104373.

OECD (2024). International migration outlook 2024. Report, OECD Publishing.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297.

Ottaviano, G. and Peri, G. (2006). The economic value of cultural diversity: Evidence from us cities. *Journal of Economic Geography*, 6:9–44.

Ottaviano, G. I. P., Peri, G., and Wright, G. C. (2018). Immigration, trade and productivity in services: Evidence from u.k. firms. *Journal of International Economics*, 112:88–108.

Ozgen, C. (2021). The economics of diversity: Innovation, productivity and the labour market. *Journal of Economic Surveys*, 35(4):1168–1216.

Page, S. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools and Societies*. Princeton University Press, Princeton.

Parrotta, P., Pozzoli, D., and Pytlikova, M. (2014a). Labor diversity and firm productivity. *European Economic Review*, 66:144–179.

Parrotta, P., Pozzoli, D., and Pytlikova, M. (2014b). The nexus between labor diversity and firm's innovation. *Journal of Population Economics*, 27(2):303–364.

Paserman, D. (2013). Do high-skill immigrants raise productivity? evidence from israeli manufacturing firms, 1990-1999. *IZA Journal of Migration*, 2(6). IZA Discussion Papers.

Pellegrino, G., Penner, O., Piguet, E., and de Rassenfosse, G. (2023). Productivity gains from migration: Evidence from inventors. *Research Policy*, 52(1):104631.

Peri, G. (2012). The effect of immigration on productivity: Evidence from u.s. states. *Review of Economics and Statistics*, 94(1):348–358.

Peri, G. and Sparber, C. (2009). Task specialization, immigration, and wages. *American Economic Journal: Applied Economics*, 1(3):135–69.

Peri, G. and Sparber, C. (2011). Highly educated immigrants and native occupational choice. *Industrial Relations*, 50(3):385–411.

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., and Aral, S. (2022). A causal test of the strength of weak ties. *Science*, 377(6612):1304–1310.

Rock, D. (2019). Engineering value: The returns to technological talent and investments in artificial intelligence. Report, MIT.

Tambe, P., Hitt, L., Rock, D., and Brynjolfsson, E. (2020). Digital capital and superstar firms. Report, NBER.

Trax, M., Brunow, S., and Suedekum, J. (2015). Cultural diversity and plant-level productivity. *Regional Science and Urban Economics*, 53(July):85–96.

Widmann, R. (2023). Immigrant inventors and local income taxes: Evidence from swiss municipalities. *Journal of Public Economics*, 219:104822.

Wigger, C. (2021). Who with whom? untangling the effect of high-skilled immigration on innovation*. *Journal of Economic Geography*, 22(2):449–476.

Woollard, M. and Schurer, K. (2000). 1881 census for england and wales, the channel islands and the isle of man (enhanced version). [data collection].

Appendices

A Key datasets

A.1 Diffbot

Diffbot is the world's largest commercial knowledge graph, built from the public web. At the start of 2025, the graph included 278.9m active companies and 231.3m individuals in employment worldwide; in the UK the graph covered 4.7m active companies and 10.6m workers.[11] This compares to totals of 5.5m active firms in the UK Business Population Estimates, and 33.9m workers aged 16+ in the UK Labour Force Survey. ==[REFS]== In Appendix B we run further diagnostics benchmarking Diffbot against administrative data.

Knowledge graphs were an established tool in early AI research, and a key concept in early visions of the semantic web. Diffbot builds its graph by continually crawling the public web, identifying key elements on webpages, then using image recognition, natural language processing and supervised learning to build a graph of entities (such as people, organisations, places), their characteristics and relationships to each other (Mesquita et al., 2019). Specifically, Diffbot builds proprietary tools based on knowledge fusion algorithms, which use supervised learning to infer properties and linkages from high-quality sources in previous versions of the graph, ranking items using a confidence score (Dong et al., 2014).

In our case, Diffbot allows us to track a company and its workforce over time, as well as seeing an array of individual and firm-level characteristics. We exploit three features of Diffbot in particular. First, as the UK Company Register is provided as open data, the graph includes detailed UK company information, including Company Registration Number (CRN) identifiers. These identifiers allow us to link companies in Diffbot to those in the Register and to other company-level data, including commercial data products like Orbis Historical and Orbis IP. Second, Diffbot provides extremely rich information on individuals and companies, including detailed education and career histories, job titles and descriptions, skills, and companies' most likely key partners, suppliers and competitors. For workers, we use these to enhance the data, building proxies for seniority, migrant status, and to map individual skills. Third, like all web data, Diffbot's core sampling frame is implicit and

[11] See www.diffbot.com for more detail.

will not be structured like a conventional sample (Dahlke et al., 2025; Nathan and Rosso, 2022). Unusually, however, Diffbot's data provision is highly transparent, allowing us to see provenance and confidence scores for every element in the graph. This provides crucial insight into sampling frames and helps with validation.

Diffbot's graph updates every four to five days. This means that the characteristics of any sample may change slightly, depending on when it was extracted. This is - very broadly - equivalent to conventional data being revised by statistical authorities in subsequent editions, a common occurrence. We timestamp our data on the dates of query and subsequent extraction. In the final build, around 10 percent of companies have an observed workforce share above our 0.25 threshold at query stage, but 'coverage' falls below the threshold at extraction stage. See Gray et al. (2025) for details.

A.2 Orbis / Historical Orbis

Our starting sample includes all medium-sized and large companies according to Companies House (CH) definitions, active at some time between 2007 and 2023. Our sample starts in 2007 because Orbis Historical is not available before that date.

Companies House distinguishes between small and micro-entities from medium and large companies based on turnover, assets and employees thresholds. Specifically, for accounting periods beginning on or after 1 January 2016, a small company (and so a micro-entity) must meet at least two of the following conditions:

- Annual turnover more than £10.2 million;
- Balance sheet total more than £5.1 million;
- Average annual number of employees more than 50.

Companies above this threshold need to provide complete, audited annual accounts. This means that company-level information is most complete and highest quality for this sample. Applying these definitions to Orbis Historical, 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.

We follow the cleaning procedures for Orbis data documented in Kalemli-Özcan et al. (2024); De Loecker et al. (2024). 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 et al. (2024):

- 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.
- When annual report and local registry filing values are not the same (and both are non-missing), we prefer annual report values.
- When annual report values are missing, and local registry filings are non-missing, we keep local registry filings.
- After the selection above, we prefer consolidated accounts to unconsolidated accounts.
- We remove remaining duplicates by taking the observations with fewer missing values for the number of employees, EBITDA and costs of employees.

Each company profile includes a Bureau van Dijk identification number and a Companies House identifier, the Company Record Number (CRN). We use these to link companies in Orbis to organisations in Diffbot (via CRN), and to patent applicants in Orbis IP (via BvD number).

A.3 Orbis IP / PATSTAT

We match patent data to our company sample at the applicant level. In this version of the paper, we pull patents information from Orbis IP, which includes a BvD identification number for each patenting company. Our match rate is is about 7.5 percent. In tests,

we find this matching rate is superior to matching using PATSTAT Global using HAN identifiers or fuzzy matching with company names.

Our main variables are dummies for whether a firm patents in a given year, and counts of patents in a year for patenting firms. In building these variables we follow standard practices in the literature (Autor et al., 2020; Arora et al., 2021):

- We use patents applied to both local (UK) and worldwide patent offices (USPTO, EPO, JPO);
- 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 $\frac{1}{2}$ to each);
- We assign patents to the priority year of application, considered the closest date to the original invention.

B Diffbot validation - minus figures and tables in this version

Benchmarking. We compare Diffbot's coverage of UK companies and workers against UK administrative data for which the sampling frame is known. We use Diffbot figures for the start of 2022, to allow us to make worker comparisons with the 2021 England and Wales Census, as well as to labour force and business data. In what follows we summarise material from Gray et al. (2025).

We first compare Diffbot company counts against those in the UK company register, Companies House (CH), and firm counts from the ONS Business Population Estimates (BPE). The intuition for this test is as follows. In CH, each company observation represents a legal entity: real-world firms may include multiple corporate entities. The BPE includes the total number of active private sector businesses, built from the population of actual firms, captured from business tax data, plus an estimate of sole proprietorships. Diffbot takes CH as an input and uses supervised learning to identify the underlying business (an

'organisation' in Diffbot's ontology).Table A1 gives results. Encouragingly, we find that our most precise Diffbot specification – counts of active companies with CRN identifiers – is significantly lower than the count of entities in Companies House, and only slightly larger than the count of enterprises in the BPE.

Table A1. Company counts in Diffbot vs UK administrative data, 2022

Sampling frame	**Diffbot**	**BPE**	**Companies House**
(1) All active for-profit and non-profit organisations	10,588,606		
(2) All active for-profit orgs	10,223,570		
(3) All active for-profit orgs, start of 2022	8,462,847	5,508,935	
(4) All active for-profit orgs, start of 2022, with CRNs	**3,497,151**	**2,947,932**	**5,012,950**

Notes: Diffbot results report variations on the query *type:Organization not (isDissolved:true) location.country.name:"United Kingdom"*. Row 1 reports this query. Row 2 adds the condition *not(isNonProfit:true)*. Row 3 adds the condition *NOT(foundingDate>"2025-01-01")*. Row 4 adds the condition *has:companiesHouseIds*. UK Business Population Estimates (BPE) include the total number of private sector businesses, including: companies, all partnerships, and estimates of sole proprietorships (including those unregistered for PAYE/VAT). Estimates in row 3 include all private sector firms at the start of 2022, including companies, all partnerships and sole proprietorships. Estimates in row 4 include companies, sole proprietors with staff and sole proprietors registered for PAYE and/or VAT. Companies House includes all registered companies, including companies, LLPs and some other partnerships. We exclude dormant and non-trading companies. Source: Diffbot, ONS, OpenCorporates.

Diffbot matching test. Diffbot scrapes Companies House profiles and identifiers, so in theory every company in our search sample should precisely match to a Diffbot organisation. We first manually explore Diffbot sources for a sample of 100 non-matches. This check shows Diffbot's workflow scrapes Companies House, but then sometimes fails to place CRN data in the relevant organisation profile cell. This pattern of ascription error appears to be random.

Next, we formally test whether company observable characteristics might plausibly influence Diffbot's matching workflow. For all 55,187 companies in the search sample (Step 2), we estimate a cross-sectional linear probability model for company i where the dependent variable is a dummy equal to one if the company is CRN-matched in Diffbot and zero otherwise:

$$\Pr(\text{match})_i = F\big(\mathbf{Observables}\big)_i \tag{A}$$

Where $\textbf{Observables}_i$ is a set of company characteristics, plus 1-digit industry dummies and region dummies. Table A2 gives results. Overall, our results suggest that CRN (non)-matching is not explained by company characteristics: predictor effect sizes are small, model fit is very low, and unobservables play little role in explaining remaining variance. The most important statistically significant predictor, being a dissolved company, reduces the probability of a match by just over 2 percent. We also show that unobserved company characteristics do not drive this result, using an Oster Test on the dissolved dummy. Unobservables need to be 43 times larger than observables to nullify the effect we find.

Table A2. Match rate predictors test

Dependent variable: CRN match in Diffbot	(1)
Dissolved	-0.212***
	(0.00667)
Average company size	0.0171***
	(0.00193)
Incorporation year	-0.00107***
	(0.000110)
Log assets	-0.0392***
	(0.00194)
EBITDA	0.00136***
	(0.000225)
Number of subs.	-0.000439
	(0.000526)
Foreign subs.	-0.100***
	(0.0357)
GUO	0.0928***
	(0.00533)
Industry FE	Yes
Region FE	Yes
Observations	47,576
R^2	0.0736
Oster delta, dissolved	43.01

Notes: The table shows the results of a linear probability model of the probability of a company matching on CRN in Diffbot, on company observable characteristics, industry and region dummies. The dependent variable is a dummy that equals one if the company is matched in Diffbot on CRN, zero otherwise. Robust standard errors are in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Diffbot selection test. For a given time period, we define the coverage rate for a company as the count of workers observed in Diffbot in that period over count of employees reported in Orbis or Orbis Historical in that period. To explore coverage rate predictors, we adapt a test by Fedyk and Hodson (2022), who explore coverage of Cognism for a sample

of large US firms. Specifically, we estimate:

$$\text{Coverage}_{ijt} = F\big(\textbf{Observables}_{ijat},\, I_i,\, J_j,\, A_a,\, T_t\big) \tag{B}$$

Where Coverage_{ijt} is the coverage ratio of company i, sector j, and region a in year t, time-varying **Observables** are as before, and I_i, J_j, A_a and T_t company, sector, region and year fixed effects. We first run pooled OLS regressions. Results are given in **Table A4**. Column (1) looks at predictors of the annual coverage ratio; column (2) the average coverage ratio across all years of our data. In both cases, selection on observables is present but trivial: some coefficients are significant, but effect sizes remain small and model fit is very low. To test for the influence of unobservables on workforce coverage we run an Oster Test on the most important predictor, in this case firm size. The delta is between 3.1 and 4.3, suggesting that unobservables would need to be three to four times more important to dominate our main result. Column (3) extends the test to a typical employer-employee setting, where researchers fit company and year fixed effects: here we can explain almost 80 percent of the coverage ratio variation.

C Diffbot skills workflow

Skills selection. We observe skills for over 80 percent of workers in our sample. This suggests that workers could be selected into skills. To test this, we run a linear probability model regressing a worker's probability of having Diffbot skills, controlling for individual characteristics, the kind of job they are doing at the time, and the year skills are observed in the data. For worker i in 4-digit occupation bin o observed in year t, we regress

$$\Pr\big(Y_{iot} = 1\big) = F\big(\mathbf{X}_{iot},\, O_o,\, T_t\big) \tag{C}$$

Where Y_{iot} is a dummy taking the value one if a worker has observed skills, $\mathbf{X}_{iot}$ is a vector of worker observables, O_o is one of 411 SOC4 fixed effects and T_t is the year skills are observed, from 2007 to 2023.

Results are shown in Table A4. Overall, model fit and coefficient effect sizes are low.

Table A3. Match rate predictors test

Dependent variable = workforce coverage ratio	(1)	(2)	(3)
Dissolved	-0.0329***	-0.0817***	
	(0.0169)	(0.0158)	
Firm size	-0.350***	-0.288***	-0.752***
	(0.0194)	(0.0157)	(0.0475)
CH Incorporation year	0.000835***	0.00128***	
	(0.000163)	(0.000147)	
Log assets	0.153***	0.131***	0.175***
	(0.0150)	(0.0131)	(0.0174)
EBITDA	-0.000417	-0.000761	-0.000317
	(0.000505)	(0.000476)	(0.000308)
Number of subs.	-0.000672	-0.00178*	0.00410***
	(0.000981)	(0.000942)	(0.000912)
Foreign subs.	-0.183***	-0.0919**	0.0338
	(0.0445)	(0.0449)	(0.0222)
GUO	0.0104*	0.00346	-0.00490
	(0.00630)	(0.00553)	(0.0119)
Industry FE	Yes	Yes	Yes
Region FE	Yes	Yes	Yes
Firm FE	No	No	Yes
Year FE	Yes	Yes	Yes
Observations	184,096	184,096	180,752
R^2	0.0444	0.0437	0.791
Oster delta, firm size	2.904	4.164	

Notes: The table shows the results of a linear probability model of the probability of a company matching on CRN in Diffbot, on company observable characteristics, industry and region dummies. The dependent variable is a dummy that equals one if the company is matched in Diffbot on CRN, zero otherwise. Robust standard errors are in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Speaking a foreign language and having a PhD, the most important predictors, increase the probability of having skills by around 7 percent and 3 percent respectively. Oster tests on these variables give deltas of 8.8 and 4.1 respectively, suggesting unobserved worker characteristics are essentially trivial in explaining whether or not Diffbot skills are observed.

LDA workflow. Diffbot's skills ontology includes roughly 32,000 professional skills. To make the raw data tractable, we use topic modelling. We select a 25-topic model based on goodness of fit. Topics are summarized using the top 10 and top 300 words. The LDA assigns topic probabilities to individuals based on their observed skills, with probabilities summing to one. For each individual, we retain the three topics with the highest probabilities; we refer to the topic with the highest probability as the 'dominant topic' for that individual. Figure A1 shows dominant topic prevalence across the preferred 25-topic model, topic distribution varies across workers but without obvious extremes.

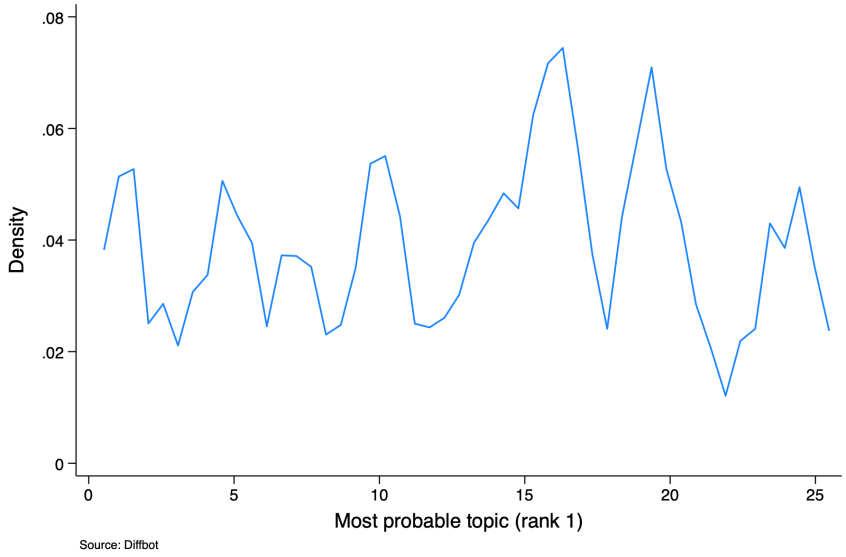

Figure A1. Topic prevalence in 25-topic model

We use Large Language Models to label each topic. We also use LLMs to rank topics by the complexity of the implied tasks, using ISCO descriptors as a benchmark. We use GPT4-o, a reasoning model, for our preferred levels. Alternate labellings using GPT-3.5 and Claude give similar rankings (Figure A2).

Table A4. Diffbot skills selection test

	(1)	(2)
migrant	0.0200***	0.0184***
	(0.00111)	(0.00120)
speaks foreign language	0.0804***	0.0714***
	(0.00123)	(0.00128)
age		-0.00195***
		(0.0000340)
female		0.00822***
		(0.00103)
has STEM degree		0.0173***
		(0.00116)
has economics degree		0.0179***
		(0.00115)
has arts / humanities degree		0.0247***
		(0.00130)
studied at oxbridge		-0.0163***
		(0.00266)
studied at russell group uni		-0.0000113
		(0.00130)
graduate		-0.00321**
		(0.00144)
has postgard degree		0.00750***
		(0.00180)
phd		0.0295***
		(0.00326)
tech occupation		0.0216***
		(0.00175)
managerial occupation		0.0242***
		(0.00124)
years of experience		0.00925***
		(0.0000730)
Constant	0.857***	0.803***
	(0.000535)	(0.00233)
Observations	531087	508266
R^2	0.0334	0.0768
Oster delta / foreign language		8.980

Notes: The table shows the results of a linear probability model of the probability of a worker having observed Diffbot skills. Robust standard errors are in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

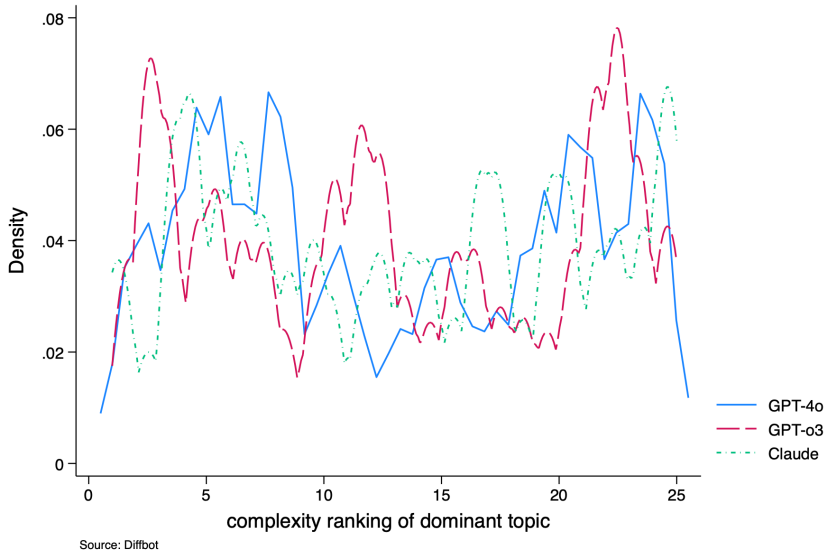

Figure A2. Comparison of topic rankings across classifiers

D Measuring migrant status

As detailed in Gray et al. (2025), Diffbot contains fields for person birth country and nationality, but in our data these are almost always blank. However, the large majority of people go to secondary school or university in their country of birth. A line with recent papers, we proxy workers' country of birth using the country of their lowest recorded education in Diffbot (Jin et al., 2025b; Lee and Glennon, 2023). This is typically an undergraduate degree, but in about 15 percent of cases it is high school qualifications (at age 16 or 18). We extract education location information from descriptive free text (e.g. 'University in London, England'); where this is not available we map institution names to UK government dictionaries of UK schools, and a global list of towns and cities with at least 1,000 inhabitants. In theory our measure is vulnerable to false positives (for example, UK-born UK-based workers educated abroad) and false negatives (non-UK born UK-based workers educated here, especially at university level). However, in countries that are net exporters of higher education, like the UK or US, our measure will understate the true number of migrant workers in our sample: in 2022/3, for example, 14.7 percent of undergraduate students in the UK were non-UK born (Cuibus et al., 2025; HESA, 2024).

We ran two checks on our processed data. First, to test the quality of imputation an

RA manually validated a random sample of 100 worker observations where we have directly observed levels and locations of education spells. We impute education location/s and compare to the observed information. Country of education is correctly imputed for 90/100 observations. Second, we test the country of lowest education proxy on a sample of staff and PhD students from a UCL Department, a setting a) where we have ground truth and b) individuals are disproportionately highly qualified, mirroring the larger sample of workers in Diffbot. We use a simple web survey, obtaining a response rate of 35 percent. All of our respondents attended university. Birth country correctly maps to country of university for 89.7 percent of respondents. Results for country of schooling are identical. As predicted, country of university predicts a lower bound for the true migrant share (51.4 versus 54.3 percent).

E Additional results

Table E1. Descriptive statistics for firm-level variables by coverage rate

	Coverage rate 0–0.5				Coverage rate 0.5–0.75				Coverage rate >0.75			
	N. obs	Mean	Sd	p50	N. obs	Mean	Sd	p50	N. obs	Mean	Sd	p50
Log TFP (Olley-Pakes)	31683	3.91	1.62	4.34	15371	3.83	1.54	3.79	17880	3.82	1.65	3.59
Number of patents	32932	0.090	1.17	0	16295	0.13	1.96	0	22963	0.074	1.27	0
Number of citations (all-time)	32932	0.30	4.62	0	16295	0.83	18.5	0	22963	0.46	10.7	0
Share migrant workers	32932	0.099	0.13	0.062	16295	0.13	0.14	0.087	22963	0.15	0.17	0.092
Share of workers with a college or higher degree	32932	0.45	0.23	0.43	16295	0.55	0.23	0.56	22963	0.54	0.27	0.56
Share migrants with degree or above	32932	0.084	0.12	0.045	16295	0.12	0.14	0.073	22963	0.13	0.16	0.074
Share UK workers with degree or above	32932	0.36	0.20	0.34	16295	0.43	0.20	0.43	22963	0.40	0.23	0.40
Share of workers in tech occupations	32932	0.074	0.089	0.053	16295	0.078	0.093	0.048	22963	0.070	0.11	0.032
Share of workers in stem occupations	32932	0.082	0.10	0.050	16295	0.087	0.11	0.048	22963	0.078	0.12	0.032
Share of workers in managerial occupations	32932	0.39	0.19	0.38	16295	0.37	0.19	0.36	22963	0.35	0.22	0.33
Share of migrant workers in tech occupations	32932	0.0082	0.028	0	16295	0.012	0.030	0	22963	0.012	0.039	0
Share of UK workers in tech occupations	32932	0.038	0.058	0.011	16295	0.041	0.060	0.019	22963	0.035	0.067	0
Share of migrant workers in stem occupations	32932	0.0094	0.027	0	16295	0.014	0.036	0	22963	0.014	0.041	0
Share of UK workers in stem occupations	32932	0.044	0.068	0.013	16295	0.047	0.072	0.019	22963	0.040	0.077	0
Share of migrant workers in managerial occupations	32932	0.035	0.072	0	16295	0.042	0.072	0.019	22963	0.048	0.095	0.0092
Share of UK workers in managerial occupations	32932	0.19	0.13	0.17	16295	0.19	0.13	0.18	22963	0.18	0.16	0.15
Workforce average age	32404	44.1	5.89	43.5	16105	43.1	5.15	42.8	22178	43.4	6.25	42.7
Share of females	32932	0.32	0.20	0.30	16295	0.35	0.20	0.33	22963	0.34	0.22	0.32
Number of employees	32932	105.1	122.9	68	16295	85.6	123.9	59	13571	51.6	67.9	34
Firm age	32932	6.35	3.45	6	16295	6.23	3.40	6	22963	5.34	3.59	5
Share of workers in non-executive occupations	32932	0.019	0.056	0	16295	0.022	0.055	0	22963	0.025	0.074	0
Company has foreign subsidiaries	32932	0.0013	0.036	0	16295	0.0013	0.036	0	22963	0.0027	0.052	0
Number of subsidiaries	32932	0.78	2.70	0	16295	0.86	2.61	0	22963	0.80	3.40	0
Log firm revenues	32608	16.8	1.02	16.7	16088	16.7	1.08	16.6	21389	16.5	1.44	16.5

Notes: This table reports descriptive statistics (number of obs., mean, standard deviation and median) for the main variables used in the regression model, divided into three coverage rate ranges.

Table E2. Worker characteristics: whole sample vs. workers with Diffbot skills

	A. Whole sample			B. Workers with Diffbot skills		
Variable	N. obs	Mean	Sd	N. obs	Mean	Sd
Migrant	538,872	0.194	0.395	466,715	0.199	0.399
Has degree or higher qualifications	538,872	0.808	0.394	520,396	0.742	0.438
Has degree	538,872	0.600	0.490	520,396	0.543	0.498
Has postgraduate degree	538,872	0.187	0.390	520,396	0.179	0.383
Has PhD	538,872	0.021	0.143	520,396	0.020	0.141
Migrant with degree or higher	538,872	0.172	0.378	520,396	0.159	0.366
UK-born with degree or higher	538,872	0.636	0.481	617,738	0.481	0.500
Migrant with degree	538,872	0.096	0.295	617,738	0.074	0.261
UK-born with degree	538,872	0.504	0.500	617,738	0.381	0.486
Migrant with postgrad degree	538,872	0.069	0.253	617,738	0.054	0.227
UK-born with postgrad degree	538,872	0.118	0.323	617,738	0.090	0.286
Migrant with PhD	538,872	0.008	0.087	617,738	0.006	0.078
UK-born with PhD	538,872	0.013	0.115	617,738	0.011	0.102
Years of labour market experience	538,872	12.806	8.792	617,738	13.994	9.098
Migrant years of experience	104,409	12.298	7.948	92,836	12.525	7.904
UK-born years of experience	434,463	12.928	8.978	373,879	13.409	8.974
Tech occupation	538,872	0.081	0.274	617,738	0.085	0.280
STEM occupation	538,872	0.095	0.294	617,738	0.097	0.296
Managerial occupation	538,872	0.301	0.459	617,738	0.331	0.471
Migrant tech occupation	538,872	0.019	0.136	617,738	0.015	0.123
UK-born tech occupation	538,872	0.063	0.242	617,738	0.049	0.215
Migrant STEM occupation	538,872	0.023	0.151	617,738	0.019	0.136
UK-born STEM occupation	538,872	0.072	0.259	617,738	0.056	0.230
Migrant manager	538,872	0.055	0.228	617,738	0.044	0.204
UK-born manager	538,872	0.246	0.431	617,738	0.192	0.394
Most probable topic (rank 1)	466,715	13.099	7.154	617,738	13.005	7.241
Migrant most probable topic	92,836	13.257	6.934	92,836	13.257	6.934
UK-born most probable topic	373,879	13.059	7.207	373,879	13.059	7.207
GPT-4o-based ranking of top1 topic	466,715	13.071	7.502	617,738	13.041	7.480
Migrant GPT-4o-based ranking	92,836	13.870	7.542	92,836	13.870	7.542
UK-born GPT-4o-based ranking	373,879	12.873	7.479	373,879	12.873	7.479

Notes: This table reports descriptive statistics (number of obs., mean and standard deviation) for worker-level variables for the whole sample (Panel A) and for the sample of workers with Diffbot skills (Panel B).

Table E3. Worker characteristics in 2023: full sample vs. Diffbot skills sample

	A. 2023 only			B. Diffbot skills, 2023 only		
Variable	count	mean	sd	count	mean	sd
Migrant	184,802	0.183	0.387	155,327	0.188	0.391
Has degree or higher qualifications	208,434	0.711	0.453	174,983	0.719	0.449
Has degree	208,434	0.530	0.499	174,983	0.533	0.499
Has postgraduate degree	208,434	0.161	0.368	174,983	0.166	0.372
Has PhD	208,434	0.019	0.137	174,983	0.021	0.142
Migrant with degree or higher	208,434	0.143	0.350	174,983	0.148	0.355
UK-born with degree or higher	275,663	0.421	0.494	214,914	0.456	0.498
Migrant with degree	275,663	0.061	0.239	214,914	0.067	0.250
UK-born with degree	275,663	0.337	0.473	214,914	0.364	0.481
Migrant with postgrad degree	275,663	0.042	0.201	214,914	0.048	0.213
UK-born with postgrad degree	275,663	0.075	0.263	214,914	0.082	0.274
Migrant with PhD	275,663	0.005	0.069	214,914	0.006	0.075
UK-born with PhD	275,663	0.009	0.096	214,914	0.011	0.102
Years of labour market experience	275,663	17.266	8.881	214,914	18.317	8.594
Migrant years of experience	33,844	16.130	7.587	29,262	16.612	7.368
UK-born years of experience	150,958	16.958	8.696	126,065	17.822	8.441
Tech occupation	275,663	0.089	0.285	214,914	0.095	0.293
STEM occupation	275,663	0.100	0.299	214,914	0.107	0.309
Managerial occupation	275,663	0.327	0.469	214,914	0.342	0.474
Migrant tech occupation	275,663	0.014	0.117	214,914	0.016	0.126
UK-born tech occupation	275,663	0.048	0.214	214,914	0.054	0.227
Migrant STEM occupation	275,663	0.016	0.127	214,914	0.019	0.137
UK-born STEM occupation	275,663	0.055	0.228	214,914	0.062	0.241
Migrant manager	275,663	0.036	0.186	214,914	0.041	0.197
UK-born manager	275,663	0.172	0.378	214,914	0.194	0.395
Most probable topic (rank 1)	214,914	12.962	7.208	214,914	12.962	7.208
Migrant most probable topic	29,262	13.271	6.837	29,262	13.271	6.837
UK-born most probable topic	126,065	13.123	7.142	126,065	13.123	7.142
GPT-4o-based ranking of top1$_t opic$	214,914	13.191	7.493	214,914	13.191	7.493
Migrant GPT-4o-based ranking	29,262	14.051	7.520	29,262	14.051	7.520
UK-born GPT-4o-based ranking	126,065	13.093	7.511	126,065	13.093	7.511

Notes: This table reports descriptive statistics (number of obs., mean and standard deviation) for worker-level variables for the whole sample (Panel A) and for the sample of workers with Diffbot skills (Panel B), in 2023.

Table E4. Top SOC4 titles (all workers vs. migrant workers) for workers with skills

A. All workers		B. Migrant workers		
SOC4 title (all workers)	Count (all)	SOC4 title (migrant workers)	Count (total)	Migrant share
financial accounts managers	24,249	fishing and other elementary agriculture occupations nec	2	50.0%
management consultants and business analysts	15,196	leisure and travel service occupations nec	28	42.9%
marketing and commercial managers	13,152	biochemists and biomedical scientists	108	40.7%
customer service managers	12,287	dental nurses	23	39.1%
office managers	11,761	generalist medical practitioners	111	37.8%
programmers and software development professionals	10,349	authors writers and translators	522	37.7%
it project managers	9,928	restaurant and catering establishment managers and proprietors	133	37.6%
sales administrators	9,452	functional managers and directors nec	94	37.2%
finance officers	9,205	butchers	111	34.2%
it managers	8,820	architects	4,300	33.3%
information technology directors	8,696	nannies and au pairs	6	33.3%
business sales executives	8,691	education advisers and school inspectors	36	33.3%
human resource managers and directors	8,412	assemblers and routine operatives nec	37	32.4%
directors in consultancy services	7,880	residential day and domiciliary care managers and proprietors	34	32.4%
mechanical engineers	7,659	importers and exporters	539	32.3%
data analysts	7,543	market and street traders and assistants	812	32.3%
senior care workers	6,865	business and related research professionals	1,095	32.1%
sales accounts and business development managers	6,694	programmers and software development professionals	10,349	32.0%
solicitors and lawyers	5,782	teaching professionals nec	22	31.8%
customer service supervisors	5,573	mobile machine drivers and operatives nec	41	31.7%
sales and retail assistants	5,518	sheet metal workers	19	31.6%
sales supervisors	5,476	bakers and flour confectioners	19	31.6%
it operations technicians	5,245	special and additional needs education teaching professionals	48	31.3%
engineering project managers and project engineers	4,973	betting shop and gambling establishment managers	74	31.1%
engineering technicians	4,846	teachers of english as a foreign language	351	31.1%

Notes: This table reports counts of workers in the top-25 occupations crosswalked to SOC4, for workers reporting skills. Panel A provides data for all workers, while Panel B for migrant workers.

Table E5. Main results

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS	(7) OLS	(8) IV
Firm migrant share	0.0178	0.0236	0.0225	0.0179	0.0149	0.0126	0.0687	0.345
	(0.0176)	(0.0182)	(0.0183)	(0.0183)	(0.0186)	(0.0187)	(0.0575)	(1.098)
Log(worker average age)		-0.00339	-0.00325	-0.00207	-0.00146	-0.00459	-0.0493	-0.0183
		(0.0173)	(0.0175)	(0.0178)	(0.0180)	(0.0186)	(0.0511)	(0.0597)
Share of graduates		-0.00753	-0.00926	-0.00200	-0.00348	-0.00116	-0.0129	0.00722
		(0.0133)	(0.0135)	(0.0145)	(0.0148)	(0.0152)	(0.0414)	(0.0435)
Share of females		-0.00750	-0.00169	-0.00840	-0.00863	-0.00804	-0.0297	-0.0111
		(0.0103)	(0.0107)	(0.0121)	(0.0127)	(0.0132)	(0.0467)	(0.0606)
Share of workers in non-exec. pos.		-0.0472	-0.0545	-0.0627	-0.0532	-0.0568	-0.113	-0.149
		(0.0588)	(0.0601)	(0.0572)	(0.0573)	(0.0581)	(0.135)	(0.139)
Foreign subsidiary dummies		-0.0508	-0.0539	-0.0515	-0.0498	-0.0616	-0.180	-0.179
		(0.0502)	(0.0503)	(0.0498)	(0.0512)	(0.0558)	(0.181)	(0.181)
Number of subsidiaries		0.00184***	0.00167**	0.00171**	0.00177***	0.00170**	0.00366**	0.00357**
		(0.000684)	(0.000669)	(0.000669)	(0.000668)	(0.000683)	(0.00151)	(0.00150)
Coverage rate		0.000826	0.000853	0.000892	0.000896	0.000998	-0.00563**	-0.00567**
		(0.00130)	(0.00133)	(0.00138)	(0.00139)	(0.00139)	(0.00271)	(0.00275)
Number of patents		-0.00193	-0.00150	-0.00111	-0.00128	-0.00132	0.000215	0.000101
		(0.00160)	(0.00159)	(0.00160)	(0.00164)	(0.00165)	(0.00330)	(0.00330)
Number of forward citations		-0.000127	-0.000140	-0.000130	-0.000126	-0.000158	-0.000313	-0.000304
		(0.000158)	(0.000155)	(0.000152)	(0.000156)	(0.000158)	(0.000219)	(0.000222)
Log(firm revenues)		-0.0204***	-0.0209***	-0.0242***	-0.0248***	-0.0260***	-0.155***	-0.153***
		(0.00246)	(0.00253)	(0.00266)	(0.00270)	(0.00273)	(0.0103)	(0.0105)
Region-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	No	No	1-digit	2-digit	3-digit	4-digit	Absorbed	Absorbed
Firm FE	No	No	No	No	No	No	Yes	Yes
N firms	6373	6373	6373	6373	6373	6373	6373	6373
Observations	46906	46906	46906	46906	46906	46906	46906	46501
R-squared	0.00786	0.0105	0.0113	0.0145	0.0178	0.0211	0.177	0.0256
First stage coefficient	-	-	-	-	-	-	-	-0.002
K-P F-stat	-	-	-	-	-	-	-	10.71

Notes: The dependent variable is TFP growth. The estimation model is indicated in the column header. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

Table E6. Correlations table

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. TFP growth	1										
2. Share migrant workers	0.0109*	1									
3. Workforce average age	0.00146	-0.103***	1								
4. Share of females	0.000693	0.0372***	-0.184***	1							
5. Share of graduates	0.00749	0.495***	-0.185***	0.245***	1						
6. Share of workers in non-executive positions	-0.00355	0.0144**	0.0163***	0.117***	0.0970***	1					
7. Foreign subsidiary dummy	-0.00368	0.0515***	-0.000950	0.0167***	0.0227***	-0.00321	1				
8. Number of subsidiaries	0.0101*	-0.0152***	0.0322***	0.0432***	0.0133**	0.0257***	0.0511***	1			
9. Coverage rate	0.00372	0.0450***	-0.0272***	0.00563	0.0370***	0.0196***	0.0114*	-0.00115	1		
10. Number of patents	-0.00883	0.00540	0.0315***	-0.0204***	0.0432***	-0.00975*	-0.00224	0.0157***	0.0232***	1	
11. Number of citations	-0.00727	0.000346	0.0291***	-0.0114*	0.0326***	-0.00933*	-0.00162	0.0149**	0.00601	0.636***	1

Notes: This table reports pairwise correlations among the variables of our regression model. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E7. Main results - TFP level

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS	(7) OLS	(8) IV
Firm migrant share	0.614***	0.494***	0.617***	0.137**	0.138**	0.128**	-0.00574	2.765
	(0.150)	(0.149)	(0.130)	(0.0666)	(0.0653)	(0.0647)	(0.0772)	(1.826)
Region-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	No	No	1-digit	2-digit	3-digit	4-digit	Absorbed	Absorbed
Firm FE	No	No	No	No	No	No	Yes	Yes
N firms	6379	6379	6379	6379	6379	6379	6379	
Observations	46974	46974	46974	46974	46974	46974	46974	46568
R-squared	0.00921	0.0432	0.306	0.870	0.873	0.876	0.946	-0.0702
First stage coefficient								-0.00215
K-P F-stat								10.99

Notes: The dependent variable is TFP growth. The estimation model is indicated in the column header. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

Table E8. Sensitivity checks

	(1) Industry-Region FE	(2) Industry-Year FE	(3) Industry-Region-Year FE	(4) Post 2010	(5) CR<1	(6) 0.5<CR <1	(7) 0.75<CR<1	(8) Weighted regressions	(9) Mig share outliers	(10) TFP outliers
Firm migrant share	0.0669	0.0709	0.0748	0.0593	0.0898	0.103	-0.144	-0.00169	0.0499	0.0196
	(0.0575)	(0.0576)	(0.0574)	(0.0566)	(0.0577)	(0.0951)	(0.183)	(0.106)	(0.0603)	(0.0533)
N firms	6373	6373	6347	6320	5960	3018	1134	6298	6316	6337
Observations	46906	46903	46677	46616	41665	17522	4672	46055	46410	46442
R-squared	0.174	0.182	0.207	0.176	0.174	0.213	0.302	0.490	0.180	0.207

Notes: The dependent variable is TFP growth. Each column reports a different specification. Column (1) adds region-by-industry (1-digit) FE; column (2) adds industry (1-digit)-by-region FE; column (3) adds industry(1-digit)-by-region-by-year FE; column (4) restricts the estimation to post 2010; column (5) reports results when the coverage rate is < 1; column (6) reports results when the coverage rate is between 0.5 and 1; column (7) reports results when the coverage rate is between 0.75 and 1; column (8) weighs regression by the coverage rate; column (9) excludes observations above the top 1% percentile of the migrant share distribution; column (10) excludes observations above the top 1% percentile of the TFP distribution. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

Table E9. Lee Bounds exercise

	(1) Main	(2) Min	(3) Max	(4) Observed only
Firm migrant share	0.0687			-0.103
	(0.058)			(0.270)
Firm migrant share (Max)		0.0000335		
		(0.017)		
Firm migrant share (Min)			-0.00422	
			(0.022)	
N firms	6,373	6,373	6,373	883
Observations	46,906	46,906	46,906	4,302
R-squared	0.177	0.177	0.177	0.283

Notes: the dependent variable is the growth rate of TFP. Column (1) is the main specification from column (7), Table 5. Column (2) assumes all unobserved workers are migrant; column (3) assumes all unobserved workers are natives; column (4) restricts the sample to firms where we observe migrant status. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diff-bot coverage rate, weighted patent count, citation count, log(revenues). All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E10. Placebo check: Lagged TFP growth on migrant share

Migrant share	**Beta**	**Obs**	R^2
L1.TFP growth (Olley and Pakes method)	-0.000855	40,128	0.915
	(0.00123)		
L2.TFP growth (Olley and Pakes method)	0.00103	33,910	0.926
	(0.00118)		
L3.TFP (Olley and Pakes method)	-0.000303	28,932	0.936
	(0.00109)		
L4.TFP (Olley and Pakes method)	-0.000476	24,494	0.945
	(0.00109)		
L5.TFP (Olley and Pakes method)	0.000920	20,273	0.952
	(0.00101)		
L6.TFP (Olley and Pakes method)	0.000622	16,282	0.962
	(0.00127)		
L7.TFP (Olley and Pakes method)	-0.00114	12,565	0.971
	(0.00107)		
L8.TFP (Olley and Pakes method)	0.000472	9,139	0.980
	(0.00140)		
L9.TFP (Olley and Pakes method)	0.00000502	6,200	0.985
	(0.00114)		
L10.TFP (Olley and Pakes method)	-0.000937	3,296	0.992
	(0.00223)		

Notes: Each row represents a regression of the migrant share on lagged TFP growth. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, dummy for number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

Table E11. Migrant share shuffled

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS	(7) OLS	(8) IV
Firm migrant share (shuffled)	-0.00726	-0.00674	-0.00671	-0.00572	-0.00474	-0.00442	-0.00710	2.255
	(0.0129)	(0.0129)	(0.0129)	(0.0129)	(0.0129)	(0.0129)	(0.0136)	(9.613)
Region-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	No	No	1-digit	2-digit	3-digit	4-digit	Absorbed	Absorbed
Firm FE	No	No	No	No	No	No	Yes	Yes
N firms	6,373	6,373	6,373	6,373	6,373	6,373	6,373	6,373
Observations	46,906	46,906	46,906	46,906	46,906	46,906	46,906	46,501
R-squared	0.00784	0.0104	0.0112	0.0145	0.0178	0.0211	0.177	-0.626
First stage coefficient								-0.000326
K-P F-stat								0.134

Notes: The dependent variable is TFP growth. The estimation model is indicated in the column header. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

Table E12. Extensions: Industry heterogeneity

	(1) main	(2) Services	(3) Manufacturing	(4) High-tech (ONS)	(5) High-tech (NESTA)
Firm migrant share	0.0687	0.00643	0.0702	0.0546	0.0758
	(0.0575)	(0.102)	(0.0604)	(0.0670)	(0.0610)
Firm migrant share × Services		0.0886			
		(0.124)			
Firm migrant share × Manufacturing			-0.0192		
			(0.186)		
Firm migrant share × High-tech (ONS)				0.0607	
				(0.128)	
Firm migrant share × High-tech (NESTA)					-0.0462
					(0.173)
N firms	6,373	6,373	6,373	6,373	6,373
Observations	46,906	46,906	46,906	46,906	46,906
R-squared	0.177	0.177	0.177	0.177	0.177

Source: Diffbot, Orbis Historical, OpenCorporates, PATSTAT. *Notes:* the dependent variable is the growth rate of TFP. The estimation model is indicated in the column header. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E13. Extensions: STEM / tech occupations

	(1) Main	(2) Migrant Tech	(3) Migrant STEM	(4) Above mean STEM	(5) Above mean Tech	(6) Share STEM	(7) Share Tech
Firm migrant share	0.0687			-0.0212	-0.0237	0.0677	0.0654
	(0.0575)			(0.0662)	(0.0662)	(0.0640)	(0.0642)
Firm migrant share in tech occupations		0.0286					
		(0.208)					
Firm migrant share in stem occupations			-0.0383				
			(0.181)				
Firm migrant share × Above mean STEM				0.208***			
				(0.0796)			
Firm migrant share × Above mean Tech					0.209***		
					(0.0776)		
Firm migrant share × Share STEM						0.00800	
						(0.415)	
Firm migrant share × Share Tech							0.0380
							(0.479)
N groups	6373	6373	6373	6373	6373	6373	6373
Observations	46906	46906	46906	46906	46906	46906	46906
R-squared	0.177	0.177	0.177	0.178	0.178	0.177	0.177

Notes: The dependent variable is TFP growth. The estimation model is indicated in the column header. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, weighted patent count, citation count, log(revenues). "K-P F-stat" refers to the Kleibergen-Paap F-statistic for first stage results. All explanatory variables are 1-period lagged. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Source:* Diffbot, Orbis Historical, OpenCorporates, PATSTAT.

Table E14. Diffbot distinctive skills test: sensitivity checks

	(1) OLS	(2) IV	(3) logit	(4) probit	(5) 2023	(6) 2022-3	(7) 2021-3	(8) 2020-3	(9) 2019-3	(10) 2011-23	(11) Top 2 topics	To
main												
migrant	0.155***	1.416***	0.0377***	0.00964**	1.262***	1.322***	1.379***	1.320***	1.344***	1.428***	0.356***	
	(0.0258)	(0.0886)	(0.00646)	(0.00388)	(0.153)	(0.141)	(0.135)	(0.127)	(0.120)	(0.0893)	(0.0911)	
Observations	437630	437630	437630	437630	145785	166571	180106	201014	225919	427947	437630	
R^2	0.122	-0.00154			0.0000690	-0.000420	-0.000829	-0.000520	-0.000873	-0.00161	0.00117	
Pseudo R^2			0.0240	0.0216								

Source: Diffbot. *Notes:* 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 controls per main paper. Column (1) fits IV per main paper. Cols (2) and (3) fit ordered logit and probit estimators. Cols (4)-(9) use alternate time windows. Cols (10)-(11) use alternate dominant topics. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E15. Diffbot complex skills test: sensitivity tests for tech migrants vs natives

	(1) OLS	(2) IV	(3) GPT-o3	(4) Claude	(5) 2023	(6) 2022-3	(7) 2021-3	(8) 2020-3	(9) 2019-3	(10) 2011-23	(11) Placebo
Migrant in tech occupation	0.895***	1.483***	1.604***	1.328***	1.477***	1.530***	1.502***	1.456***	1.478***	1.500***	0.156*
	(0.0812)	(0.221)	(0.211)	(0.233)	(0.353)	(0.329)	(0.319)	(0.301)	(0.287)	(0.222)	(0.0920
Observations	37224	37224	37224	37275	14216	16105	17383	19312	21258	36460	32761
R^2	0.215	0.00749	0.0100	0.00802	0.0106	0.0111	0.0101	0.0101	0.0108	0.00761	0.00900
Underidentification test		3513.7	3513.7	3546.5	1391.2	1602.8	1705.2	1907.9	2104.7	3470.0	
Under-identification test p-value		0	0	0	1.70e-304	0	0	0	0	0	
K-P F-statistic		4251.6	4251.6	4327.9	1693.4	1957.8	2078.8	2332.3	2570.1	4203.2	

Source: Diffbot. *Notes:* Sample is workers with Diffbot skills and observables. Regressions compare the ISCO-complexity of skills for migrants versus migrants and natives doing the same role types. Columns (1) and (2) fit OLS and IV, per Table 8. Columns (3) and (4) fit alternate topic complexity rankings, from GPT-o3 and Claude respectively. Columns (5)-(10) use alternate time windows. Column (11) runs a placebo test as before. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E16. Summary statistics for specialisation subsamples

(a) A. All firms

	N	Mean	sd	Median
Management roles (migrant share / native share) ratio	24616	0.24	0.46	0.12
Tech roles (migrant share / native share) ratio	24616	0.27	0.61	0
stem roles (migrant share / native share) ratio	24616	0.28	0.61	0
Migrant tech role specialisation	15961	1.41	2.72	0.34
Migrant stem role specialisation	15961	1.45	2.78	0.44
Migrant management role specialisation	10538	0.73	1.03	0.44
Observations	24616			

(a) B. Patenting firms

	N	Mean	sd	Median
Management roles (migrant share / native share) ratio	896	0.20	0.25	0.13
Tech roles (migrant share / native share) ratio	896	0.21	0.38	0
Stem roles (migrant share / native share) ratio	896	0.22	0.36	0.100
Migrant tech role specialisation	594	1.15	1.84	0.50
Migrant stem role specialisation	594	1.29	2.11	0.61
Migrant management role specialisation	493	1.04	1.36	0.57
Observations	896			

Source: Diffbot, Orbis Historical OpenCorporates, PATSTAT.

Table E18. Placebo check: migrant share on lagged patent count

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
L.Number of patents	0.000474							
	(0.006)							
L2.Number of patents		0.00254						
		(0.007)						
L3.Number of patents			0.0110					
			(0.009)					
L4.Number of patents				-0.00792				
				(0.009)				
L5.Number of patents					-0.00338			
					(0.007)			
L6.Number of patents						-0.00311		
						(0.010)		
L7.Number of patents							0.0151	
							(0.013)	
L8.Number of patents								-0.00170
								(0.004)
L9.Number of patents								
L10.Number of patents								
N groups	2243	1975	1757	1578	1430	1241	1065	861
Observations	12912	10907	9182	7733	6389	5077	3899	2779
R^2	0.582	0.606	0.646	0.674	0.706	0.746	0.785	0.824

Source: Diffbot, Orbis Historical, OpenCorporates. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E19. Placebo check: randomised migrant tech specialisation

	(1)
	est1
L.ts_tech_shuffled	0.00839
	(0.015)
N groups	129
Observations	378
Pseudo R^2	0.553

Source: Diffbot, Orbis Historical, OpenCorporates. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E20. Migrant specialisation and firm patent quality

	(1)	(2)	(3)
A: Citations			
L.migrant tech role specialisation	0.0663		
	(0.0879)		
L.migrant stem role specialisation		-0.00711	
		(0.0675)	
L.migrant management role specialisation			0.134
			(0.123)
N companies	142	142	115
Observations	754	759	601
Pseudo R^2	0.729	0.732	0.721
B. Cites extensive margin			
L.migrant tech role specialisation	0.000410		
	(0.000434)		
L.migrant stem role specialisation		0.000452	
		(0.000536)	
L.migrant management role specialisation			-0.000495
			(0.00368)
N companies	2243	2192	1506
Observations	12912	12619	8384
R^2	0.449	0.448	0.457
C. Cites intensive margin			
L.migrant tech role specialisation	0.0140		
	(0.0870)		
L.migrant stem role specialisation		-0.0366	
		(0.0742)	
L.migrant management role specialisation			0.0455
			(0.0612)
N companies	95	96	80
Observations	332	334	286
Pseudo R^2	0.789	0.793	0.806

Source: Diffbot, Orbis Historical, OpenCorporates. *Notes:* Panels A and C fit PPML estimators. Panel B fits an LPM estimator. Other notes per Table 10.

Table E21. Horse race between specialisation and other measures

	(1)	(2)	(3)	(4)	(5)
L.migrant tech role specialisation	0.115**	0.110**	0.109**	0.111**	0.110**
	(0.047)	(0.049)	(0.050)	(0.050)	(0.050)
L.firm share high-skilled migrants		-0.430	-1.148	1.123	-1.037
		(2.628)	(2.652)	(2.615)	(2.654)
L.Share of high-skilled natives		-1.769	-1.469	-1.227	-2.481
		(1.965)	(1.910)	(2.182)	(3.626)
L.Average migrants' experience			-0.0579		-0.0597
			(0.036)		(0.038)
L.Average natives' experience			-0.0305		-0.0279
			(0.106)		(0.108)
L.Share of migrant workers with above-median experience				-4.716	
				(3.247)	
L.Share of UK worker with above-median experience				-0.633	
				(1.809)	
L.Fractionalization index, birth country					-1.137
					(3.615)
N groups	189	189	189	189	189
Observations	987	987	987	987	987
Pseudo-R^2	0.618	0.618	0.619	0.619	0.619

Source: Diffbot, Orbis Historical, OpenCorporates. *Notes: Notes:* Regressions estimate firm outcomes on migrant specialisation in tech, STEM or management roles. All regressions fit controls, area, region, year and firm FE. Controls include: log(workforce mean age), share of graduates, share of females, share of workers in non-executive positions, dummy for number of subsidiaries, dummy for foreign subsidiaries, Diffbot coverage rate, log(revenues). All explanatory variables are 1-period lagged. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$