



ARTICLE



<https://doi.org/10.1057/s41599-021-00862-2>

OPEN

Developing China's workforce skill taxonomy reveals extent of labor market polarization

Weipan Xu¹, Xiaozhen Qin¹, Xun Li^{1✉}, Haohui Chen², Morgan Frank^{3,4}, Alex Rutherford⁵, Andrew Reeson² & Iyad Rahwan⁵

China, the world's second largest economy, is transitioning into an advanced, knowledge-based economy after four decades of rapid economic development. However, China still lacks a detailed understanding of the skills that underly the Chinese labor force, and the development and spatial distribution of these skills. Similar data has proven essential in other contexts; for example, the US standardized skill taxonomy, Occupational Information Network (O*NET), played an important role in understanding the dynamics of manufacturing and knowledge-based work, and the potential risks from automation and outsourcing. Here, we use Machine Learning techniques to bridge this gap, creating China's first workforce skill taxonomy, and map it to O*NET. This enables us to reveal workforce skill polarization into social-cognitive skills and sensory-physical skills, and to explore China's regional inequality in light of workforce skills, and compare it to traditional metrics such as education. We build an online tool for the public and policy makers to explore the skill taxonomy: skills.sysu.edu.cn. We also make the taxonomy dataset publicly available for other researchers.

¹Department of Urban and Regional Planning, School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. ²Data61, Commonwealth Scientific and Industrial Research Organisation, Melbourne, Australia. ³Department of Informatics and Networked Systems, University of Pittsburgh, Pittsburgh, PA, USA. ⁴Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁵Center for Humans & Machines, Max-Planck Institute for Human Development, Berlin, Germany. ✉email: lixun@mail.sysu.edu.cn

Introduction

Workers rely on their skills to earn a living. Fluctuations in the demand for particular skills translate to changes in wages and employment opportunities for individual workers. Technology has long driven such changes, which can result in significant economic and social upheavals (Mokyr et al., 2015). Technology acts as a substitute for some skills, for example calculation, but complements others, for example accountants who use the outputs of those calculations. In general, routine work has proven most susceptible to substitution, while non-routine tasks have been complemented by technology (Autor et al., 2003). For those with complementary skills, technology increases their productivity, and hence the wages their skills are likely to command (Brynjolfsson and McAfee, 2014); for those without such skills the outlook is less promising.

For example, in the US, the ‘college premium’, the wage advantage enjoyed by more educated (skilled) workers, has increased over the last 40 years while low skilled workers have seen their real earnings decline (Acemoglu and Autor, 2011). Over a similar time period many European countries also show evidence of job polarization, with declines in the middle third of the wage distribution (Goos et al., 2009). Unfortunately, it is not a simple matter for workers to move from declining to growing occupations. A recent analysis of US skills data shows a marked polarization between two distinct clusters of skills, one (social-cognitive) associated with high-income jobs and the other (sensory-physical) with low-income jobs (Alabdulkareem et al., 2018).

With 758 million workers China represents by far the world’s largest labor market. Despite relatively low wages by international standards, automation is occurring at an increasing rate in China (Cheng et al., 2019) and has the potential to cause significant labor disruption. Understanding the distribution of skills across the workforce is important to help manage these impacts on individual workers and their communities. However, China lacks a skill taxonomy, which limits the assessment of the distribution of skills across occupations and regions. China’s National Occupation Classification Code (NOCC) includes occupation titles and corresponding task descriptions. By contrast, the US Department of Labor’s Occupational Information Network (O*NET) describes the importance of 161 skills, knowledge and abilities for each of its 967 recognized occupations (www.onetcenter.org). The O*NET database has become essential to labor market research into topics including changing task composition of occupations (Price and Price, 2013), earnings inequality (Acemoglu and Autor, 2011), and the impact of artificial intelligence on labor (Frank et al., 2019; Frey and Osborne, 2017).

The aim of this paper is to build China’s first workforce skill taxonomy and use it to investigate potential skill polarization. We build a Naïve Bayes Model to infer the relationships between O*NET’s occupational tasks and skills, and use the model to predict skills needed across Chinese occupations based on their job descriptions. The resulting skill taxonomy can provide a new perspective for studying the growing regional inequality in China if we view cities as “abstract bundles of workplace tasks and skills” (Autor et al., 2003). The industrial structure of Chinese cities varies extensively. Major cities in China have transformed from manufacturing hubs to service centers, while large numbers of medium and small cities still specialize in mining, manufacturing and farming. This heterogeneous industrial structure suggests varying needs for workers capable of specific workplace tasks, which, in turn, suggests distinct city skill profiles, in which large cities might have significant social-cognitive skill stocks and medium or small cities have sensory-physical skill stocks. Understanding city skill profiles helps policy makers better understand how skills determine a city’s economic performance. For example, targeted policy could reskill workers in a particular

city to help their workforce adapt to emerging labor demands. Therefore, we use the skill taxonomy to compute city skill profiles, to explain the regional inequality of economic growth, and to quantify the polarization of Chinese workforce skills. Our findings based on China’s first workforce skill taxonomy could open a new research agenda for China’s labor market research.

Materials and methods

O*NET discussions. The O*NET database is updated yearly, providing information about the relationship among occupations, skills, and tasks. Additionally, it provides information about the importance of abilities, knowledge, skills, work activities, work context and work values of each occupation on a scale from 1 (not important at all) to 5 (extremely important). Amongst this information, we use abilities, knowledge, skills, work activities to characterize occupational skills (161 in total). According to O*NET, 161 skills are categorized into 15 groups: cognitive abilities, complex problem-solving skills, information input, interacting with others, knowledge, mental processes, physical abilities, psychomotor abilities, resource management skills, sensory abilities, social skills, systems skills, technical skills, work output, and basic skills. The O*NET 23.0 Database is applied in this paper.

We denote $onet(o, s)$ as the importance of skill $s \in S$ to occupation $o \in O$. Some skills are ubiquitous amongst occupations, such as Identifying Objects. Therefore, referring to the benchmark work (Alabdulkareem et al., 2018), we determine whether or not a specific skill is ‘effectively used’ by an occupation by calculating the corresponding revealed comparative advantage (RCA) as depicted in Eq. (1).

$$RCA(o, s) = \frac{onet(o, s) / \sum_{s' \in S} onet(o, s')}{\sum_{o' \in O} onet(o', s) / \sum_{o' \in O, s' \in S} onet(o', s')} \quad (1)$$

We denote effective use of skill s in o using $e(o, s) = 1$ if $RCA(o, s) > 1$, and $e(o, s) = 0$ otherwise.

China’s National Occupation Classification Code (NOCC). NOCC provides us the main material to understand the occupation work content in China, which can be found through <http://ms.osta.org.cn>. NOCC defines occupational titles and job descriptions, but unlike O*NET does not provide data on skills and abilities. Although NOCC lacks detail about workforce skill, it contains task information from the job description, which could be used to deduce what skills a specific occupation needs. The occupation classification has 8 major classes, 75 middle classes, 434 minor classes, and 1481 unit classes. In order to be consistent with the classification of employment data in the 2010 census, we selected the 434 minor classes as the unit of analysis through <http://ms.osta.org.cn/nvqdbApp/htm/fenlei/index.html>.

From task text to skill. An occupation can be conceptualized as a bundle of tasks (Autor et al., 2003; Kok and Weel, 2014) and a task requires a series of skills. Therefore, the relationships between skills and tasks can be inferred based on their co-location probability in the same occupation. We extract $K = 1273$ tasks from O*NET occupation descriptions in the form of word tokens and use $I = 161$ skills from O*NET to build a task-skill relationship matrix $A = (a_{k,i}) \in \mathbb{R}^{K \times I}$, where $a_{k,i}$ indicates the mutual information of the task-skill pair based on their distribution across O*NET’s 967 occupations. The mutual information is a measure of stochastic interdependence between two variables (Darbellay, 1999). The heat map (Fig. 1a) addressing the task-skill relationships demonstrates some tasks are highly related to some specific skills. For example, a skill called Medicine

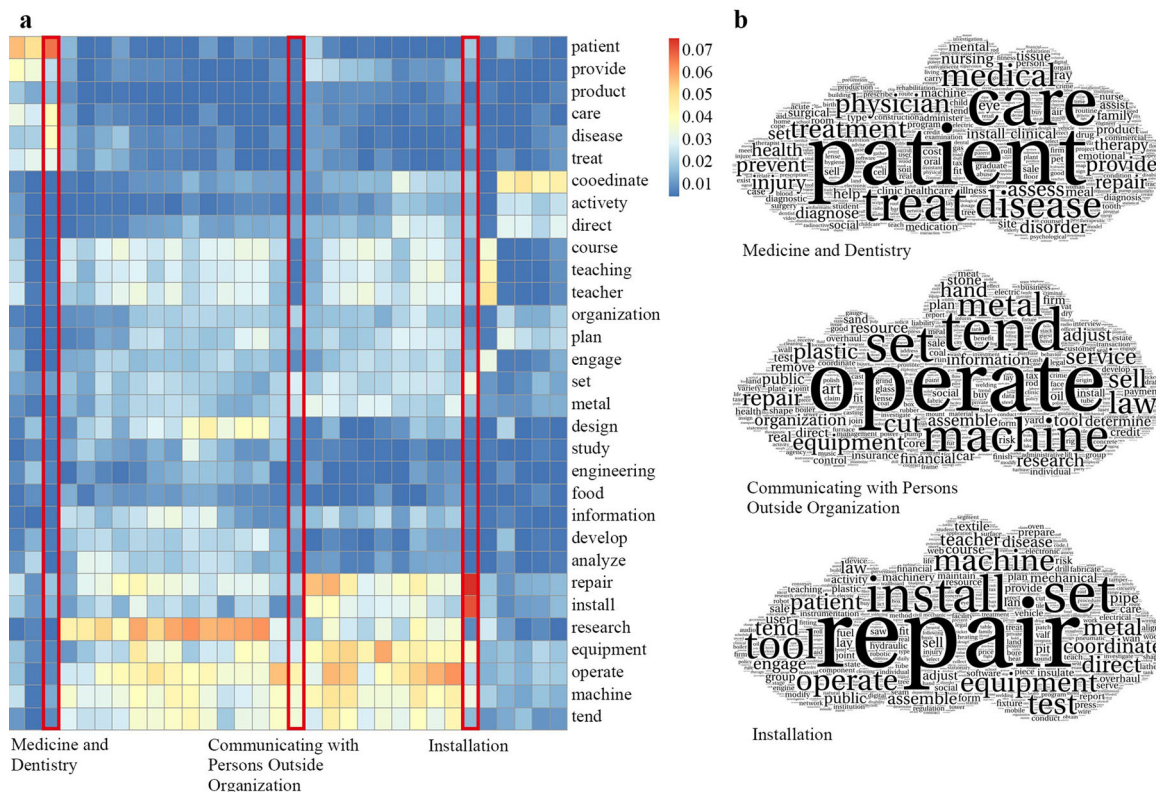


Fig. 1 Mutual information between O*NET tasks and skills. a The heat map is drawn based on the mutual information of the task-skill pairs, showing the heterogenous structure of the task-skill relationships. **b** Task clouds of three skills: Medicine and Dentistry, Communicating with Persons Outside Organization and Installation. The word size is proportional to the value of the mutual information criterion.

and Dentistry, is more likely to co-occur with tasks like patient, treat, care, disease, and medical (Fig. 1b).

Machine learning methods can be used to map the task tokens (labels) to skills (labels) so as to predict whether or not an occupation needs a particular skill to carry out the specific tasks, such as deep learning, random forest, supported vector machine, etc. However, most are correlation analysis rather than causal inference, and deep learning is more of a black box. Compared to these methods, Bayesian approach can unpack the causal relationship between tasks and skills in a much more straightforward and tractable way. The task is the prior condition on which a job depends to apply a particular skill. Therefore, we use Naïve Bayes Model to predict the probability of each particular skill s that an occupation o would employ based on the underlying tasks $t_{i,o} \in K$. Given that NOCC addresses t_i of each Chinese occupation, we can infer the probability of each occupational skill, that is, $P(s^{O*NET} | o^{Chinese})$. To address the relationship between skills s^{O*NET} and occupations $o^{Chinese}$, we decompose the conditional probability as Eq. (2) based on Bayes' theorem.

$$P(s^{O*NET} | o^{Chinese}) = \frac{P(s^{O*NET})P(o^{Chinese} | s^{O*NET})}{P(o^{Chinese})} \quad (2)$$

As occupations are a bundle of tasks, we derive the conditional probability $P(o^{Chinese} | s^{O*NET})$ as Eq. (3), where each $o^{Chinese}$ is regarded as a set of tasks $t_i \in \{0, 1\}$.

$$P(o^{Chinese} | s^{O*NET}) = \prod_{i \in K} P(t_i | s^{O*NET}) \quad (3)$$

A tripartite network (Fig. 2b) is built to link O*NET 1273 tasks, 161 skills and 696 occupations, so the probability of whether or not a certain task t_i would occur given a set of skills s^{O*NET} can be derived (Fig. 2c). Because a task t_i is transversal

across the U.S. and Chinese workforce, the task-skill relationships $P(t_i | s^{O*NET})$ can be applied in the Chinese context, thus solving Eq. (3) and then Eq. (2) as shown in Fig. 2d. The resulting Chinese skill taxonomy addresses occupational skill distributions in binary values. For example, skills like Finger Dexterity, Equipment Maintenance and Near Vision are important to sewing workers but not so important to entrepreneurs. To the best of our knowledge, this represents the first time that Chinese occupations can be quantitatively measured in terms of skill importance. Details about skill distributions of all Chinese occupations can be found in Data S1 in the appendix.

City skill profile. Based on the workforce skill taxonomy obtained in this study, we compute a skill profile for each Chinese city. These skill profiles are comprised of two parts: the numbers of skills effectively used and socio-cognitive scores.

The skills effectively used by Chinese cities. The determination of the skills effectively used by a city involves three steps. First, the skill taxonomy addresses whether or not a particular skill out of 161 skills is effectively used by any of the 353 occupations. We construct a matrix with 353 rows (occupations) and 161 columns (skills). The cell value of the matrix, 1 or 0, indicates whether the corresponding skill is effectively used by the corresponding occupation. Second, we address CS(c,s) which represents the numbers of workers in city c that use skill s using Eq. (4), where census(c,o) depicts the numbers of workers in city c with occupation o and $e(o,s)$ indicates whether or not skill s is effectively used by occupation o . Finally, we can find out which skills are effectively used in the city by calculating the revealed comparative advantage (RCA) of each skill in 102 cities according to Eq. (5). RCA(c,s) means the revealed comparative advantage of a specific

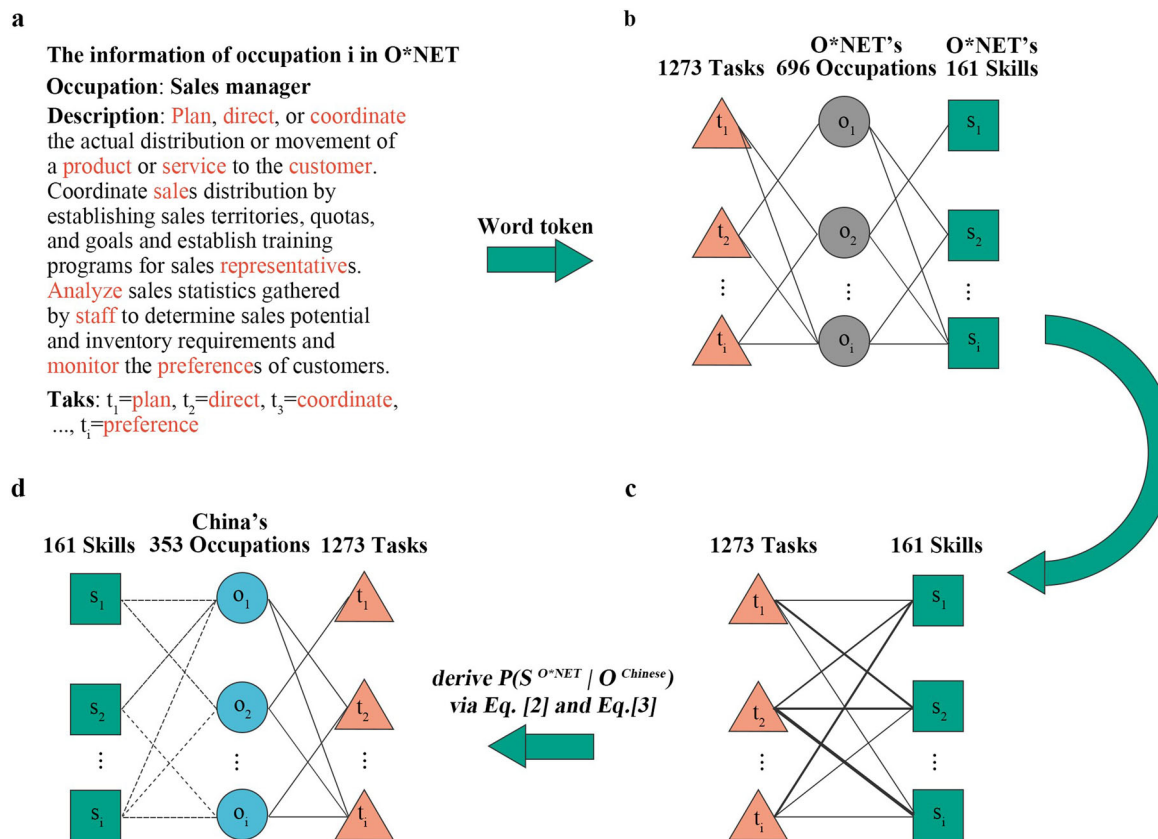


Fig. 2 The workflow of building China's first workforce skill taxonomy. **a** Sales manager job descriptions in O*NET. We derive tasks by extracting word tokens from the job descriptions that are highlighted in red. All tasks can be found in Table S1. **b** A tripartite network links O*NET 1273 tasks, 161 skills and 696 occupations. **c** The relationships between tasks and O*NET skills, represented by the conditional probability of whether or not a particular task t_i would exist given a set of skills s^{O*NET} Deriving the relationships between O*NET skills and Chinese occupations by using Naïve Bayesian inference.

city c in a specific skill s . If $RCA(c,s)$ is >1 , then skill s is used effectively by city c . The higher the value of $RCA(c,s)$, the more city c has a comparative advantage in the skill s . Accordingly, the number of skills effectively used by city c is $Skills_c$ (see Eq. (6))

$$CS(c, s) = \sum_{o \in O} census(c, o) * e(o, s) \tag{4}$$

$$RCA(c, s) = \frac{CS(c, s) / \sum_{s \in S} CS(c, s)}{\sum_{c \in C} CS(c, s) / \sum_{c \in C, s \in S} CS(c, s)} \tag{5}$$

$$Skills_c = \sum_{s \in S^{O*NET}} \begin{cases} 1, & RCA(c, s) \geq 1 \\ 0, & RCA(c, s) < 1 \end{cases} \tag{6}$$

The socio-cognitive scores. Skills are divided into two clusters through community detection (Figs. 4 and S1). There are 97 socio-cognitive skills and 64 sensory-physical skills among the 161 skills in total. The socio-cognitive level of an occupation can be defined as the percentage of socio-cognitive skills to the total number of cognitive skills, that is 97, (see Eq. (7)). Taking the occupation Entrepreneur for example, the number of the socio-cognitive skills is 66, so the socio-cognitive level is 0.68. Second, we determine whether or not an occupation is a socio-cognitive occupation by introducing a socio-cognitive level threshold. In this study, we take 0.6 as the threshold, all the occupations whose socio-cognitive level is higher than that are regarded as the socio-cognitive occupations while the others are non-socio-cognitive occupations. We also tried 0.7 and 0.8 as thresholds to test the robustness of the model, respectively (Table S3). The results are robust in models using different thresholds. Finally, we determine a city's socio-cognitive scores $cognitive_c$ by the proportion of jobs that are socio-cognitive occupations to the total

employment in the city (Eq. (8)).

$$cognitive_o = \frac{\sum_{s \in S^{Cognitive}} e(o, s)}{97} \tag{7}$$

$$cognitive_c = \frac{\sum_{j \in cognitive_o > threshold} census(c, o)}{\sum_{o \in O} census(c, o)} \tag{8}$$

Results

China's polarized labor market. Labor market polarizations have been observed in the U.S. and many European countries since the 1980s (Autor et al., 2003; Alabdulkareem et al., 2018). Accordingly, Chinese scholars have started to pay attention to China's labor market polarization problem in recent years. Lacking the occupation-specific data, scholars can only draw conclusions upon the macro-data, like employment data and manufacturing sector data, or the microdata, like workforce survey. This data gap has produced conflicting results: one study (Lv and Zhang, 2015) observed labor polarization while the other (Du et al., 2017) was not conclusive. Using our skill taxonomy, we revisit this question of labor polarization in China, which is of significance for policymakers.

Figure 3a addresses the skill distributions over six major occupational groups, demonstrating extensive heterogeneities, in which white collar workers, like Department and Enterprise Heads and Technicians and Professionals, heavily depend on soft skills, such as social, mental process, and complex problem-solving. On the other hand, blue collar workers, like Manufacturing Workers, depend on hard skills, such as Psychomotor Abilities, Work Output, and Technical Skills. Moreover, the number of relevant skills vary significantly across occupations, with some

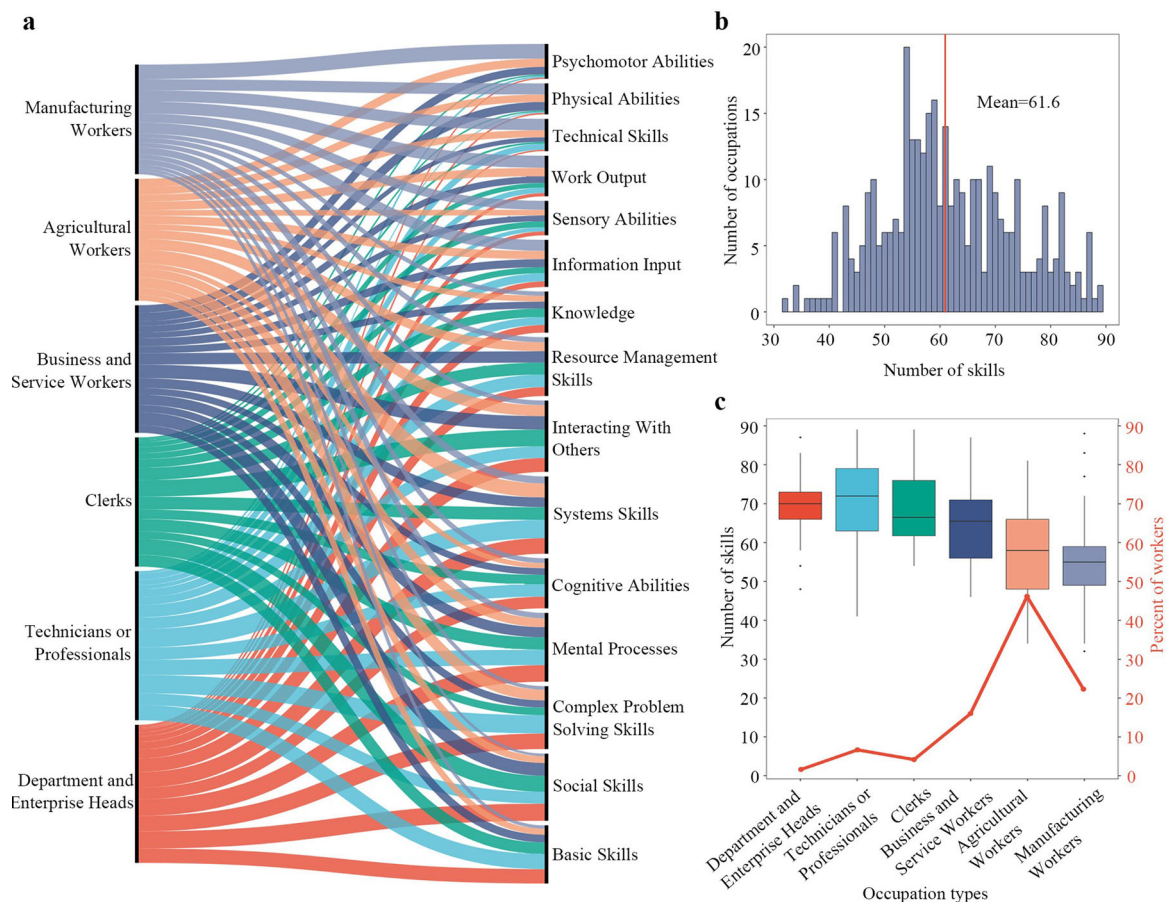


Fig. 3 Skill distributions over Chinese occupations. **a** The Alluvial Diagram indicates the degree to which the occupational groups depend on different skill sets. We aggregate the skill taxonomy into the 6 major occupational groups, and 15 skill sets defined by the O*NET. The thickness of the line is proportional to the aggregated values of skill importance. **b** The frequency distribution shows skill distributions across all 353 Chinese occupations. The vertical red line describes the average numbers of skills that are required by Chinese workers. **c** The box-plot shows the numbers of skills that are important to the six major occupation groups; the red line shows the corresponding employment shares.

occupations requiring more than 80 skills while others require less than 40 skills (Fig. 3b). Technicians and Professionals require the highest average numbers of skills at 71 while Manufacturing Workers and Agricultural Workers both require the least at 55 and 58, respectively (Fig. 3c). Further, while the latter two major groups account for the most significant shares of employment at 23% and 47%, respectively, and their skill contents are primarily composed of skills that are susceptible to automation, the technological unemployment risks are elevated. Additionally, the majority of Manufacturing Workers and Agricultural Workers reside in small and medium cities whose economies rely on one or a few industries and are relatively weaker than megacities. It is more challenging for small and medium cities to implement policies to alleviate automation impacts such as career migration and reskilling.

Since 2016, China has adopted the most industrial robots (Cheng et al., 2019), thus signaling the potential to replace large numbers of low-skilled manufacturing workers. A worker might benefit from reskilling to gain more career mobility and economic opportunity. For instance, according to the World Economic Forum, with appropriate reskilling, there exists a career pathway for assembly line workers transiting to construction laborers, and First-Line Supervisors of Construction Trades and Extraction Workers (BCG, 2018). Therefore, we built a skill space (Fig. 4a) that addresses relationships between skills to understand the career mobility opportunities for all occupations. The skill-pair proximity (the edge of the skill space) is the minimum probability

of the corresponding skill-pair $\theta(s,s')$ co-occurring in the same occupations as depicted in Eq. (9). We denote whether or not skill s is important in occupation o as

$$\theta(s, s') = \frac{\sum_{o \in O} e(o, s) \cdot e(o, s')}{\max\left(\sum_{o \in O} e(o, s), \sum_{j \in J} e(o, s')\right)} \quad (9)$$

We identified two skill clusters using *fast unfolding* community detection in the skill space (Blondel et al., 2008), revealing the skill polarization. Skills such as *Knowledge*, *Social* and *Cognitive* constitute a socio-cognitive cluster and skills such as *Physical*, *Sensory* and *Work Out* constitute another sensory-physical cluster (Fig. S1 and Table. S2). The skill polarization in China aligns with findings in the U.S. (Alabdulkareem et al., 2018). even though the skills linking two clusters, that is the bridging nodes, are significantly different. The transition corridor between the two clusters comprises of skills, such as *Mathematics*, *Judging the Qualities of Things, Services, or People*, and *Estimating the Quantifiable Characteristics of Products, Events, or Information* (Fig. 4a). Workers can master the last two skills by accumulating work experience, but *Mathematics* can only be derived through systematic learning, e.g., attending higher education. However, China’s 4-year degree or higher education attainment rate at age 25–34 is only 14% in 2018, far behind countries of the Organization for Economic Co-operation and Development (OECD) and other developing countries (OECD, 2021), which

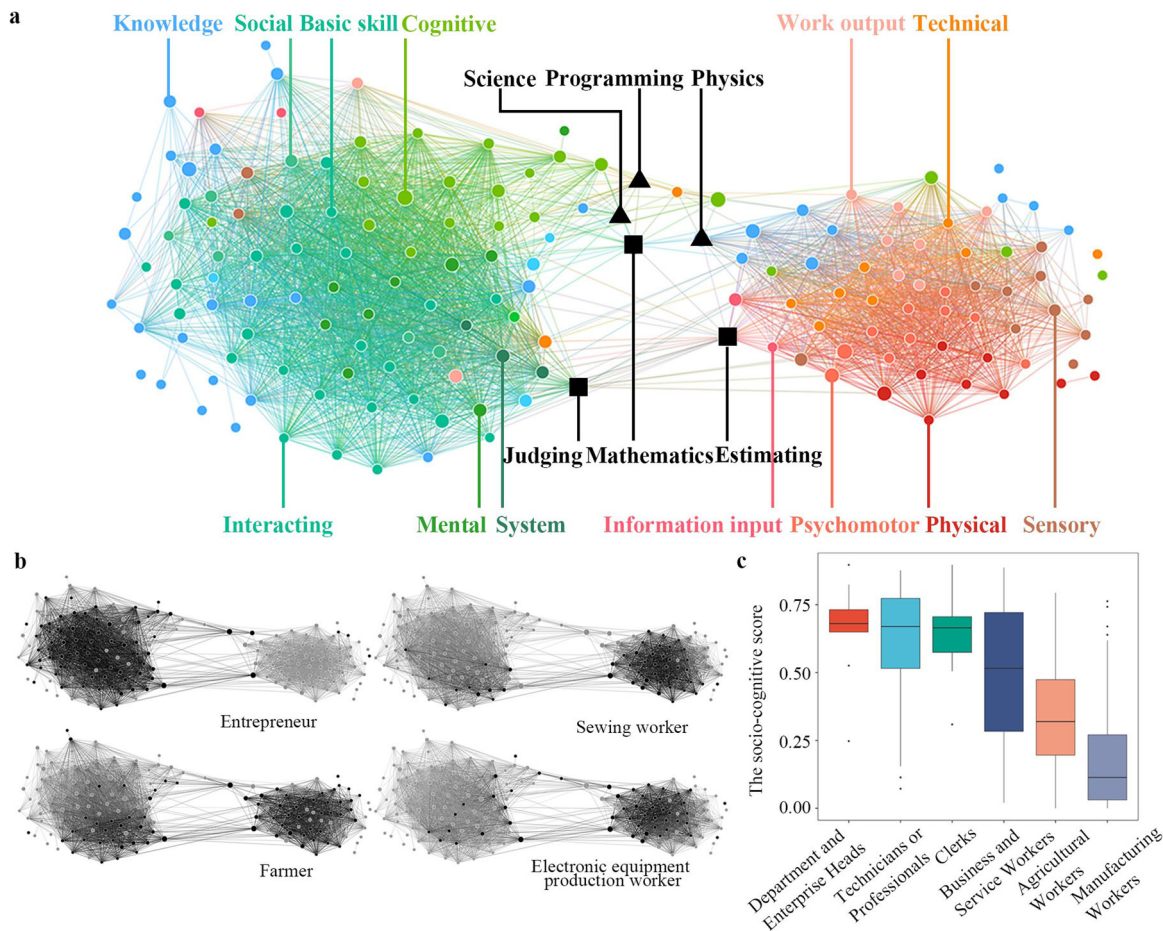


Fig. 4 China's workforce skill space. **a** A skill space addresses the relationships between skill pairs. The edges represent skill-pair proximities that are the minimum probability of the corresponding pair co-occurring in the same occupation. Node colors represent the skill categories defined by O*NET and node sizes reflect the betweenness centrality. We name the cluster on the left "socio-cognitive" skill cluster and the right "sensory-physical" skill cluster, given the types of skills that constitute them. Three skills *Judging the Quality of Things*, *Mathematics*, and *Estimating the Quantifiable Characteristics of Product, Event or Information*, are highlighted in black squares, as they have the greatest betweenness centrality in the network indicating their transition positions between the two clusters. *Mathematics* is a complementary skill to *Science*, *Physics* and *Programming* which are highlighted in the black triangles. **b** Four occupations' positions in the skill space. The darker nodes the more important those skills to the corresponding occupations. **c** The socio-cognitive scores of 6 major occupation groups. Department and Enterprise Heads and Technicians and Professionals have greater socio-cognitive scores than the other occupation groups.

means reskilling and career mobility is challenging for most blue-collar workers. Besides, *Mathematics* is a vital complementary skill to *Science*, *Physics* and *Programming* (Fig. 4a), and all four skills are integral parts of STEM (Science, Technology, Engineering and Mathematics) defining a nation's competitiveness (Atkinson and Mayo, 2010). In China's labor market, they also work as bridging nodes or obstacles of career mobility.

While highlighting some occupations' positions on the skill space (Fig. 4b), we found distinct skill sets in white- and blue-collar workers. For instance, *entrepreneur* has 66 out of a total of 72 skills belonging to the socio-cognitive cluster, while *farmer* has 40 out of a total of 71 skills belonging to the sensory-physical cluster. Reskilling the latter is of particular significance, as it accounted for 41% of total employment in 2010. Moreover, China's rapid urbanization, which would exceed 70% in 2030, would further suppress the demands of this workforce. Fortunately, *farmers* are on top of a few socio-cognitive skills such as *Systems Evaluation* and *Management of Personnel Resources*, which would make career mobility feasible. Over the past decades, industries such as electronic equipment production and textiles have made major contributions to China's industrialization. The most representative jobs in these two industries

are Electronic Equipment Production Workers and Sewing Workers, accounting for 0.3% and 1.7% of total employment respectively (Population Census Office, 2010). Both have skills concentrated in the sensory-physical clusters, so reskilling them is challenging yet of significant importance. Computing career pathways for specific occupations is beyond the scope of this study, but the skill space provides a good starting point.

From the examples above, one might assume blue-collar workers could be reliant on the sensory-physical skills and white-collar workers on socio-cognitive skills. Therefore, we compute the occupation-specific socio-cognitive score based on the proportion of the socio-cognitive skills to a job's overall skills (see Eq. (7) for the computation and Data S3 for socio-cognitive scores for each occupation) to test for such a relationship (Fig. 4c). In the U.S., jobs that have higher socio-cognitive scores tend to have higher wages (Alabdulkareem et al., 2018). Even though China's wage statistics are only available at the major group level, we still observe similar findings. The socio-cognitive jobs like Managers and Professionals (earning 131,929 CNYs and 83,148 CNYs per year respectively) are paid significantly more than sensory-physical jobs like Business and Service Workers and Manufacturing Workers (earning 49,502 CNYs and 50,703 CNYs

per year respectively). Clerks have relatively high socio-cognitive scores, but they earn just 58,211 CNYs per year. The wage statistics use different occupation coding than the NOCC, but they are still comparable. See Table S2 for the detailed wage statistics and explanation. The polarized structure of the skill space provides us with a new perspective for studying China’s wage inequality problem, in which unequal access to education resources is only one part of the story (Li et al., 2015), and the distinct skill sets between the high- and low-paid workers have always been neglected. The skill space also informs two reskilling pathways to bridge the wage gap. However, learning *Mathematics* and all other STEM skills might not be feasible for the overall under-educated workforce unless a more inclusive education system, e.g., vocational training, is implemented.

City skill profile and its effects on economic growth. If we view a job as a bundle of tasks and a task requiring a series of skills (Acemoglu and Autor, 2011), a city, as a container of jobs, can be viewed as bundles of tasks and skills. In this part, we build the Chinese city skill profile using the skill taxonomy and study the regional inequality problem from the perspective of workforce skills.

A city skill profile is defined by two parts; one is the numbers of skills effectively used by the city *c*, i.e., $Skill_c$ (Eq. (6)), and the other is the proportion of socio-cognitive jobs that the city has, i.e., $Cognitive_c$ (Eq. (8)). We determine whether or not a particular skill is effectively used by a city by its importance to the corresponding city over the importance of that skill to all cities. The numbers of skills used effectively by cities range from 55 to 103 out of a total of 161 skills. The range between 70 and 80 skills accounts for more than half of the cities. $Skill_c$ might represent a city’s skill diversity, and we found a significant correlation between skill diversity and GDP per capita (Model 2 in Table 1).

In terms of skill content, some cities like Guangzhou, China’s third largest city, dominate in socio-cognitive skills while others like Putian, famous for its shoemaking industry, mainly rely on sensory-physical skills (Fig. 5a). The city skill profile might address China’s underlying distinct industrial structure, where service centers rely on socio-cognitive skills while manufacturing, mining and farming hubs rely on sensory-physical skills. If a city has a large number of jobs that depend on socio-cognitive skills, which are more resilient to digital disruption, one might wonder if it also has stronger economic performance. Therefore, we calculate the city’s cognitive scores by the proportion of jobs that mainly rely on socio-cognitive skills (Eq. (8)) and use it to

explain economic performance. Beijing has the highest score of 0.52 while Nanyang has the lowest score of 0.09 (Data. S2 for cognitive scores of all cities). Moreover, the cognitive score is consistent with the education levels: a city of a relatively high cognitive score also has a relatively large number of well-educated workers (Fig. S3).

Given education levels were used as a main explanatory variable for economic performance (Schultz, 1961; Fleisher et al., 2010), we use it as the benchmark to explore whether or not skills have stronger explanatory power (Table 1). Model 1 is the basic model covering two main control variables, capital per capita and population density, which explain 61% variation of GDP per capita. Three variables, *skill number*, *university degree* and *social-cognitive score* are added to model 1, respectively, so as to get model 2–4. Model 3 addresses the effect of the education level of cities (*university degree*) on urban economic performance, while model 2 and 4 indicates the effect of skill endowments. The results show that all of them could benefit economic growth, but *social-cognitive score* ($R^2 = 0.761$) has better explanatory power than *university degree* ($R^2 = 0.645$), as does skill number ($R^2 = 0.693$). When *social-cognitive score* and *university degree* are considered simultaneously in Model5, only the *social-cognitive score* is significantly positive. Therefore, city skill profile has stronger explanatory power in explaining economic performance than the conventional explanatory variable education factor. Additional regression analyses on per capita wage show similar findings (Table S4). Interestingly, we found that “sub-provincial or above” cities, which enjoy higher administrative powers than their peers, tend to have higher cognitive scores (Fig. 5b). The “sub-provincial or above” cities, largely comprised of the provincial capitals, are a legacy of the planned economy, in which a city of higher administrative power tends to own more premium resources such as centrally funded universities and state-owned corporation. Those cities normally serve as service hubs within the region for the surrounding areas which are dedicated to manufacturing, mining and farming. In this regard, we might be able to explain China’s regional inequality from a new combined geography and skill perspective, in which there exists a core-periphery layout where provincial capital cities provide socio-cognitive skills and periphery cities provide sensory-physical skills.

Skill effects on labor immigration. China’s regional inequality also stems from differential attraction of migrant workers amongst cities. Population growth has been an important indicator for local government performance from the mid-1990s

Table 1 Estimating effects of education and skill profile on urban economies.

	GDP per capita				
	(1)	(2)	(3)	(4)	(5)
Capital per capita	0.362*** (0.032)	0.304*** (0.031)	0.314*** (0.034)	0.264*** (0.028)	0.278*** (0.028)
Population density	0.076*** (0.019)	0.057*** (0.018)	0.061*** (0.019)	0.053*** (0.015)	0.059*** (0.015)
Skill number		0.008*** (0.001)			
University degree			1.370*** (0.445)		-1.295** (0.504)
Socio-cognitive score				1.021*** (0.132)	1.353*** (0.182)
Constant	3.793*** (0.119)	3.363*** (0.136)	3.860*** (0.116)	3.764*** (0.094)	3.691*** (0.095)
Observations	99	99	99	99	99
R^2	0.609	0.693	0.645	0.761	0.777
Adjusted R^2	0.601	0.683	0.634	0.753	0.767
Residual Std. Error	0.146	0.130	0.140	0.115	0.112
F Statistic	74.873***	71.322***	57.498***	100.807***	81.724***

t statistics in parentheses; GDP per capita is measured in 2010; Skill number is the number of skills effectively used in each city. Capital per capita and Population density take the logarithm. *represent the significance of the effects $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

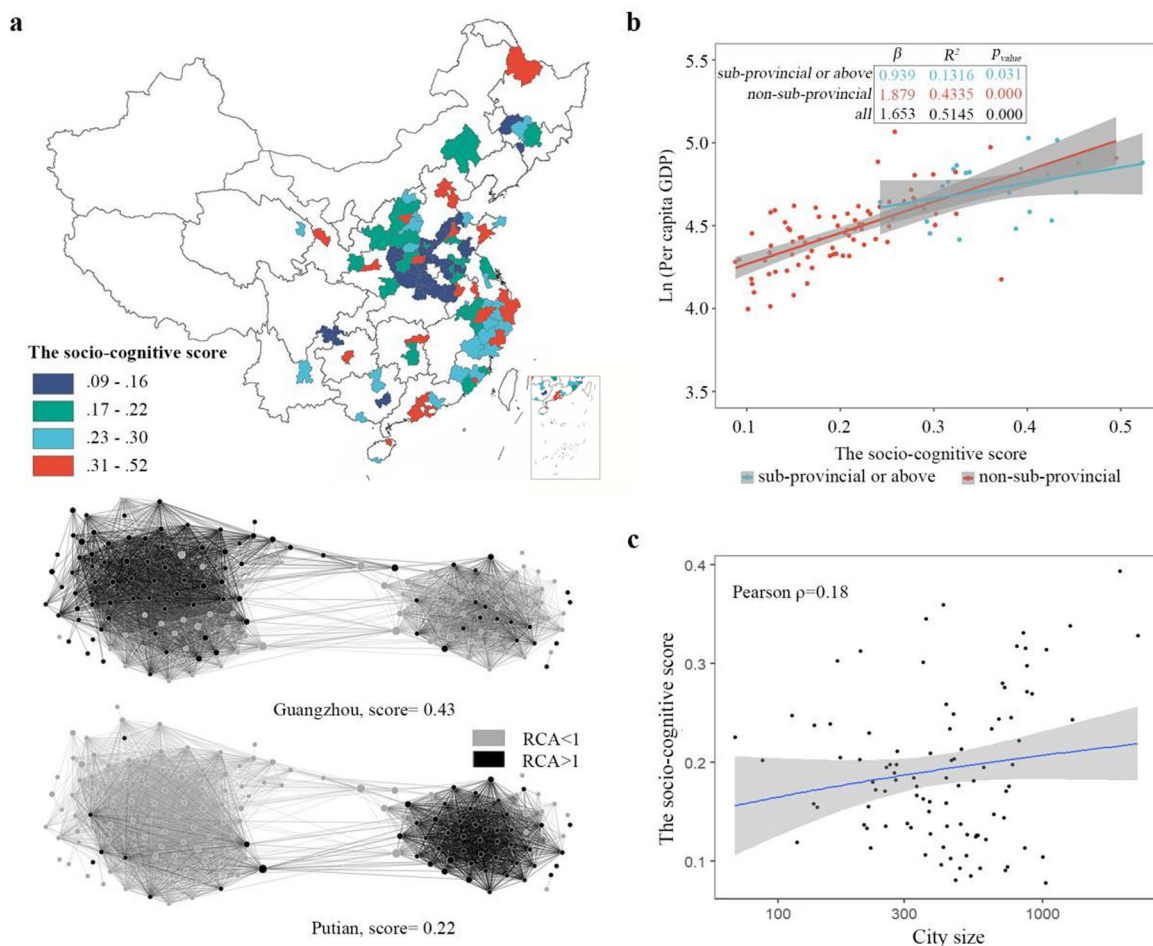


Fig. 5 City skill profile. **a** The skills that are effectively used by a city are highlighted in the skill space. Two cities representing high- and low-level cognitive scores were selected as examples. **b** Strong correlation is found between the GDP per capita and the city cognitive scores. Cities of higher administrative power, so called “sub-provincial or above” cities, highlighted in blue, tend to have greater cognitive scores and GDP per capita than their peers. **c** The scatter diagram shows a weak relationship between city’s cognitive scores and its population size.

(Chan, 2013) when China started to enter its manufacturing export prime. Job opportunity is the primary motivation for immigration. Assuming job opportunities are proportional to the employment size, radiation model is applied to accurately predict the migration patterns amongst U.S. counties (Simini et al., 2012). However, in China, job opportunity is not necessarily proportional to labor market size due to the extensive heterogeneity in local labor market structure. In contrast to the U.S. (Alabdulkareem et al., 2018), in China the size of a city’s labor market little correlation with its socio-cognitive score (Fig. 5c). For example, despite Shenzhen and Nanyang having labor markets of similar size at 10 million workers, their socio-cognitive scores differ significantly, at 0.31 and 0.07, respectively. Therefore, assuming a worker would get similar level of job opportunity in cities of similar employment size would not be an appropriate prior to apply radiation model in China. Instead, we assumed skilled population size is the major human capital of a city, given cities with more skilled workers tend to attract more immigrants (Moretti, 2012; Glaeser and Saiz, 2003). Therefore, we adapt the radiation model assuming job opportunities are proportional to a city’s number of skilled workers to test whether or not migration patterns between cities are better explained. To be more specific, we input the number of college-educated workers, skilled workers and overall employment respectively to the radiation model (Eq. (10)), deriving the predictive number of labor immigrants

between each two cities. We denote the ratio of immigrants from city i into city j to all immigrants from city i as

$$T_{ij} = \frac{m_i n_j}{(m_i + n_j)(m_i + n_j + s_{ij})} \tag{10}$$

where m_i, n_j is the population of city i, j respectively. s_{ij} is the total population in circle of radius s_{ij} , distance from city i to j , centered at city i . Population data is represented by the number of skilled workers, all workers one by one.

We collected daily city-to-city migration data (at 24 July 2019) from Baidu Map® to verify the accuracy of the prediction. As the migration data only shows the top 10 destination cities for each source city, we use the Normalized Discounted Cumulative Gain (NDCG) to measure to what extent the radiation model can predict migration patterns. As showed in Fig. 6a, the average NDCG values of radiation model based on the skilled worker, the educated worker and overall employment are 0.65, 0.67 and 0.61 respectively. It turns out that the model based on skilled worker population is not significantly different from the one based on college-educated workers ($p_{t-test} = 0.19$). However, it is significantly better than the model based on total employment ($p_{t-test} = 0.02$). Taking Tianjin as an example (Fig. 6b-d), both radiation models (using total employment and skilled workers) can successfully predict 7 out of the top 10 destinations. However, when it comes to the rank accuracy, the skilled worker model performs better than the baseline

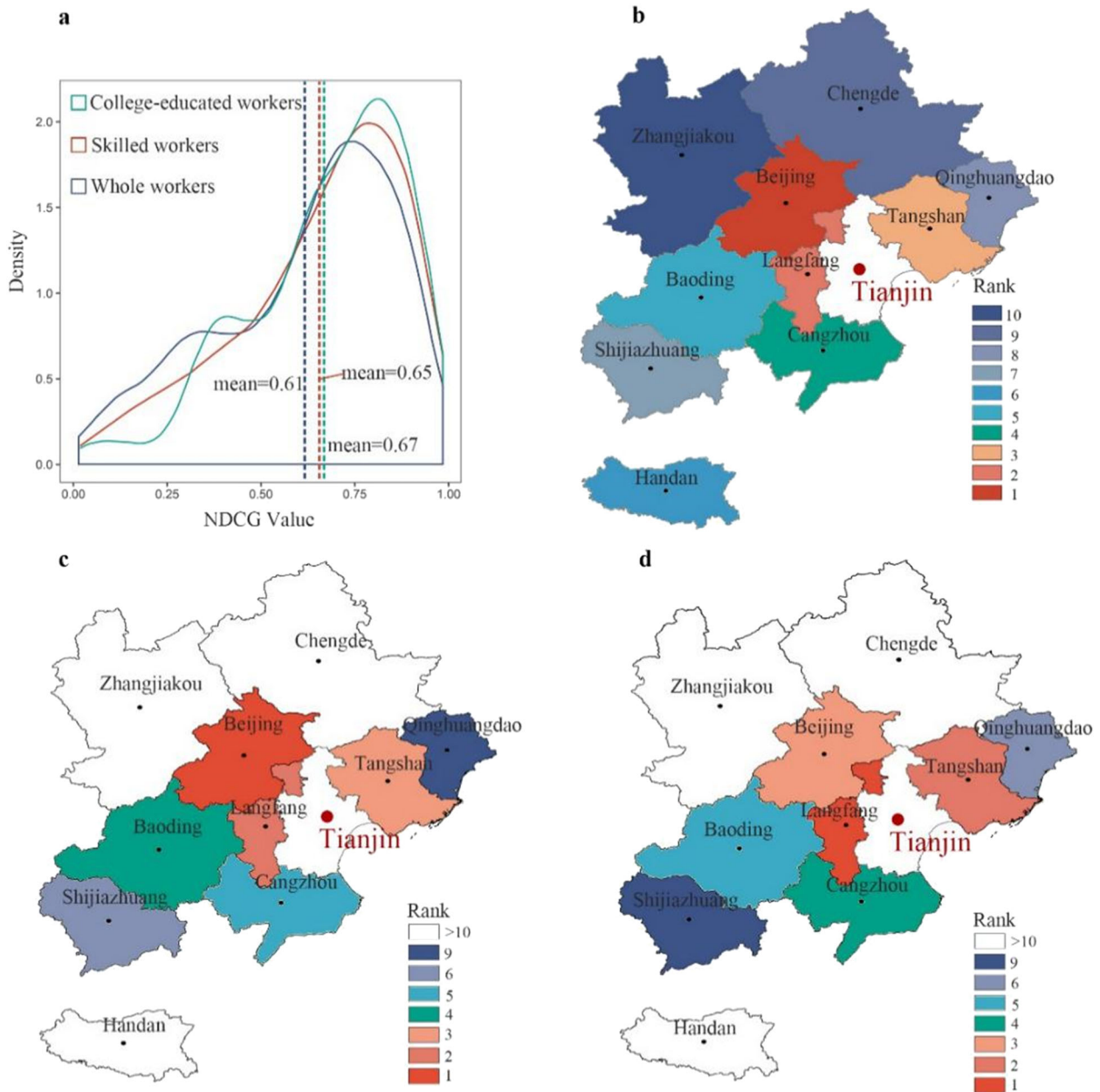


Fig. 6 Radiation models on predicting city-to-city migration patterns. **a** The distribution of NDCG values of the radiation models based on college-educated workers, skilled workers and total employment respectively. **b** The top 10 destinations from Tianjin city based on the daily migration data from Baidu Map®. **c** The predicted destinations from Tianjin city based on the baseline radiation model using total employment. **d** The predicted destinations from Tianjin city based on the radiation model using skilled worker population.

model. Specifically, the skilled worker model predicts Beijing as the very top destination, aligning perfectly with the migration data, while the baseline model predicts Langfang, which only ranks 4th. Despite Tianjin being closer to Langfang than Beijing, more workers will migrate to Beijing due to the latter's larger skilled worker population. Therefore, skilled population could be a better indicator of job opportunity or city attractiveness to predict labor migration.

Conclusion and discussion

China's skill taxonomy could open a new agenda for researching both labor force and regional inequality issues in China. It is striking that our skills taxonomy better predicted city level GDP than even education, suggesting it is a useful research and policy

tool. Applying the taxonomy across cities reveals considerable variation in skills profiles. Consistent with (Alabdulkareem et al., 2018) for the U.S., we find marked skills polarization across at both the occupational and geographical level in China. However, while in the USA it tended to be the smaller cities which had the lowest level of socio-cognitive skills, and hence incomes, the legacy of a centrally planned economy means that in China even relatively large cities can have low levels of these skills.

Over recent decades both employment and wage have tended to be higher for jobs involving socio-cognitive skills. Given the highly uneven distribution of these skills among Chinese cities, this suggests geographic inequality is increasing, which is illustrated by divergent population growth rates among cities (Fig. S5). Those cities which are heavily dependent on socio-cognitive skills are likely to see higher

wages per capita (Table. S4). The stark polarization of skills across occupations indicates that workers may struggle to transition from occupations in declining demand to those for which demand is growing. And even if they could transition, the lack of occupational diversity in many Chinese cities reduces the opportunities to do so.

Cities with a diversified skills base will be more resilient in the face of technological and other social and economic changes. Globally the decline of routine jobs such as in manufacturing has coincided with the growth of the service sector. However, service sector opportunities will be limited in cities with few highly skilled workers to drive demand. This suggests a key priority for policy will be to find ways to help both workers and cities find ways to transition from being dependent on the sensory-physical cluster of skills. A better understanding of occupations and skills in the Chinese context is an important step on this pathway.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data and code used in our study are uploaded onto <https://github.com/Tutu-wq/China-s-first-skill-taxonomy>.

Received: 25 July 2020; Accepted: 30 June 2021;

Published online: 29 July 2021

References

- Acemoglu D, Autor D (2011) Skills, tasks and technologies: implications for employment and earnings. In *Handbook of labor economics*, Elsevier, 4:1043–1171.
- Alabdulkareem A, Frank MR, Sun L, AlShebli B, Hidalgo C, Rahwan I (2018) Unpacking the polarization of workplace skills. *Sci Adv* 4(7):eaao6030.
- Atkinson RD, Mayo M (2010) Refueling the US Innovation Economy: Fresh Approaches to Science, Technology, Engineering and Mathematics (STEM) Education. *Information Technology and Innovation Foundation*.
- Autor DH, Levy F, Murnane RJ (2003) The skill content of recent technological change: an empirical exploration. *Q J Econ* 118(4):1279–1333.
- Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 2008(10):P10008.
- Brynjolfsson E, McAfee A (2014) *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.
- Chan KW (2013) China, internal migration. *The encyclopedia of global human migration*. Blackwell Publishing Ltd., 571–594.
- Cheng H, Jia R, Li D, Li H (2019) The rise of robots in China. *J Econ Perspect* 33(2):71–88.
- Darbellay GA (1999) An estimator of the mutual information based on a criterion for conditional independence. *Comput Stat Data Anal* 32(1):1–17.
- Du Y, Jia P, Cheng J (2017) Structural changes in labor market. *Job tasks and skill demand*. *Stud Labor Econ* 5(3):30–49.
- Fleisher B, Li H, Zhao MQ (2010) Human capital, economic growth, and regional inequality in China. *J Dev Econ* 92(2):215–231.
- Frank MR et al. (2019) Toward understanding the impact of artificial intelligence on labor. *Proc Natl Acad Sci USA* 116(14):6531–6539.
- Frey CB, Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation? *Technol Forecast Soc Change* 114:254–280.

- Glaeser EL, Saiz A (2003) The rise of the skilled city (No. w10191). National Bureau of Economic Research.
- Goos M, Manning A, Salomons A (2009) Job polarization in Europe. *Am Econ Rev* 99(2):58–63.
- Kok S, Weel BT (2014) Cities, tasks, and skills. *J Reg Sci* 54(5):856–892.
- Li H, Loyalka P, Rozelle S, Wu B, Xie J (2015) Unequal access to college in China: how far have poor, rural students been left behind? *China Q* 185–207.
- Lv S, Zhang S (2015) Employment “polarization” in China: an empirical research. *China Econ Q* 12(2):56–62.
- Mokyr J, Vickers C, Ziebarth NL (2015) The history of technological anxiety and the future of economic growth: is this time different? *J Econ Perspect* 29(3):31–50.
- Moretti E (2012) *The new geography of jobs*. Houghton Mifflin Harcourt.
- OECD (2021) Population with tertiary education (indicator). <https://doi.org/10.1787/0b8f90e9-en>. Accessed 29 June 2021.
- Population Census Office (2010) *China Statistics Press*. <https://data.cnki.net/Yearbook/Single/N2012060680>.
- Price B, Price B (2013) The changing task composition of the US labor market: an update of Autor, Levy, and Murnane (2003). Mimeo.
- Schultz TW (1961) Investment in human capital. *Am Econ Rev* 51(1):1–17.
- Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.
- World Economic Forum Boston Consulting Group (BCG) (2018) *Towards a reskilling revolution: a future of jobs for all*. World Economic Forum, Geneva.

Acknowledgements

The team acknowledges the support from the National Science Foundation (No. 41971157). The authors also acknowledge Junhao Jiang and Jingyuan Hu for technology support on the website of China’s skill space.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-021-00862-2>.

Correspondence and requests for materials should be addressed to X.L.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021