UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Are occupations "bundles of skills"?

Identifying latent skill profiles in the labor market using topic modeling

Labussière, M.; Bol, T.

DOI
10.31219/osf.io/5zwmt
Publication date
2024
Document Version
Final published version

License
CC BY

Link to publication

Citation for published version (APA):
Labussière, M., & Bol, T. (2024). *Are occupations "bundles of skills"? Identifying latent skill profiles in the labor market using topic modeling*. (v1 ed.) Open Science Framework. https://doi.org/10.31219/osf.io/5zwmt

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Nov 2026

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

Marie Labussière[1] and Thijs Bol[1]

[1]University of Amsterdam

Preprint Appril 22, 2024.

Abstract

Occupations are a central unit for understanding inequality in the labor market, yet we know little about *why* occupations matter. The existing literature often assumes that occupations are distinct bundles of skills, so skills are rarely conceptualized and measured independently of occupations. How does this limit our understanding of wage inequality? In this article, we use a unique dataset of millions of online job postings in the United Kingdom to capture the skill content of work at the job level and analyze its relationship to existing occupational classifications. While previous literature has often defined different skills as uni-dimensional and independent from each other, we propose a skill profile approach to capture the combination of skills that workers need on the job. Using topic modeling on highly detailed job skill requirements, we identify the skill profiles of job postings and analyze the extent to which they affect wages *within* or *across* occupational categories. Our results reveal substantial heterogeneity in skill content within occupations, and show that both job-level skills and occupations are key to explaining wage differentials across jobs. These findings challenge the often assumed role of occupations as bundles of skills, and offer new perspectives for analyzing labor market stratification.

Keywords occupations; skills; social stratification; topic modeling; online job ads

*This work was supported by the ERC starting grant From School to Career: Towards A Career Perspective on the Labor Market Returns to Education (ID: 950189).

1 Introduction

Since the start of their respective disciplines, sociologists and economists alike have used occupations to understand why some workers get a lot, and others get very little (Blau and Duncan, 1967; Parkin, 1971; Weber, 1979). Designed to group jobs that perform similar types of work, occupations have been shown to underlie various forms of inequality (Haupt and Ebner, 2020), such as wage differentials between workers (Weeden et al., 2007), unequal positions in status hierarchies (Treiman, 1977; Ganzeboom et al., 1992), or differences in work precarity (Kalleberg, 2009).

While differences across occupations are the topic of a long-standing research tradition, much less is known about *why* occupations are so important for understanding inequality in the labor market (Haupt and Ebner, 2020). Several studies argue that formal and informal institutions play an important role in generating inequality in the labor market (Weeden, 2002; Charles and Grusky, 2004; Bol and Weeden, 2015), but a dominant explanation is to be found in the skills used at work. In particular, recent economic research has argued that the key to understanding inequality across occupations lies in the nature of the tasks performed at work (Autor et al., 2003; Acemoglu and Autor, 2010; Goos et al., 2014; Deming, 2017): workers earn different wages because they are not equally capable of performing the tasks required in occupations. And with technological change, wage premiums go to those who have the skills to meet the growing need for analytical, complex and non-routine tasks (e.g., Autor and Handel, 2013; Haslberger, 2021).

This skill-based argument is also present in sociology, albeit less explicitly. Stratification scholars often explain wage inequality across occupations by the differences in the types and levels of skills that occupations require (e.g., Mouw and Kalleberg, 2010; Williams, 2013; Janietz and Bol, 2020) Similarly, research on career mobility emphasizes that the ability of workers to change occupations over the course of their careers depends heavily on the (lack of) transferability of their skills across occupations (e.g., Sacchi et al., 2016; Kalleberg and Mouw, 2018). All of this suggests that workers in the same occu-

pation possess a distinctive set of skills that is relevant for explaining their labor market outcomes, such as the wage they receive or their ability to change jobs. Previous research has thus tended to conceptually conflate the effects of skills and occupations and to use occupations as an empirical proxy for skills (e.g., Liu and Grusky, 2013; Goos et al., 2014; Fernández-Macías and Hurley, 2016; Mouw et al., 2024).

However, the implicit assumption that occupations bundle similar job skills is increasingly challenged by empirical evidence. First, workers appear to transfer a substantial part of their skills portfolios when they change occupations (Poletaev and Robinson, 2008; Gathmann and Schönberg, 2010; Cheng and Park, 2020), suggesting that there are overlaps in the skills required for *different* occupations. Second, the content of jobs is found to vary considerably *within* occupations (Yamaguchi, 2012), even when using detailed occupational categories (Autor and Handel, 2013; Cassidy, 2017; Freeman et al., 2020; Martin-Caughey, 2021). Does this residual heterogeneity matter, or is it just noise? Scholars have seemingly been willing to assume that within-occupation heterogeneity mainly captures irrelevant, idiosyncratic differences between jobs. However, as Weeden and Grusky (2005) acknowledge, the hypothesis that these differences have too little explanatory power to be considered "has yet to be tested" (p.191).

Examining this hypothesis is important not only for empirical reasons. If the skill content of jobs shapes the labor market outcomes of workers beyond of what we currently measure with occupational categories, this would have important implications for how we conceptualize the respective roles of skills and occupations in labor market inequality.

We explore this issue by analyzing the relationship between occupations and skills, and disentangling their effects on wage inequality: Are occupations bundles of skills? And if not, to what extent are the skills used on the job an important predictor of wages irrespective of occupation? Answering these questions requires a comprehensive conceptualization and direct measure of skills used on the job, which has been strikingly lacking in previous literature. We develop the notion of skill profiles to reflect the combination of skills that workers need to perform on the job, and draw on unique data from millions of job postings

to identify them.

Specifically, our study addresses three challenges in the existing literature. First, previous research on skills has tended to focus on narrow sets of skills and to analyze them in isolation. Although this was already noted by Liu and Grusky (2013), recent studies have persisted in analyzing specific skill types as independent factors regardless of their context of use, such as ICT skills (Buchmann et al., 2020), social skills (Deming, 2017), and cognitive skills (Caines et al., 2017; Deming and Kahn, 2018). This ignores that the relevance and meaning of a particular skill depends on the other skills required for the job. For example, the ability to solve problems is unlikely to refer to the same competence for a welder as it does for a manager. To perform their jobs effectively, workers need to master a combination of different types of skills that form a *skill profile* (Poletaev and Robinson, 2008; Yamaguchi, 2012; Liu and Grusky, 2013). We use computational methods on highly detailed job posting texts to create the skill profiles of jobs, as a mix of both generic and specialized skills. This contributes to a better conceptualization and operationalization of the skill content of occupations, recognizing that skills are not independent factors but are "coherently bundled into jobs" (Fernández-Macías and Bisello, 2020, 6).

Second, virtually all existing studies rely on preconceived categories of skills, often created by experts. Since the seminal work of Autor et al. (2003), the literature has been heavily influenced by the dichotomies of routine vs. non-routine tasks on the one hand, and manual vs. cognitive tasks on the other (e.g., Goos et al., 2009; Deming, 2017; Haslberger, 2021). However, this two-factor representation has proven difficult to operationalize (Fernández-Macías and Hurley, 2016; Haslberger, 2022) and may fail to capture important dynamics in workplace skills (Liu and Grusky, 2013). And while more detailed classifications have been used (e.g., Liu and Grusky, 2013), such a priori definitions make strong assumptions about the skill structure of the labor market, by emphasizing certain skill characteristics to the detriment of others. We avoid this pitfall by following and extending previous bottom-up approaches to mapping the skill structure of the labor market (Anderson, 2017; Alabdulkareem et al., 2018; Djumalieva and Sleeman, 2018).

Our methodology based on natural language processing allows us to identify a data-driven classification of skills that reflects both their multidimensionality and interdependence.

Third, existing data provide limited leverage to analyze variation in skill content within and between occupations. Scholars have mostly relied on either expert-coded classifications or employee surveys. Important contributions have been made, for example showing the returns to occupational skills over time (e.g., Liu and Grusky, 2013) or how, even holding occupations constant, skills are still positively associated with wages (e.g., Cassidy, 2017). Yet, both sources of data also have downsides. Expert-coded data, for example by O*NET, assumes within-occupation homogeneity of skills, and cannot be easily updated, which is particularly problematic in times of technological change when skills are expected to change rapidly (see, e.g., Buchmann et al., 2020). Employee surveys solve some of these problems but tend to focus on pre-defined categories of generic skills, providing very little information about specialized knowledge and know-how. To overcome these issues, we rely on real-life job vacancy postings, where employers define the skill requirements for the jobs they are hiring for. We thereby introduce a relatively novel data source and methods to the study of occupations that allow us to examine – at a more detailed level than before – the skill content of occupations and how this affects the wage offered for jobs.

We draw on a unique dataset of online job postings collected by Lightcast, a labor market analytics company (formerly known as Burning Glass Technologies). In addition to providing near-complete coverage of online job postings, Lightcast's data includes detailed information about the characteristics of jobs, including their coding into the Standard Occupational Classification (SOC). Crucially, the data contain granular information about the set of skill requirements listed in the job postings, making it an ideal data source for exploring the underlying skill structure of the labor market without relying on pre-existing skill classifications. We use topic modeling to identify latent categories of skills and describe each job posting as a weighted combination of these latent skill categories, which we define as a skill profile. We aggregate and compare the skill profiles at different levels of the SOC occupational classification to assess the extent to which standard occupational categories

are coherent and distinct bundles of skills. To validate our empirical approach, we revisit the long-standing issue of wage inequality and analyze how the skill profiles of jobs affect the wage offered for them within and across occupations.

The use of job ad data introduces a change in perspective from the previous literature. We focus on the sets of skills that employers require for a given job and how they would reward them, regardless of the skills of the workers who will occupy that job. In this paper, the term "skills" therefore refers to the skills that are required to perform a given job, rather than the skills that workers possess (but may not use). This differs from a common distinction in the literature, which defines skills as characteristics of workers and tasks as characteristics of jobs (see Haupt and Ebner, 2020). Specifically, our definition of skills should be understood as "skills used on the job", which largely overlaps with the notion of job tasks. It is relevant in the context of our analytical approach, which seeks to develop a pure measure of skills that is agnostic to how occupations and workers are socially valued.

The context for our study is the United Kingdom (UK), a country where occupational wage inequality is large (Bol and Weeden, 2015) and has increased in recent decades (Williams, 2013). The UK has a relatively deregulated labor market (Hall and Soskice, 2001b) and a generalist educational system, which is only loosely linked to the labor market. This implies that many jobs do not require specific educational degrees or qualifications (Bol and Weeden, 2015). Accordingly, in our data, the overwhelming majority of employers (>80%) do not specify a preferred qualification requirement. For the current paper, where we analyze skill requirements in job postings, we see this as an advantage: the presence of educational requirements may lead employers to omit certain skills that are tacitly included in formal qualifications (e.g., repairing a boiler for a plumber). Employers need to be explicit about the skills they are looking for, which is why we consider UK job postings to provide relatively high-quality information about skill requirements. The UK is therefore a relevant case for analyzing the extent to which wage differences between occupations can be explained by their similarity in skill content.

Using 6.9 million job postings and computational methods, we generate skill profiles for a representative sample of 600,000 jobs in the UK. We find that jobs require much more diverse skill profiles than is usually assumed, highlighting the need to account for both the multidimensionality of skills and their interdependence at the job level. Further, we show that most of the skill similarity across occupations occurs between rather than within occupational boundaries, challenging the common conceptual conflation between occupations and skills. Importantly, we show that skill content measured at the job level does matter: our comprehensive measure of skill profiles is an important predictor of the wage offered in job postings, complementing rather than supplanting the effect of occupations. This suggests that analyses using occupations as proxies for skills are in fact aggregating jobs that differ significantly in their skill content and are missing an important factor for understanding wage differentials. These findings raise important questions about the nature of occupations and the independent role of skills in generating inequality in the labor market. They challenge our current understanding of wage and job mobility processes and invite the development of new hypotheses about how skills and occupations are intertwined in workers' career paths, which we discuss at the end of the article.

2 Background

2.1 What is an occupation?

In understanding inequality in society, occupations have taken a central position in sociological research. Since the beginning of the discipline (Durkheim, 1893; Weber, 1979), scholars have studied the type of work that people do and its implications for stratification in society. The interests of sociologists coincided with those of governments, which began systematically measuring and categorizing jobs at the end of the 19^{th} Century (Leicht, 2020).

Occupations are typically defined as sets of jobs requiring similar skills (Hauser and Warren, 1997, 180), where skills refer to the ability to carry out the tasks and duties of

a given job. The most widely used occupational classification, the International Standard Classification of Occupations (ISCO), is based on both the type of work performed by workers and the skill level of the job, which reflects the complexity and range of the tasks involved (International Labour Office, 2012). National occupational classifications often follow this line of reasoning and separate occupations by the distinct tasks that are performed in that occupation and their skill level (see, e.g., Office for National Statistics, 2010, 2).

Occupations, as measured and used in social scientific research, are thus explicitly defined as bundles of skills that are used on the job. This is the dominant way they are understood in economics; occupations are used as proxies for the task content of jobs (Autor et al., 2003, e.g.,) or for task-specific human capital (Gathmann and Schönberg, 2010, e.g.,). Sociologists tend to use a broader conceptualization, and see occupations as status groupings (Haupt and Ebner, 2020), with the idea that the role of workers within the division of labor determines their socioeconomic position and the prestige attached to it. This "status dimension" (Avent-Holt et al., 2020) is particularly prevalent in the status attainment literature, which ranks occupations along continuous socioeconomic indices (Blau and Duncan, 1967; Ganzeboom and Treiman, 1996). Such a ranking relies heavily on the functionalist argument that rewards attached to occupational positions mirror the contributions of those occupying them (Treiman, 1977). In the status attainment literature, occupations are positions that transform a person's primary resource (their skills, often defined by their education) into their main reward (wage returns in the labor market). Occupational hierarchies thereby explain but also legitimate inequality in society.

Occupations have also been a central unit of analysis in stratification research where they are used to define social classes. For example, the operationalization of the EGP-like class schemes (Erikson and Goldthorpe, 1992) that dominate the field (Barone et al., 2022), is based on occupations. Scholars disagree about the level of aggregation at which occupations influence social stratification. Research has long focused on the most aggregated levels of the occupational classification, which became known as the "big class"

approach (Erikson and Goldthorpe, 1992). The microclass theory, championed by Weeden and Grusky (2005), also use occupations as their key unit, but advocates the use of detailed occupational categories to "capture the available structure at the site of production" (p.144). Irrespective of the level of aggregation, virtually all class approaches see occupations as central units that capture homogeneous working conditions, which are then linked to labor market outcomes.

Occupations are indeed seen as an important (Avent-Holt et al., 2020; Janietz and Bol, 2020), or even crucial (Mouw and Kalleberg, 2010; Williams, 2013), structural unit to understand wage inequality, and a major structure shaping workers' opportunities for intergenerational labor market mobility (Kalleberg and Mouw, 2018). However, the precise mechanisms linking occupations and inequality are often implicit, with occupations being used as "a placeholder for complex social processes" (Haupt and Ebner, 2020, 21). This ambiguity mirrors the lack of conceptualization of what an occupation entails. Most often, this conceptualization is loosely based on the premise that there exists some natural or objective ordering of skills, and that occupations capture this ordering. With a few rare exceptions (Sakamoto and Wang, 2020; Martin-Caughey, 2021), scholars have taken for granted the definition and measures of occupational categories, without trying to "pin down what exactly occupations proxy" (Williams, 2013, 853).

However, as Braverman (1998) notes in his historical account of occupational categories in the US, "classifications of workers (...) are neither 'natural' nor self-evident" (p.296). In particular, the assessment of the skill level of jobs has historically been derived from normative conceptions rather than empirical investigation. From their inception by the Bureau of the Census, occupations have been classified according to an implicit hierarchy of unskilled versus skilled workers, without regard to the actual distribution of skills within these categories (Braverman, 1998, 297). This has already been noted by Caplow (1954), who shows that the ranking of occupations along a one-dimensional skill axis does not always depend on the degree of skill required, but is rather influenced by cultural norms, such as the superiority of white-collar work over manual work (p.45). This implicit

hierarchy between occupations is still prevalent today: professional occupations are seen as having the highest skill requirements, while those in elementary occupations have the lowest. Such normative understandings of skills are apparent in a number of studies that consider skills only hierarchically, reducing workers' skills to their educational attainment (e.g., Goos et al., 2009; le Grand and Tåhlin, 2013; Oesch and Piccitto, 2019).

Given these conceptual ambiguities, it seems crucial to open the black box of occupations and measure the skill content of jobs directly (Martin-Caughey, 2021). However, researchers rarely have access to disaggregated information on the skills of jobs within occupational categories. Moreover, the frequent conflation of occupations and skills in past research has seemingly relieved researchers of the need to conceptualize how and in what form skills play a role in generating inequalities at work. As a result, while some studies have highlighted skill heterogeneity within occupations (Autor and Handel, 2013; Cassidy, 2017; Freeman et al., 2020), current attempts to directly measure the skill content of jobs face important limitations, related to a poor conceptualization of the notion of skill itself.

2.2 From skills to skill profiles

While sociologists and economists alike have been keen to see occupations as bundles of skills (e.g., Goos et al., 2009; Mouw and Kalleberg, 2010), the precise nature of these skills remains largely implicit. There is no discussion, and even less consensus, about which skills are (most) important for workers' labor market outcomes. On the one hand, since the seminal work of Autor et al. (2003), research has overwhelmingly relied on crude dichotomies to characterize job tasks, such as routine/non-routine or abstract/manual (see Rohrbach-Schmidt, 2019; Schulz et al., 2023, for recent examples in sociology). On the other hand, a few studies have used more detailed skill classifications (e.g., Liu and Grusky, 2013; Cheng and Park, 2020) and explored the relevance of other job characteristics for labor market outcomes, such as task complexity (Caines et al., 2017) or so-called social skills (Deming, 2017). What these different approaches have in common, however, is that they all assume that the skill composition of jobs can be safely reduced to a single or a few

independent dimensions, without considering how skills are embedded at the job level.

Indeed, much of the literature has assumed that skills function in isolation. Researchers typically nominate a few skill categories as particularly useful for measuring the skill content of jobs and analyze them separately from the work context in which the skills are performed. This ignores the fact that the meaning and relevance of a skill depends on the other skills required for the job. Jobs combine tasks of different natures (e.g., cognitive, interpersonal, physical, creative) that can only be performed competently if workers develop and use the right mix of skills. This creates interactions and interdependencies among skills that factor into workers' productivity and wages. As Anderson (2017) shows, we can better understand wage differentials across workers by looking at the combination of workers' skills, rather than looking at them separately. In particular, some skills are synergistic in that they enhance the effectiveness of others, giving workers a comparative wage advantage (Anderson, 2017). For example, programming languages or generalist legal skills can be combined with a wide range of high-value skills, making them particularly profitable in the labor market (Stephany and Teutloff, 2024).

We are certainly not the first to point out the lack of attention to the multidimensionality of skills. As early as 1990, Spenner noted that "many analysts treat the issue as unimportant" (p.402). Other scholars have outlined the need for a "truly multidimensional model of skill" (Liu and Grusky, 2013, 1333) and acknowledged that the core tasks of a job cannot be unbundled in isolated skills (Sostero and Fernández-Macías, 2021; Autor and Handel, 2013, 64). Yet, these studies have often confined themselves to this observation without proposing a corresponding conceptualization or operationalization of multidimensional skills.

In this paper, we shift the focus away from isolated skills and develop the concept of *skill profiles*, which describes the full combination of skills required to perform a given job. This notion recognizes the interdependence and complementarity of skills, by capturing not only which skills are involved but also their specific configuration. Skill profiles further highlight the contextual meaning of skills, recognizing that the same skills may be used

and valued differently depending on the other skills required for the job (e.g., Hurrell et al., 2013). Working with the concept of a skill profile therefore calls for the exploration of the full spectrum of skills that workers develop and use, from the very general or transversal skills that are relevant in many work contexts (e.g., communication and organizational skills) to the highly specialized knowledge and technical know-how that are field-specific.

A skill profile approach also invites us to explore other facets of the relationship between skills and inequality, highlighting mechanisms that arise from the interdependencies of skills at the job level. The limited conceptualization of the relationship between skills in the previous sociological and economic literature indeed obscures important dynamics. The effects of labor market changes are mainly analyzed through the lens of the growing or declining demand for isolated skills, which obscures how these changes can reshape the complementarities of workers' skills and their associated values and returns (Stephany and Teutloff, 2024). For example, Deming (2017) finds that the puzzling stagnation in the demand for high-skilled labor since 2000 masks a growing complementarity between cognitive and social skills: the return to math skills has increased mainly for workers with high levels of interpersonal skills. A skill profile approach seems all the more relevant as recent research suggests a growing demand for skills complementarity in the labor market. There is evidence that new forms of work organization increasingly require workers to develop skills that were traditionally associated with different occupations (Hénaut et al., 2023). Consistently, workers who synergistically combine skills from different domains are found to have the strongest labor market advantage (Anderson, 2017; Stephany and Teutloff, 2024).

Shifting the focus from isolated skills to skill profiles is not just important for capturing new patterns of income inequality. More generally, it renews our understanding of the effects of skills over the life course of workers (Kalleberg and Mouw, 2018). Earlier research has shown that the specific skill combination that workers use on the job shapes their opportunities for mobility between occupations (Gathmann and Schönberg, 2010; Geel and Backes-Gellner, 2011). Yet, prevailing definitions and measures of workers' skill levels,

such as the dichotomy between vocational and general education (e.g., Hanushek et al., 2017), capture only a fraction of workers' skill profiles and emphasize their differences over their similarities. This can lead to both a) an overestimation of skill-based differences between educational groups and b) an underestimation of the potential for occupational mobility. Indeed, past research indicates that skills are more portable across occupations than is commonly assumed (Poletaev and Robinson, 2008; Gathmann and Schönberg, 2010; Yamaguchi, 2012), and that there is more room for upward occupational mobility than actual worker mobility flows suggest (DeMaria et al., 2020). Such findings highlight the limitations of a narrow understanding of human capital and skill mismatches, while challenging the idea that occupations are distinct bundles of skills.

The notion of skill profile builds on previous attempts to map occupations in a multidimensional skill space (Yamaguchi, 2012; Liu and Grusky, 2013) and identify occupational skill portfolios (Poletaev and Robinson, 2008; Geel and Backes-Gellner, 2011). We extend these studies in three important ways. First, we define skill profiles at the job level rather than at the occupational level in order to examine potential heterogeneity in skill content within detailed occupational categories. Due to data constraints, previous research has often examined the skill content of occupations at a fairly aggregated level (e.g., Gathmann and Schönberg, 2010; Geel and Backes-Gellner, 2011), where some heterogeneity is arguably unavoidable. We use granular information about the skill content of jobs to see whether skill similarity within occupations increases with the level of disaggregation.

Second, previous approaches tend to include only transversal skills (e.g., Poletaev and Robinson, 2008; Yamaguchi, 2012), neglecting the role of job-specific knowledge and know-how. This is problematic because jobs requiring similar generic skills may differ significantly in terms of specialized or technical skills. In contrast, we are able to include the full set of skill requirements of jobs to model them as a mix of (different types of) generic and job-specific skills.

Third, we depart from top-down approaches that determine a priori which are the relevant skill dimensions to construct the skill space. While some data reduction is un-

doubtedly necessary due to the intractably large number of skills, most studies assume that a few dimensions are appropriate without empirical investigation.

To uncover the skill profiles of jobs, we take a bottom-up approach that categorizes skills based on their co-occurrence at the job level. In doing so, we build on previous approaches that used network analysis to identify data-driven communities of similar skills (Anderson, 2017; Alabdulkareem et al., 2018; Djumalieva and Sleeman, 2018; Stephany and Teutloff, 2024). These studies have made significant progress in mapping the skill structure of the labor market, but their methodology implies that the division of skills is a partition problem: each skill belongs to one and only one community. This restriction is problematic because a large number of skills may be used differently in various jobs. This is especially true for so-called "soft skills" or "basic skills", which by definition are relevant in a wide variety of work contexts (Burning Glass Technologies, 2015). Imposing the constraint of mutually exclusive categories can lead to arbitrary choices and contribute to underestimating the degree of overlap between skill profiles. Therefore, a more realistic representation of the topology of skills requires a method where the same skill can belong to different skill sets simultaneously, which we present in section 4.

2.3 Occupations, skills and inequality in the labor market

A large body of literature has documented the relationship between occupations and wages. Across highly different countries, studies have shown that a large share of the total wage inequality is found between occupations (Weeden et al., 2007; Mouw and Kalleberg, 2010; Williams, 2013; Bol and Weeden, 2015). These studies estimate that roughly 40 to 50% of all variance in wages is between occupations, and a growing proportion of wage inequality occurs between occupations (see Williams (2013) for the UK, Mouw and Kalleberg (2010) and Weeden et al. (2007) for the US). While the empirical relationship is clear, the mechanisms underlying this association are less well-specified. As discussed above, in much of the literature occupations are relevant to wages only insofar as they encapsulate skill differentials between workers – especially skill differentials that cannot be captured with

standard measures of education (Juhn et al., 1993; Böhm et al., 2024).

With our direct measures of the skill profiles of jobs, we are able to assess whether the relationship between occupations and wages is mainly driven by differences in skill profiles between occupations. If occupations are indeed bundles of skills, then our measure of skill content at the job level should be largely redundant and not allow for a better explanation of the variance in wages than occupations alone. That would be in line with the expectation of Weeden and Grusky (2005) that job-level work conditions have little explanatory power beyond occupations (p.191–192). More generally, it would support the emphasis on between-occupation wage inequality in the literature (Mouw and Kalleberg, 2010; Williams, 2013; Liu and Grusky, 2013), and the aggregation of task or skill information at the occupation level (e.g., Cortes, 2016; Cheng and Park, 2020; Haslberger, 2021).

If, instead, occupations are not homogeneous in skill content, this has important implications for how we understand the respective roles of skills and occupations in generating wage inequality. With respect to skills, the extent to which heterogeneity in skill profiles within occupations is an important driver of income inequality remains largely an open question, given the lack of prior research using detailed job-level information. Existing studies rely on crude measures of job task content (e.g., use of routine, abstract or manual tasks) and find sizable effects of task content net of occupational categories (Autor and Handel, 2013; Cassidy, 2017; Rohrbach-Schmidt, 2019). However, their results show a moderately small, if not zero, increase in the overall predictive power of the models when job-level task characteristics are included (at best, R^2 goes from 0.65 to 0.67 in Autor and Handel (2013)). With a much more comprehensive and fine-grained measure of the skills required at work, we will be able to assess whether these studies tend to underestimate the explanatory value of job-level characteristics due to poor measurement.

But what are occupations, if not bundles of skills? While it is beyond the scope of this paper to fully analyze the nature of occupations beyond their skill content, it is important to mention the sociological literature that treats occupations as much more

than bundles of skills. Here, occupations are defined as institutionalized units that affect a variety of inequality outcomes, including the wage distribution (Parkin, 1971; Weeden, 2002). Recent studies show that occupational wage differentials arise from the ability of occupational groups to erect and defend barriers to entry into labor market positions (Bol and Weeden, 2015; Damelang et al., 2018; Haupt, 2023). These institutional barriers (e.g., occupational licenses, educational credentials) generate economic rents for occupational incumbents; wages in excess of what would prevail in a fully competitive market.

According to this closure argument, occupations affect wages not just because they vary in skill content, but also in institutions that regulate access to them. Research indeed shows that occupational closure affects wage outcomes even when the skill composition of occupations is held constant (Bol and Weeden, 2015; Weeden, 2002). With this study, we are able to assess whether the strong explanatory power of occupations in understanding wage inequality persists when the skill content of jobs is comprehensively controlled for. Hereby, this paper takes an important step forward in clarifying the respective roles of skills and occupations, both analytically and empirically.

3 Data

3.1 Lightcast job postings data

We use a unique source of online job ads collected in the UK by Lightcast, a job market analytics company formerly known as Burning Glass Technologies[1]. Lightcast spiders more than 51,000 job-related websites daily to identify online job postings. After deduplicating the identified job postings, it extracts, parses, and codes their data into a machine-readable format to create labor market analytics (see Carnevale et al., 2014, 3–5 for details). The breadth and level of detail of Lightcast data offer important advantages for our research. First, because of its broad coverage, the dataset is claimed to encompass the near universe of jobs posted online, avoiding reliance on a single job platform (Cammeraat and Squicciarini, 2021, 11). Second, Lightcast data extracts or derives fine-grained information

from job postings, including the job title, salary offer, and occupational code, defined at the four-digit level of the UK Standard Occupational Classification (SOC). Of particular interest to our study, Lightcast extracts the skill requirements from the open text of each job posting, using a dynamic taxonomy of over 32,000 distinct skill keywords.[2] Skill requirements are standardized[3] but unique, offering granular information about the skill profile of jobs. This data is ideal for exploring the underlying skill structure in the labor market without depending on pre-existing skill classifications. Furthermore, unlike DOT or O*NET, Lightcast data allows us to analyze skills at the job level and thus compare the skill profiles of jobs *within* occupational groups.

Lightcast data has been increasingly used in task- or skill-based approaches to the labor market, albeit mostly in economics (e.g., Deming and Kahn, 2018; Hershbein and Kahn, 2018; Modestino et al., 2020). The use of job postings in general has been very limited in sociology (see Salvisberg and Sacchi, 2014; Buchmann et al., 2020, for exceptions), with most studies opting for expert-coded or worker self-report data (e.g., Liu and Grusky, 2013; Fernández-Macías and Hurley, 2016; Martin-Caughey, 2021). Job postings offer a different facet of the labor market, as part of the "official front stage of the recruitment process" (Hamann and Beljean, 2021, 50). Employers write them in such a way as to attract the right candidates for the position, mobilizing explicit categories that *they* perceive as most relevant. Such categories are not only based on the specific requirements of the job but are also shaped by broader organizational factors, norms, and institutions (Sostero and Fernández-Macías, 2021, 9–12). This can lead to a number of biases, such as when employers use job postings to market their company brand (Backhaus, 2004). One potential concern for our study is that the skills listed in job postings may not accurately reflect the skill composition of occupations. However, Sostero and Fernández-Macías (2021) found that job postings do offer a relatively precise picture of what people report doing on the job. For each 2-digit occupation, the authors compared the skill requirements in the UK Lightcast data with the tasks documented by workers in the JRC-Eurofound tasks database. Results show that the skills mentioned in job postings serve as a reasonable

proxy for the task content of occupations. Yet, there is a lower level of overlap between occupations in the Lightcast data compared to the JRC-Eurofound data. Indeed, job postings tend to include the most salient skills for the job, excluding the ones that are implicit or implied (Sostero and Fernández-Macías, 2021, 26–27). This is important to consider while interpreting the results: by design, the job posting data emphasize the differences between occupations rather than their similarities.

The advantages of Lightcast data come with a few limitations. Notably, Lightcast data only includes online job postings, which can provide an inaccurate representation of labor demand. According to Carnevale et al. (2014, 11), between 60 to 70% of job openings were posted online in 2013. Although this share has likely increased over the past decade, recent research shows that Lightcast data is still significantly skewed toward certain occupations (Cammeraat and Squicciarini, 2021). This is because in some industries, such as small business and construction, the hiring process is primarily offline. Elementary and skilled trades occupations are thus typically underrepresented in online job postings, to the advantage of professional occupations (Cammeraat and Squicciarini, 2021). While this lack of representativeness is often acknowledged in previous literature (e.g., Hershbein and Kahn, 2018; Modestino et al., 2020), few studies address this issue in practice (see Blair and Deming, 2020; Cammeraat and Squicciarini, 2021, for exceptions). In our study, using a sample of job postings that is representative of the current occupational structure of the labor market is crucial to avoid the topic model being biased by the over- or under-representation of certain occupations and their associated skill requirements. We therefore use a stratified random sample from the UK Lightcast data, as we describe in more detail below.

Another source of bias stems from the way Lightcast data is extracted and coded. First, characteristics that are explicitly included in the job postings may not be correctly identified by Lightcast's parsers (see Carnevale et al., 2014, 16). Second, some characteristics such as occupational coding are derived or imputed using custom machine-learning algorithms, which can lead to misclassifications. Quality tests showed that the accuracy level

for the most granular occupation classification (6-digit) was relatively low (73%) in the US Lightcast data in 2013 (Carnevale et al., 2014, 16). However, Sostero and Fernández-Macías (2021, 20–21) have shown that inconsistencies in the occupational coding of job titles are primarily due to job titles that are too generic (e.g. "Administrator" or "Project manager"), causing Lightcast's algorithms to infer the occupational code from other elements of the job description. Moreover, as Lightcast's algorithms have been continuously trained on new data, their accuracy has likely improved over the past decade. Nevertheless, it is important to bear in mind that some job postings may be misclassified. We do not expect the measurement error to be greater than in other data sources: the mismatch rates between two independent coders be as high as 50% for the 4-digit occupational level in traditional surveys (Kim et al., 2020).

3.2 Sample description

We use the 2019 Lightcast data, which includes more than 6.9 million job postings. For a portion of these, Lightcast did not identify any skill requirements (N=599,541, 8.6%). The text of these job postings may not mention any skills because the job title or job description conveys sufficient information, or they may include skill requirements that are not recognized by Lightcast's algorithms. While the latter case may result in selection bias, we consider that the large taxonomy used by Lightcast captures the most essential information about the skill content of jobs. This is supported by the analysis of a sample of the original job posting texts, which shows that those without identified skill requirements are typically for jobs that do not require specific qualifications.[4] Finally, we excluded job postings for which Lightcast could not infer the occupational code (N=5220, 0.1%), yielding a total of 6.3 million job postings.

Because the methods used in this paper are computationally intensive, we draw a random sample of 600,000 job postings to perform the analysis, representing about 10% of the original data. This sample size ensures optimal statistical properties and preserves the diversity of skill requirements in the sample: the random sample includes 9,065 of

the 11,583 (78.2%) distinct skill categories present in the job postings in 2019.[5] We use stratified sampling to make the sample of job postings representative of the occupational structure of the UK labor market, using employment figures from the 2019 UK Annual Population Survey (APS).[6] While previous studies have performed comparisons at the one-digit level only (Sostero and Fernández-Macías, 2021; Cammeraat and Squicciarini, 2021), we stratify the Lightcast data at the two-digit level to correct for imbalances within some one-digit groups.[7] Figures S1 and S2 in the Supplementary Materials indeed show that the over- or underrepresentation of some major (1-digit) groups in job postings is not uniformly distributed at the sub-major (2-digit) level. An overview of the variables used in our study is shown in Table 1.

Variable	**Characteristics**	**Coverage** (% of job ads)
Job title	Clean job title, $N = 1,9$ million distinct job titles	100
Occupation	UK SOC Classification (1–4 digits)	99
Skill requirements	Lightcast taxonomy of skills, $N = 11,583$ distinct skill requirements	91
Wage	Minimum and maximum hourly wage	62

Table 1: Variables used in the Lightcast UK dataset, January-December 2019

To analyze the skill content of occupations, we use the skill requirements extracted by Lightcast from each job posting. The list of identified skills is generally quite short, with job postings in the sample having an average of five skill requirements. However, the distribution is skewed to the right (Figure S3): while 75% have 7 or fewer skill requirements, a few job postings (1.6%) have more than 20. The distribution of the number of skill requirements is not uniform across occupations. Occupational groups such as plant and machine operators and elementary occupations tend to feature fewer skill requirements than the others ($Q2 = 2$), especially compared to managers, directors, and senior officials ($Q2 = 6$, see Figure S4). The limited information on skill content in the job posting data – and in some occupational groups in particular – can affect the learning of topics. In the next section, we explain how we address this problem by leveraging the co-occurrence

patterns of skills at the aggregate level.

To analyze skill content within and between occupations, we use the different levels of the UK Standard Occupation Classification (SOC): major groups (1-digit, N=9), sub-major groups (2-digit, N=25), minor groups (3-digit, N=90) and unit groups (4-digit, N=369). In the last part of the analysis, we restrict the sample to job postings that propose a wage (62.2%), which results in a restricted sample of about 373,500 job postings. Descriptive statistics do not reveal significant differences between the restricted and full samples in terms of occupational or industrial composition (see comparison in Tables S1-S2). However, job postings that include a wage tend to offer more stable jobs than those that do not, with more permanent positions (+ 17,7pp) and more full-time contracts (+12.7pp). This means that the results of our wage predictions should not be generalized beyond the available sample – a point we will return to in our discussion.

Lightcast data include both the minimum and maximum hourly wages, but the distributions of these two variables are fairly similar (see Figures S5-S6). A significant proportion of job postings (37.4%) have the same value for both the minimum and maximum hourly wages, suggesting that recruiters listed a single wage rather than a range. We use the logarithm of the minimum hourly wage in our main analysis and include the results for the maximum wage in the supplementary materials. As expected given the similarity of the two variables, the results are very comparable.

4 Analytical strategy

4.1 Identifying skill profiles

The LDA approach

Text mining has been increasingly used in sociology in the last decade (Macanovic, 2022), showing the great potential of computational text analysis for exploring the meaning of large amounts of textual data (e.g., Kozlowski et al., 2019). In particular, topic models can

be used to identify hidden thematic structures in a set of documents – latent "topics"– that are both interpretable and sociologically meaningful (DiMaggio et al., 2013; Nelson, 2021). Our empirical strategy relies on a widely used Bayesian statistical model, Latent Dirichlet Allocation (LDA), which is based on two underlying principles (Blei et al., 2003). First, LDA assumes that in a given text corpus, documents "share the same set of topics, but each document exhibits those topics in different proportion" (Blei, 2012, 79). Documents are more likely to contain certain topics than others, and they also vary in the extent to which they mix different topics. Second, each topic is defined by a distribution over the corpus vocabulary. Words with high probability on the same topic tend to co-occur frequently in the documents – i.e., more often than expected by chance. Following this specification, LDA estimation yields two key outputs: the document-topic matrix, which provides the distributions of documents over the set of topics, and the topic-term matrix, which provides the distributions of topics over the set of words (see Tables 2-3).

The LDA approach is well-suited for analyzing the skill composition of occupations. By using the job postings as documents and skill requirements as words, we can accomplish two objectives: a) identify a data-driven categorization of skills by detecting latent skill topics (henceforth referred to as skill categories) and b) precisely operationalize the notion of skill profile. Regarding the first objective, LDA reduces the complexity of job skills data inductively, based on the relations between skills within the job postings. This data-driven dimension reduction avoids the need to *a priori* decide which skills are the most relevant. In particular, we are able to include both general *and* job-specific skills, rather than solely concentrating on the former (see e.g., Poletaev and Robinson, 2008; Yamaguchi, 2012). Additionally, LDA identifies latent skill categories based on the relations between skills in job postings and their shared context. This approach goes beyond the mere categorization of skills that are similar in standard skill taxonomies, as it also clusters skills that are *complementary*, enabling the study of different combinations of skills (e.g., see the topic "Graphic Design" in Table 3b). Moreover, through the topic-term matrix, LDA provides a flexible classification where skills belong to all categories with different probabilities. This

	t_1	t_2	.	.	.	t_K
d_1	p_{d_1,t_1}	.	.	.	.	p_{d_1,t_K}
d_2	.	.				.
.	.		.			.
.	.			.		.
.	.				.	.
d_D	p_{d_D,t_1}	.	.	.	.	p_{d_D,t_K}

(a) Document-topic matrix, definition

	"Software Engineering"	"Management"	"Graphic Design"
Web developer	0.7	0.1	0.2
Software manager	0.5	0.4	0.1
Graphic designer	0.1	0.1	0.8

(b) Document-topic matrix, illustration

Table 2: Definition (a) and illustration (b) of the document-topic matrix.
Note: We note $d_1, ...d_D$ the D documents, $t_1, ...t_K$ the K topics, and p_{d_i,t_i} the probability that document d_i contains topic t_i. The distribution of each document over the topics adds up to one, i.e., $\sum_{k=1}^{K} p_{d_m,t_k} = 1$ for all topic d_m in $d_1, ...d_D$. The example is based on a fictional case with job postings for three job titles (Web developer, Software manager, Graphic designer) and three topics ("Software engineering", " Management", " Graphic Design").

	w_1	w_2	.	.	.	w_N
t_1	p_{t_1,w_1}	.	.	.	.	p_{t_1,w_W}
t_2	.	.				.
.	.		.			.
.	.			.		.
.	.				.	.
t_K	p_{t_K,w_1}	.	.	.	.	p_{t_K,w_W}

(a) Topic-term matrix, definition

	Budgeting	Creativity	Java	Adobe InDesign
"Management"	0.9	0.1	0.0	0.0
"Graphic Design"	0.0	0.4	0.0	0.6
"Software Engineering"	0.0	0.0	0.8	0.2

(b) Topic-term matrix, illustration

Table 3: Definition (a) and illustration (b) of the topic-term matrix.
Note: We note $t_1, ...t_K$ the K topics, $w_1, ...w_N$ the N words from the corpus vocabulary and p_{t_i,w_i} the probability that topic t_i contains word w_i. The distribution of each topic over the words adds up to one, i.e., $\sum_{n=1}^{N} p_{t_k,w_n} = 1$ for all topic t_k in $t_1, ...t_K$. The example is based on a fictional case with three topics ("Software engineering", " Management", "Graphic design") and a vocabulary of four skill words (Budgeting, Creativity, Java, AdobeInDesign).

departs from the unrealistic assumption that each skill fits into a single category, which is made in virtually all existing skill classifications (e.g., Tippins and Hilton, 2010; Le Vrang et al., 2014; Djumalieva and Sleeman, 2018). Instead, it highlights whether and how the same skill can be used differently in different job contexts.

Regarding the second objective, LDA enables us to represent the job postings in a skill space that is both low-dimensional and meaningful. Specifically, the document-topic matrix describes each job posting by its probabilities of containing the latent skill categories. This reduces the dimension of the skill space from 9,065 – the number of distinct skills in the

data – to k the number of topics (see illustration Figure 1). Besides being parsimonious, this representation of job postings provides a rich description of the type of skills desired by employers. It not only indicates what categories of skill are most prevalent in a given job posting but also the extent to which it mixes multiple skill categories. For instance, certain job postings may mainly contain one skill category (e.g., the Graphic designer in Table 2b), while others may require a hybrid set of skills that span several skill categories (e.g., the Software manager). Such a representation of job postings is consistent with the conceptualization of job skills outlined above; it acknowledges that the execution of jobs depends on a specific combination of skills that cannot be unbundled (Lazear, 2009; Autor and Handel, 2013). We therefore measure the skill profiles of job postings by their probability distributions across the latent skill categories.

The biterm topic model

The strengths of the LDA model also come with an important limitation. To infer topics, LDA implicitly relies on word co-occurrence patterns at the document level. However, when the documents are short, as in the Lightcast data, such co-occurrence patterns are very sparse. This makes it difficult for the model to identify the context in which the words are used, which affects topic learning. To circumvent this problem, we use a variant of LDA designed specifically for short texts by Yan et al. (2013), the biterm topic model (BTM). BTM differs from LDA in two important ways. First, it uses unordered word pairs, or *biterms*, as semantic units, instead of words. For example, a job posting with three skill requirements (Creativity, Animation, Editing) has six associated biterms (Creativity-Animation, Creativity-Editing, Animation-Editing). The use of biterms allows for direct modeling of word co-occurrence patterns (Yan et al., 2013). Second, while LDA captures word co-occurrence patterns in individual documents, BTM aggregates the biterms from the entire corpus to learn the topics. Using the global word co-occurrence patterns solves the problem of sparse patterns at the document level, and significantly improves topic learning (Yan et al., 2013).[8] For a technical presentation of the biterm model

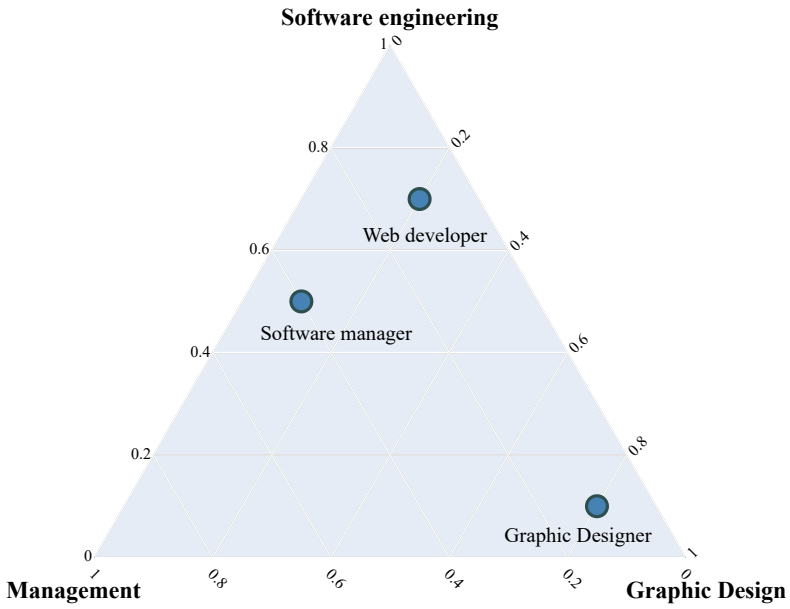


Figure 1: Latent skill space, illustration
Note: We represent the latent skill space defined by the three topics "Software Engineering", " Management", " Graphic Design" defined in Table 3b. This takes the shape of a triangle, each corner of which corresponds to a topic. Job postings can be positioned in this space (blue circles) depending on their distribution over the topics, as defined in Table 2b. The scale on each side indicates the probability of containing each topic, from 0 to 1. For example, the job posting for a Graphic designer is in the bottom right corner because it contains almost exclusively the topic "Graphic Design", with a probability of 0.8.

and a comparison with the standard LDA approach, see section 2.1 in the Supplementary Materials.

Model parameters

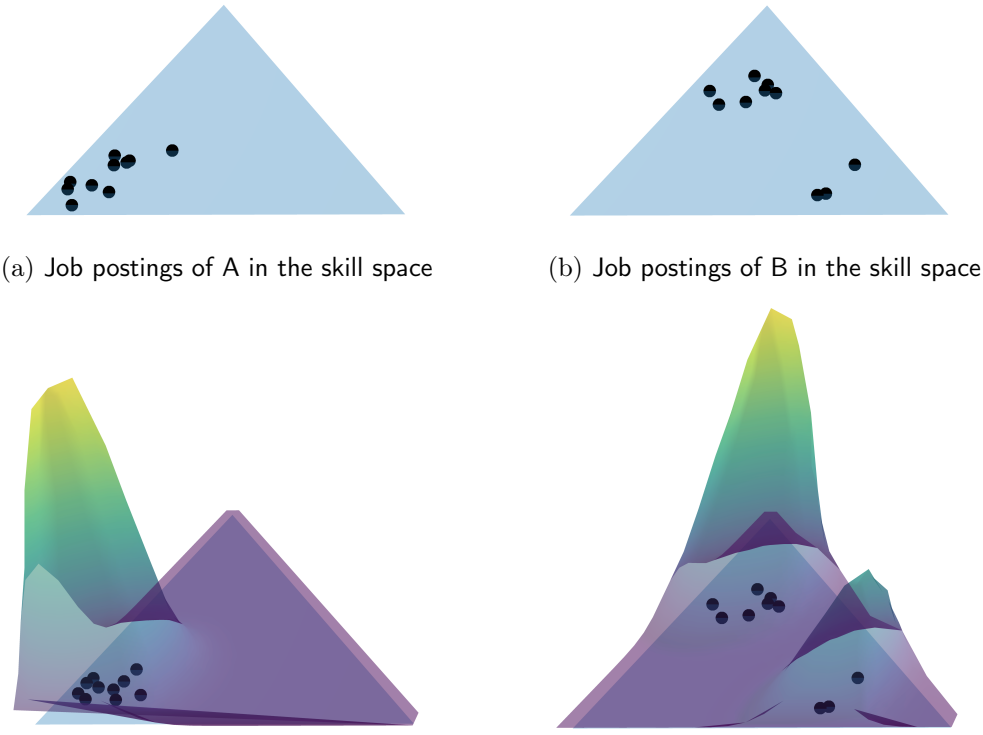
Several factors influence the learning of topics in a given dataset. The most prominent one is the number of topics K, which needs to be specified in advance. As DiMaggio et al. (2013, 582) point out, choosing K is a matter of finding the right lens for interpreting the data, depending on the interests of the analyst. A given number of topics is not inherently good or bad; its relevance depends on the desired level of granularity. The selection of K may be guided by quantitative measures that assess the model's predictive quality, such as perplexity metrics. Yet, optimizing these measures does not ensure semantically meaningful topics (Chang et al., 2009). Other measures to evaluate topic coherence using reference

corpora like Wikipedia (Röder et al., 2015) pose challenges since we lack a suitable benchmark corpus for assessing the interpretability of skill categories.[9]. We primarily evaluated the models based on their capacity to identify a collection of semantically *significant* and *distinct* topics. This was done by manually analyzing the topic-word matrix and using statistical measures to confirm our perceptions. In section 2.1.2 in the Supplementary Material, we explain in more detail how we selected our 19-topic solution and the other parameters of the model.

4.2 Comparing occupations

The biterm topic model enables us to characterize each job posting by their skill profile, i.e., their probability distribution over our 19 latent skill categories. To analyze whether occupations are coherent bundles of skills, we need to aggregate the skill profiles of job postings at the level of their respective occupations. The characteristics of the skill space need to be taken into consideration in this process of aggregation. The job postings' distributions over the topics are indeed defined in a 19-dimension space, which is not Euclidean due to the constraint that all topic probabilities sum to one. This means that standard measures of central tendency such as the mean are incoherent (Greenacre, 2021). Furthermore, the aggregation should not flatten significant variation in skill profiles *within* occupations, since the aim of the paper is precisely to study the potential heterogeneity of occupational categories.

Against this backdrop, we compare the skill profiles of occupations in two steps. First, we construct empirical distributions over the job postings at the occupational level using kernel density estimation. For each occupation, we construct an empirical distribution that gives the probability of occurrence of each skill profile in the job postings of this occupation. These empirical distributions enable us to retain a maximum amount of information on the diversity of skill combinations within each occupational category, without making specific assumptions about the shape of those distributions. Second, we compare these occupational-specific empirical distributions using the Maximum mean discrepancy



(a) Job postings of A in the skill space

(b) Job postings of B in the skill space

(c) 3D empirical distribution of A

(d) 3D empirical distribution of B

Figure 2: Illustrations of the empirical distributions of two occupations A and B, defined on a latent skill space with three topics.
Note: Flat triangles in (a) and (b) represent the latent skill space, with each corner corresponding to a topic (see Figure 1). There are 10 job postings per occupation, whose positions in the skill space are represented by black dots. Using kernel density estimation, we construct 3-dimensional empirical distributions in (c) and (d) that give the probability of occurrence of each combination of topics (or skill profile) in the job postings of each occupation. The color gradient is only there to facilitate the interpretation of the 3D distributions.

(MMD) distance, which can be technically defined as a kernel-based metric on the space of probability distributions (Schrab et al., 2023, 3).[10] The MMD distance is very common in machine learning because of its interesting features: it is flexible, robust, and implementable in high dimensions (Gretton et al., 2012). Intuitively, the MMD distance measures the amount of shared volume between two distributions.[11] The more two occupational categories contain job postings with similar skill profiles, the smaller the MMD distance between them.

Our approach is illustrated in Figure 2 with a latent skill space of three topics. We

distinguish between two occupations, A and B, each defined by 10 job postings with contrasting skill profiles. While the job postings in occupation A are all focused on the left-hand topic (a), the distribution of job postings in occupation B is less concentrated: most job postings contain the upper topic, but a small fraction is more focused on the right-hand topic (b). The empirical distributions shown in the lower panels capture these different patterns, with a unimodal distribution centered in the left-hand topic for A (c) and a bimodal distribution spread over the other two topics for B (d). In this example, the MMD distance between the two occupations is expected to be very small because there is only marginal overlap in the empirical distributions of A and B. We construct such empirical distributions for each occupation on our 19-dimensional skill space and calculate the MMD distance between each pair of occupations to assess the extent to which they have distinct skill profiles.

This approach does not provide direct information about whether the skill profiles are homogeneous *within* occupational categories (cohesion) but rather indicates the extent to which the skill profiles are distinguishable between them (separation). However, we compute and compare the MMD distance between occupations at different levels of the occupational classification to approximate the cohesion of occupational categories. We rely on the idea that if occupational categories are good clusters of skills, the skill profiles in a given sub-occupational category (e.g. sub-major group 21) should be closer to sub-occupational categories in the same occupational category (the major group 2, i.e, 22, 23, 24) than from other sub-occupational categories (e.g. 31, 32...). This mimics the logic behind standard measures of cluster consistency such as the silhouette width (Rousseeuw, 1987), without making specific assumptions about the shapes of the empirical distributions of the occupational categories.[12]

4.3 Predicting the wage offered in job postings

Our topic modeling approach leverages the granular information contained in the job postings to characterize the skill content of jobs. The skill profiles we identify are therefore

better able than previous approaches to measure the different types of generic and specialized skills that employers consider necessary for the job. While this allows us to analyze within-occupation variation in job skills, we need to ensure that the observed differences in skill profiles "capture substantive differences, rather than noise, in job content", to use the words of Autor and Handel (2013, 81). Our precise measure of skill content may capture idiosyncratic differences between job postings that are not economically or sociologically meaningful.

To validate our empirical approach, we assess the effect of the skill profiles of jobs on the wage offered in the job posting. In a standard OLS regression model, we predict the logged minimum hourly wage [13] offered in the job postings by their probabilities (or *loads*) on the K-1=18 topics. This leads to the model:

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

where $log(w)$ is the logged hourly wage offered in the job posting, and p_k is the probability of the job posting to contain the latent skill category k. We use the coefficient of determination R-squared (R^2) to assess the predictive power of the skill profiles given by $p_1, ...p_{18}$. As a benchmark, we also estimate the model with the occupational categories, which are often identified as a major source of wage inequality across workers (e.g., Mouw and Kalleberg, 2010; Williams, 2013):

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

where *soc* is the Standard Occupational Classification (SOC) coding of the job postings. We estimate model (2) with each of the four levels of the occupational classification, from 1 ($n_{soc} = 9$) to 4 digits ($n_{soc} = 369$), resulting in four regressions. In a final step, we use the skill profiles and occupational coding of jobs as simultaneous predictors to analyze the extent to which they provide redundant or complementary information in predicting the

offered wage:

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

As in model (2), we estimate a separate regression for each level of the occupational classification, resulting in four regressions. If, as is often assumed, the share of wage variation explained by occupational categories is mainly due to skill differences between occupational groups, the coefficients of determination of the regressions from model (3) should improve only marginally over those from model (2). Conversely, a significant increase in the R^2 coefficients would provide evidence that the explanatory value of occupations is mainly driven by characteristics *other than* skill content. We therefore take advantage of our detailed and direct measure of skill content to determine whether occupations predict wages either *because* or *independently* of the different skills that jobs require.

5 Analyses

5.1 Describing job ads with latent skill categories

A data-driven classification of skills

Table 4 displays the 19 latent skill categories (or topics) identified by the LDA biterm model, listing the 15 highest-ranked skill requirements (or terms) for each category.[14] To facilitate the interpretation, the ranking is based on the *relevance* of terms for each topic, which corrects the probability of terms for topics by their prevalence in the whole corpus (see Sievert and Shirley, 2014, 66). This limits the overrepresentation of the most common skill requirements in the job ads, such as "communication skills" (present in 24.0% of the sampled job postings) and "organizational skills" (11.2%). The latent skill categories have been labeled by ChatGPT using the top-100 highest term probabilities for each topic.[15]

We observe a few categories of mainly generic skills, such as the categories "Project Management" and "Communication and Interpersonal Abilities". They include both in-

	Latent skill category	**The 15 most relevant skills**, by decreasing order ($\lambda = 0.7$)
1	Project Management	budgeting, project management, planning, communication skills, stakeholder management, people management, building effective relationships, key performance indicators KPIs, quality management, staff management, problem solving, teamwork/collaboration, performance management, change management, organizational skills
2	Office administration and management	Microsoft Excel, administrative support, organizational skills, Microsoft Office, communication skills, detail-orientated, Microsoft Word, secretarial skills, typing, Microsoft Powerpoint, customer service, spreadsheets, data entry, Microsoft Outlook, general office duties
3	Communication and Interpersonal Abilities	communication skills, detail-orientated, teamwork/collaboration, customer service, organizational skills, problem solving, writing, English, verbal/oral communication, time management, Microsoft Excel, listening, French, German, Spanish
4	Sales and Business Development	sales, business development, sales management, sales goals, account management, customer service, product sales, communication skills, business-to-business, prospective clients, building effective relationships, customer contact, client-base retention, telesales, teamwork/collaboration
5	Caregiving and Support Services	teaching, working with patient and/or condition: mental health, care planning, childcare, dementia knowledge, communication skills, nursing home, autism diagnosis treatment care, creativity, planning, English, home management, social services, learning disability, organizational skills
6	Customer Service and Retail Operations	cleaning, customer service, communication skills, cooking, retail industry knowledge, food safety, teamwork/collaboration, organizational skills, stock control, cash handling, food preparation, store management, English, housekeeping, detail-orientated
7	Digital Marketing and Content Strategy	social media, marketing, digital marketing, creativity, marketing management, Google Analytics, market strategy, content management, copy writing, editing, writing, email marketing, budgeting, social media tools, research
8	Financial Operations	accounting, finance, account reconciliation, balance sheet, Microsoft Excel, budgeting, VAT returns, financial reporting, financial accounting, bank reconciliation, accruals, statutory accounts, communication skills, detail-orientated, variance analysis
9	Web Development and Software Engineering	Javascript, Microsoft hash, software development, Java, NET, SQL, Git, software engineering, DevOps, Scrum, Python, active server pages ASP, ASP.NET, continuous integration CI, AngularJS
10	Logistics and Supply Chain Management	procurement, purchasing, supply chain knowledge, logistics, supply chain management, enterprise resource planning ERP, SAP, procurement contracts, planning, manufacturing resource planning MRP, contract management, Microsoft Excel, key performance indicators KPIs, material requirement planning MRP, communication skills

Continued on next page

Table 4: Description of the 19 latent skill categories obtained from the biterm topic model, using the 15 most relevant skills by decreasing order.

	Latent skill category	**The 15 most relevant skills**, by decreasing order ($\lambda = 0.7$)
11	Facility Maintenance	plumbing, preventive maintenance, predictive preventative maintenance, electrical work, carpentry, painting, HVAC, boilers, wiring, communication skills, hand tools, heating systems, customer service, cleaning, emergency lighting
12	Healthcare and Patient Care	patient care, surgery, anaesthesiology, communication skills, working with patient and/or condition: trauma, dentistry, orthopaedics, xrays, critical care, paediatrics, primary care, rehabilitation, gynecology, blood pressure measurement, urology
13	Business strategy	risk management, communication skills, business development, research, due diligence, project management, building effective relationships, analytical skills, economics, accounting, Microsoft Excel, asset management industry knowledge, insurance underwriting, teamwork/collaboration, planning
14	Engineering and Technical Expertise	AutoCAD, mechanical engineering, engineering design and installation, commissioning, engineering design, calculation, civil engineering, project management, mechanical design, Revit, systems engineering, simulation, communication skills, electronics industry knowledge, planning
15	Manufacturing and Engineering	computer numerical control CNC, machining, engineering drawings, welding, lathes, manufacturing processes, quality assurance and control, ISO9001standards, MIG and TIG welding, lean manufacturing, milling cutters, machine tools, computerised numerical control lathes, quality management, problem solving
16	Data Management and Analysis	SQL, python, machine learning, data science, tableau, data warehousing, Microsoft Power BI, extraction transformation and loading ETL, data analysis, business intelligence, big data, Apache Hadoop, SQL server reporting services SSRS, data architecture, Microsoft SQL server integration services SSIS
17	Technical Support and Troubleshooting	Microsoft Active Directory, VMware, Windows Server, Cisco, troubleshooting, ITIL, domain name system DNS, Microsoft Windows, IT support, wide area network WAN, technical support, dynamic host configuration protocol DHCP, transmission control protocol/internet protocol TCP/IP, Linux, Microsoft Exchange
18	Graphic Design and Creative Media	Adobe Photoshop, Adobe InDesign, Adobe Acrobat, Adobe Creative Suite, Adobe Illustrator, creativity, graphic design, Adobe After Effects, digital design, typesetting, animation, editing, creative design, detail-orientated, teamwork/collaboration
19	Scientific Research and Laboratory Work	research, chemistry, biology, clinical trials, biotechnology, biochemistry, molecular biology, experiments, clinical research, cancer knowledge, oncology, drug development, bioinformatics, high performance liquid chromatography HPLC, drug discovery

Table 4 Continued: Description of the 19 latent skill categories obtained from the biterm topic model, using the 15 most relevant skills by decreasing order.
Note: We use the *relevance* of the skills as defined by Sievert and Shirley (2014) to account for the fact that some skill requirements (e.g. communication skills) are listed much more frequently than others, which mechanically increases their probability of belonging to any latent skill category. The parameter λ determines the weight given to the probability of a term under a given topic relative to its marginal probability in the corpus (Sievert and Shirley, 2014, 66). We observed that $\lambda = 0.7$ provides optimal interpretability, but note that the latent topics are easily interpretable without this correction (see Table S3 in the Supplementary Materials).

terpersonal (e.g. people management) and intrapersonal (e.g. problem-solving) skills, as well as basic ICT tools (e.g., Microsoft Word, Excel) and basic activities (e.g., planning, writing, typing). The other latent categories focus on job-specific skill requirements including specialized knowledge (e.g. supply chain knowledge, category 10), specific techniques (e.g. MIG and TIG welding, 15), and know-how (e.g. blood pressure measurement, 12). Specific software and ICT tools often act as discriminating skill requirements, even in non-ICT-focused categories (e.g. AutoCAD in 14, or the Adobe Suite in 18).

Importantly, the latent categories that emphasize specialized and technical skills are also characterized by specific generic skills. Indeed, we can observe that some intrapersonal skills are prominent in some categories but not in others, such as "creativity" for the categories related to marketing and graphic design (categories 7 and 18). This is consistent with the idea that specialized and generic skills are closely intertwined at the job level (Fernández-Macías and Bisello, 2020). Other generic skills are much less distinctive, such as "communication skills" and "collaboration/teamwork", which rank high in most latent skill categories. While employers may include such requirements in their job descriptions for strategic purposes (Backhaus, 2004), the ubiquity of these terms may also reflect the growing importance of so-called "social skills" in the labor market (Deming, 2017).

Table 4 suggests that the latent skill categories vary greatly in the type of skill requirements they contain, yet the table only includes the most discriminating terms for defining the topics. To get a more comprehensive overview, we analyze the composition of latent skill categories by looking at their full distribution over the skills. For the sake of illustration, we use simple dummies from Lightcast that identify two types of skills: software skills and baseline skills. The latter are defined as skills that employers seek across various occupations and that are not typically taught in training programs, such as "organizational skills" or "meeting deadlines" (Burning Glass Technologies, 2015, 7). We refer to skills that do not fit within either of these categories as "occupation-specific".

Figure 3 sums the probabilities of skill requirements for each latent skill category based on their categorization into the three skill types. Even with this crude measure, we can

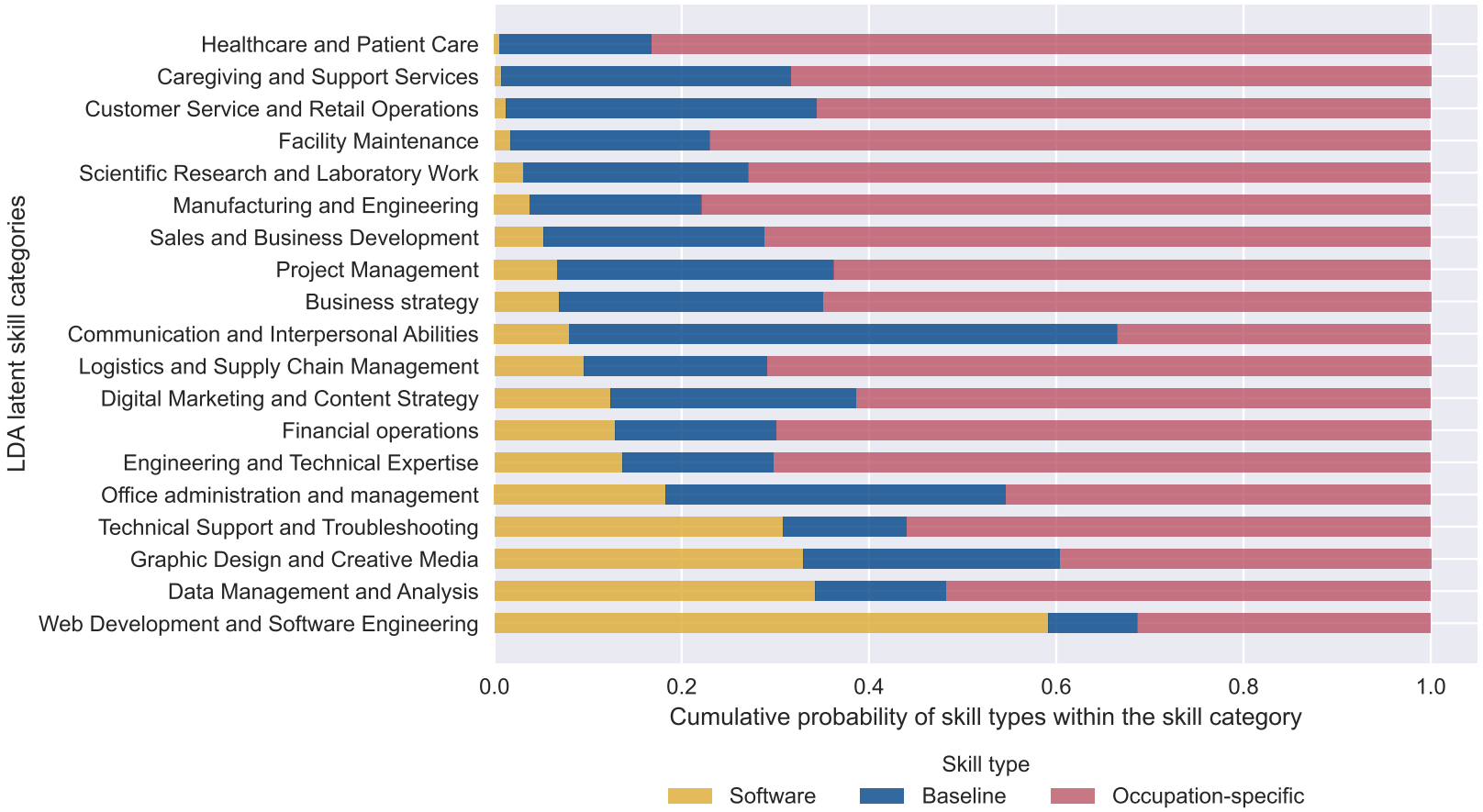


Figure 3: Cumulative probability of three skill types within each latent skill category, based on Lightcast coding of skill requirements.

Note: Lightcast data includes dummy variables for identifying software and baseline skill requirements. For a definition of the latter, see Burning Glass Technologies (2015, 7).

observe important variations in the composition of the latent skill categories. In particular, categories that appear similar at first glance, such as "Web development and Software engineering" and "Technical Support and Troubleshooting", differ substantially in their proportions of software skills, with 0.59 and 0.31 respectively. Another example is the skill category "Healthcare and Patient Care", which is significantly less likely to contain a baseline skill (0.16) than the other health-related category "Caregiving and Support Services" (0.30). This means that the same health-specific skill (e.g., "dementia knowledge") will be more likely to be featured with other expert knowledge in the category "Healthcare and Patient Care" than in the category "Caregiving and Support Services", which contains more non-health-specific general skills. This highlights the heterogeneity of latent skill categories, which are characterized not only by different most likely skill requirements as illustrated in Table 4 but also by how and to what extent they articulate different types of skills.

Jobs as skill profiles

The LDA biterm model provides meaningful latent categories that capture both the similarity and complementarity of skills at the job level. We can then position job postings within this latent skill space. In other words, we can analyze whether and how job postings tend to draw on different latent skill categories. Figure S13 gives the distribution of the highest probability for topics within the sample. The median for the highest probability is 0.54, which means that half of the sampled job postings have a majority skill category that is more likely to be featured than all the others combined. The probability for the most likely category even exceeds 0.8 for about 20% of the job postings, suggesting that they are highly specialized. At the other end of the spectrum, a few job postings (about 5%) have a maximum probability below 0.2, which implies that their distribution across topics is roughly uniform. Overall, these findings point to substantial heterogeneity in the skill profiles of jobs. While half of the job postings focus predominantly on a single latent skill category, the other half mix several categories in varying proportions.

Water Treatment
Cooling Towers
Working With Patient And/Or Condition: Legionella
Equipment / Instrument Sterilization
Organizational Skills
Water Sampling

(a) Extracted skill requirements

Latent skill category	p
Facility Maintenance	0.89
Healthcare and Patient Care	0.03
Office administration and Management	0.02
Customer Service and Retail Operations	0.02
Sales and Business Development	0.02
Project Management	0.01
Engineering and Technical Expertise	0.01
Other skill categories	0.00

(b) Corresponding skill profile

Table 5: Example of a specialized skill profile for a Legionella Risk Assessor, with the skill requirements extracted from the original job posting (a), and the corresponding skill profile (b). p gives the probability that the job posting contains a given latent skill category; categories with a probability below 0.00 are grouped together in the last row.

HSG (Health & Safety Guidance)
Microsoft Word
Site Surveys
Microsoft Excel
Quality Assurance and Control
Working With Patient And/Or Condition
Legionella

(a) Extracted skill requirements

Latent skill category	p
Facility Maintenance	0.39
Office administration and Management	0.25
Project Management	0.19
Engineering and Technical Expertise	0.09
Sales and Business Development	0.02
Communication and Interpersonal Abilities	0.02
Logistics and Supply Chain Management	0.01
Manufacturing and Engineering	0.01
Caregiving and Support Services	0.01
Other skill categories	0.00

(b) Corresponding skill profile

Table 6: Example of a generalist skill profile for a Legionella Risk Assessor, with the skill requirements extracted from the original job posting (a), and the corresponding skill profile (b). p gives the probability that the job posting contains a given latent skill category; categories with a probability below 0.00 are grouped together in the last row.

To further illustrate the differences between specialized and mixed skill profiles, we provide examples in Tables 5-6 from two job postings advertising a job within the unit group 8133 ("Routine inspectors and testers"). The job postings advertise the same job ("Legionella Risk Assessor") and have the same number of skill requirements (n=6), yet their content differs significantly: while the first job posting focuses primarily on technical knowledge and know-how (e.g,. water sampling, equipment sterilization, Table 5), the second includes office tools and emphasizes regulatory aspects, such as the Health & Safety Guidance (Table 6). This suggests that the first position requires on-site work, while the second one involves more office work and coordination. These differences are reflected in the latent skill profiles of the job postings. The first job posting gets a very high load on the topic "Facility Maintenance" ($p = 0.9$), while the probability of the second job posting is distributed over three main topics: "Facility Maintenance" ($p = 0.4$), "Office Administration and Management" ($p = 0.2$) and "Project Management" ($p = 0.2$). This illustration shows that our measure of skill profile captures granular differences in the skill content of job postings, reflecting whether and how they articulate different categories of skills.

Figure S15 shows the marginal probability distributions per topic, which allows us to analyze in more detail how skill categories are distributed within the job postings. As we could expect, the distributions of the generic skill categories (1, 2, 3 in Table 4) have most of their mass between 0 and 0.4, which indicates that they mainly *complement* other categories and rarely characterize a job posting on their own. This is especially true for the category "Communication and Interpersonal Abilities", which has no mass above 0.8. Other distributions (e.g. for 9, 11, 12) follow a U-curve: these skill categories tend to play either a negligible or a predominant role in the definition of job postings. As one might expect, these are the ones that include the most job-specific skills, such as the category "Healthcare and Patient Care" (see Figure 3). In contrast, some topic distributions are fairly flat, which implies that skill categories can either define a job by themselves or complement other skills. This is for example the case for "Financial Operations", which

can be either the primary or secondary activity of a job, such as for a corporate accountant or business manager respectively. Finally, some topic distributions have most of their mass around 0, suggesting that the skill category may only play a role in a small fraction of ads (e.g. "Scientific Research and Laboratory Work").

These findings underscore the relevance of our conceptualization of jobs as skill profiles. First, they highlight the heterogeneous nature of skill requirements: job postings typically bundle different types of skills (e.g., knowledge, know-how, professional attitudes) and vary in their combinations of different skill domains (e.g., office administration, logistics, manufacturing). This heterogeneity is largely hidden when mostly generic skills are considered, as is often done in the literature, at the expense of occupation-specific skills (e.g., Liu and Grusky, 2013; Cheng and Park, 2020). Second, the topic distributions indicate that the same skill category can play different roles depending on the type of job: while some skills are the primary feature of some job postings, they appear only secondarily in others. This confirms the importance of context and supports the idea that jobs differ in the extent to which they require different and/or complementary skills (Anderson, 2017; Stephany and Teutloff, 2024). With the biterm topic model, we can map the skill content of jobs without obscuring these heterogeneous patterns, which may appear not only between but also within occupations.

5.2 Skill profiles within and between occupations

We analyze whether occupations are distinct and homogeneous skill bundles based on the representation of their respective job postings in our 19-dimension latent skill space. We calculate the MMD pairwise distance for all pairs of occupational categories, starting with the most aggregated level of the SOC classification, the nine major groups. Figure 4 shows the distance between each major group with a heat map, where the bluer the smaller and the redder the greater the MMD distance. We can see from the predominant light blue pattern that the skill distributions of the major groups are overall at a similar distance from each other. We mainly observe a distinctive pattern with the major group 6 "Caring,

Leisure and Other Service", which tends to be further away from all other major groups, notably 4 ("Administrative and Secretarial") and 7 ("Sales and Customer Service"). These patterns do not align with the implicit hierarchy of occupational categories, which would imply that the categories furthest apart in the SOC classification, such as the managers and directors (1) and elementary occupations (9), are also the furthest apart in terms of skill profiles. However, we lack a tangible benchmark to assess the significance of these (dis)similarities. Furthermore, Figure 4 does not give information about the variation of skill profiles *within* major groups.

To explore the internal consistency of major groups, we delve deeper into the classification, examining the MMD distance between the 25 sub-major groups. Figure 5(a) provides a simulated heat map to facilitate the interpretation. The simulation is based on the theoretical expectation that major groups are coherent clusters of skills, with sub-major occupations in the same major group being closer to each other than to occupations in other major groups. When we compare the simulated heat map with the empirical one in Figure 5(b), we notice very distinct patterns. Most sub-major occupations are at fairly similar distances from each other, regardless of whether they are part of the same major group. Similar to what we observed in Figure 4, there are a few sub-major groups that stand out and are far away from most of the other sub-major groups. This is particularly the case for occupations 22 ("Health Professionals"), 23 ("Teaching and Educational Professionals"), and 53 ("Skilled Construction and Building Trades"). Conversely, we observe that some sub-major groups are remarkably close to each other even though they are far apart in the SOC classification, such as the Health and Teaching professionals (22, 23) with 61 "Caring Personal service occupations". This is coherent from a skill perspective, as workers in sub-major group 61 are notably described as assisting health professionals and teachers in the care of patients and children.

These first results challenge the view that major groups form bundles of skills. Figure 4 shows no evidence that the skill profiles of 1-digit occupations are well separated, while Figure 5 highlights a low degree of cohesion within these categories when looking at the

skill profiles of the 2-digit occupations they contain. However, it could be argued that due to the high level of aggregation of the major groups, their skill profiles are necessarily highly heterogeneous. In fact, most researchers rely on the 2- or 3-digit level when analyzing the effect of occupations on labor market outcomes (e.g. Mouw and Kalleberg, 2010; Avent-Holt et al., 2020; Haslberger, 2022). We therefore replicate the analysis at a more detailed level and examine the relevance of the sub-major and minor groups for inferring the skill content of jobs.

To assess the heterogeneity in skill profiles within sub-major groups, we examine the MMD distance between the 90 minor groups in Figure 6. As previously, we simulate a heat map (a) to give a visual representation of good cluster quality for major and sub-major groups: minor groups should be very close to minor groups in the same sub-major group, relatively close to other minor groups in the same major group, and further away from minor groups in a different major group. The empirical heat map (b) shows starkly different patterns. While there is evidence of some clusters of minor groups, they do not intersect with the sub-major groups. The most salient example is provided by the minor groups 221 ("Health Professionals") to 231 ("Teaching and Education professionals"), which are not only close to each other but also have similar pairwise distances with all other minor groups. This suggests that these minor groups do share a distinctive skill profile, despite being placed in different sub-major groups (22 and 23). Another example is sub-major group 53, where two out of the three minor groups show clustering patterns. These are the occupations 531 ("Construction and Building Trades") and 532 ("Building Finishing Trades"), which both include different types of construction workers, such as roofers, carpenters, and plasterers. In contrast, occupation 533 groups workers with a supervisory and control role ("Construction and Building Trades Supervisors"), leading to a distinct skill profile.

Other minor groups tend to be far away from most other groups, such as 244 ("Welfare Professionals"), 521 ("Metal Forming, Welding and Related Trades"), and 612 ("Childcare and Related Personal Services"). However, beyond these few examples, Figure 6(b) shows

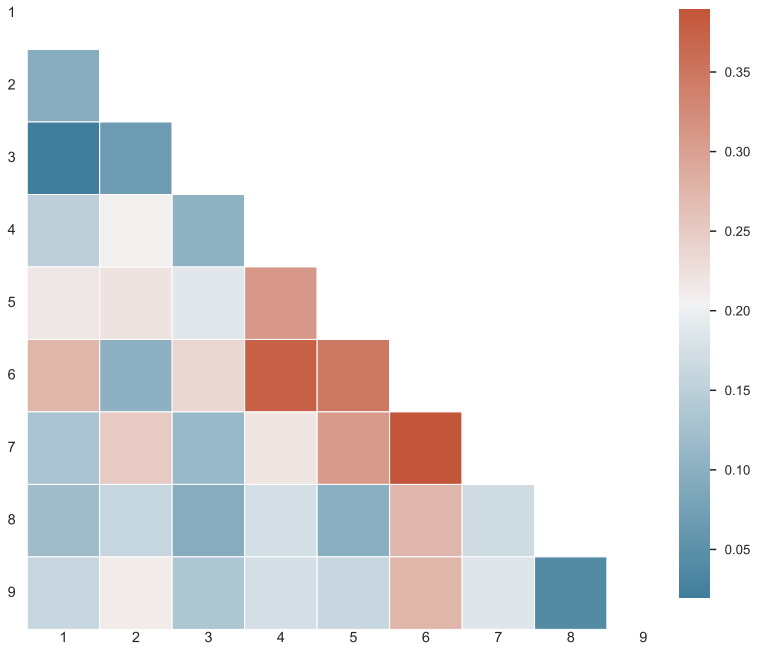
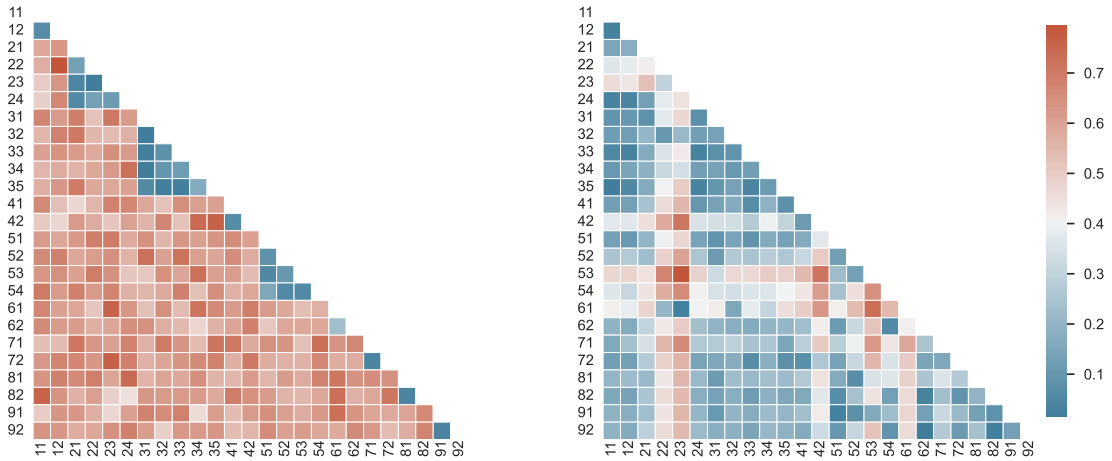


Figure 4: Empirical heat map for the MMD distance between major groups.
Note: each cell represents the Maximum mean discrepancy pairwise distance between major groups, with a kernel bandwidth of $\lambda = 1.49$.



(a) Simulated heat map (b) Empirical heat map

Figure 5: Simulated and empirical heat maps for the MMD distance between sub-major groups.
Note: The x- and y-axes are labeled from the coding in the SOC occupational classification. The simulated heat map (a) represents the theoretical expectation that sub-major occupations in the same major group should be closer to each other than to sub-major occupations in other major groups. It is based on a simulation with normal Gaussian distributions and is for illustrative purposes only. In the empirical heat map (b), each cell indicates the maximum mean discrepancy pairwise distance between sub-major groups based on their representation in the latent skill space. We use a kernel bandwidth of $\lambda = 1.49$.

(a) Simulated heat map

(b) Empirical heat map

Figure 6: Simulated and empirical heat maps for the MMD distance between minor groups.
Note: The x- and y-axes are labeled from the coding in the SOC occupational classification. The simulated heat map (a) represents the theoretical expectation that minor occupations in the same sub-major group should be closer to each other than to minor occupations in other minor groups. It is based on a simulation with normal Gaussian distributions and is for illustrative purposes only. In the empirical heat map (b), each cell indicates the maximum mean discrepancy pairwise distance between minor groups based on their representation in the latent skill space. We use a kernel bandwidth of $\lambda = 1.49$.

that most minor groups are at relatively similar distances from each other. This does not mean that their skill content is comparable, but rather that the extent to which the skill content varies between minor groups remains relatively stable across the occupational structure. In other words, most 3-digit occupations are as (dis)similar to each other when they are in the same 1- or 2-digit occupational category as when they are in different occupational categories. This indicates that the skill structure of the labor market is largely misaligned with the occupational structure as measured by the major or sub-major SOC groups.

Importantly, the same observation holds when we examine the MMD distance between the 369 unit groups to assess the cluster quality of minor groups (see Figure S18): we observe similar clusters to those identified in Figure 5, but they do not coincide with the boundaries of the minor group classification. This indicates that even at the 3-digit level, occupational categories are not a good proxy for the skill content of jobs. While some previous research has argued that there are homogeneity-inducing mechanisms at the detailed occupational level (Weeden and Grusky, 2005, 149), our results suggest that these mechanisms are not related to the type and combination of skills that workers need to perform on the job. On the contrary, the homogeneity in skill content across jobs appears to be related to other dynamics that have yet to be explored.

Taken together, these results provide converging evidence that occupations are neither distinct nor coherent bundles of skills. This is true for both the major groups, which represent a big-class map of the labor market (Erikson and Goldthorpe, 1992), and the detailed minor groups, which come close to the microclass approach championed by Weeden and Grusky (2005). Jobs in the same occupational category tend to require dissimilar skill profiles, which may reflect jobs of varying levels of complexity and authority. This is particularly true when occupational categories include jobs with different levels of seniority, some of which consist of managing or supervising the work of others (see for example sub-major groups 53, 62, 71). Conversely, jobs that are far apart in the occupational hierarchy may still consist of similar bundles of tasks, such as the elementary cleaning occupations

(minor group 923) and the housekeeping and related services (minor group 623).

Our bottom-up approach thus shows that occupations are not reliable categories for conveying information about the skill content of jobs. Furthermore, we show that our latent skill space, agnostic to pre-existing occupational classifications, leads to a different picture of the skill structure of the labor market than is usually assumed. While it is beyond the scope of this paper to analyze the actual skill structure of the labor market (see Anderson, 2017; Alabdulkareem et al., 2018), our results support the idea that workers navigate a dense network of skills across jobs, with substantial overlap in their skill profiles (Poletaev and Robinson, 2008).

5.3 Predicting wage with occupations and skill profiles

Our measure of skill profiles captures granular differences in the skill requirements of job postings. This allows us to measure virtually all observable variation in skill content within occupational categories. However, we do not yet know whether these observed differences between job postings are sociologically relevant. While jobs may require highly specific bundles of skills, to what extent does this translate into unequal socioeconomic outcomes for workers? To test the relevance of our measure, we assess the predictive power of skill profiles on the wage offered in the job postings. If our measure mainly captures idiosyncratic variations between job postings, skill profiles should explain a negligible fraction of the variation in wage across jobs. Table 7 compares the results of the OLS regressions of the logged minimum hourly wage using the equations (1)-(3) defined on p.28.

Model (1) includes the skill profile of job postings, that is their probabilities on the K-1=18 latent skill categories. This can be contrasted with Models (2), which are four regressions including the occupational coding of the job posting for each level of the SOC classification, from the Major (1-digit) to the Unit (4-digit) groups. We find that our measure of skill content accounts for slightly less than a third of the variance in wages across job postings ($R^2 = 0.30$). This level of predictive power is similar to that of major ($R^2 = 0.28$) and sub-major groups ($R^2 = 0.32$) but is less precise than the most detailed

levels of the SOC classification. Specifically, the 369 unit groups explain about 40% of the variance in the minimum hourly wage. This indicates that improving the predictive quality of Model (1) comes at the cost of a sharp increase in the number of parameters, expending large degrees of freedom. Overall, these findings show that skill profiles capture relevant differences between job postings and compete well with traditional measures. At the same time, they provide much more direct information about the skills people need to perform on the job.

Our estimates also confirm the strong predictive power of occupations. They align with studies in other contexts, such as the work of Mouw and Kalleberg (2010), who found that the 469 US Census occupational codes accounted for 43% of the wage gap between workers in 2008. However, the extensive literature on the explanatory power of occupations has tended to leave aside the question of *why* occupations are important in the first place (Williams, 2013, 852). Our detailed measure of skill content gives us a unique opportunity to better understand what occupations actually proxy. In particular, it allows us to assess whether part or most of the effect of occupations on wage inequality is attributable to differences in skill content between occupational groups, as is often assumed (e.g., Goos et al., 2009; Liu and Grusky, 2013). If occupations mainly proxy the type of skills and tasks that are required in jobs, skill profiles should mostly add redundant information to occupations in predicting wages.

Models (3) include the skill profiles of the job postings as well as their occupational coding for each level of the SOC classification. The skill profiles increase the R^2 from 0.29 to 0.40 when the major groups are used as predictors, representing a 40% increase in the predictive power of the model. The contribution of skill profiles decreases as the level of granularity of the SOC classification increases. Yet, the skill profiles still increase the R^2 of the minor and unit groups by 25% and 16% respectively. Likelihood ratio tests confirm that the addition of skill profiles significantly improves the goodness of fit for all levels (Table S6). Regarding the regression coefficients, they decrease moderately when compared to models (1) and (2), but remain in the same order of magnitude (see the

		Model (1) *Skill profiles*		**Models (2)** *SOC coding*		**Models (3)** *(1) and (2) combined*	
		R^2	Df	R^2	Df	R^2	Df
		0.301	18				
SOC level	Major groups			0.289	8	0.403	26
	Sub-major groups			0.320	24	0.414	42
	Minor groups			0.356	89	0.432	107
	Unit groups			0.396	368	0.460	386

Table 7: Adjusted coefficient of determination R-squared (R^2) and degrees of freedom (Df) for the prediction of the logged minimum hourly wage included in job postings, comparing the models (1)-(3) defined on p.28.
Notes: Degrees of freedom represent the number of effective parameters in the models.

full models in Tables S4-S5). Overall, these findings suggest that the amount of shared information between skill profiles and occupations is relatively limited. Contrary to what we would expect if occupations would mainly capture differences in skills and tasks between jobs, we find that occupations and skill profiles are *both* important predictors of wages.

Taken together, these results provide two main insights. First, a direct measure of the skill content of jobs significantly improves the prediction of wages compared to including the occupational code alone. Strikingly, combining the skill profiles with the 9 major groups has as much predictive power as the 4-digit unit groups alone, with a much smaller number of parameters ($Df = 26$ vs. 468). This supports the idea that our latent skill space captures relevant differences in the skill content of job postings that translate into different socioeconomic outcomes for workers. This is also consistent with the theoretical expectation that wage differentials between workers at least partly arise from the different skills required by their jobs (Haupt and Ebner, 2020), although the moderate overall predictive quality of our models ($R^2 = 0.46$ at most) suggests that other important predictors of wages remain unobserved.

Second, occupations and skill profiles appear to be mainly complementary predictors of wages, providing distinct rather than redundant information. This confirms earlier studies showing, with less precise measures, that job tasks at the individual level are informative

about wages even within detailed occupational groups (Autor and Handel, 2013; Cassidy, 2017; Rohrbach-Schmidt, 2019). Our results suggest that while occupations were designed to group sets of jobs with highly similar tasks and duties (International Labour Office, 2012, 11), these categories are only weakly related to the very tangible skills that jobs require. As we will elaborate in the discussion, this raises the important question of what occupations capture, if not differences in skills and tasks across jobs.

5.4 Limitations

The Lightcast job posting data used in this paper provide valuable insights into the skill content of jobs. However, it is important to note that these data were not initially developed for research purposes and have several limitations. First, the data indicate whether a particular skill is required for a job, but do not provide information on the level of skill intensity or ability required. This implies that we consider skills in a binary way, ignoring the fact that jobs with similar skill requirements may necessitate varying levels of proficiency. We measure skills at such a granular level that we do not expect this to introduce substantial bias, but we cannot rule out the possibility that we are underestimating both the heterogeneity in skill content within occupational categories and the predictive power of skill profiles.

Second, our data capture only the demand side of the labor market. We measure the skill content of jobs as it is described by employers, which may not correspond to what jobs may require in practice and may be even further removed from what workers actually do on the job. More generally, we have largely taken skills for granted, as if they were objectively and unambiguously defined in relation to the task content of jobs. Yet, this neglects the role of institutions in shaping how employers perceive and use skills to construct a given job, and how workers respond to them in return. Our exploratory approach has shed light on the great heterogeneity of skills within occupational categories, but falls short in contextualizing how the skill profiles of jobs are created and socially validated. We see this as a very fruitful avenue for further research, as we might expect the skill content of jobs

to vary greatly depending on the institutional context in which they are embedded. For example, the presence of unions, occupational licensing, or gender norms may all affect the definition of skills and their strategic use in the labor market (Bol and Weeden, 2015; Martin-Caughey, 2021).

Third, the wage offered in job postings may differ from the wage paid to incumbents. Our wage equations should therefore be interpreted with caution, as the mechanisms generating wage inequality between jobs may differ from those generating inequality between workers. For a given wage offered, workers have different bargaining power depending on the scarcity and suitability of their skill profiles. Replicating this study with employee surveys would therefore be worthwhile, provided that they include detailed task data. Moreover, our wage equations are only estimated for a subset of job postings that do propose a wage. As discussed in the data section, these jobs postings tend to advertise more permanent, full-time contracts, which could affect the relationship between skills, occupations, and wages. While we have no strong intuition about how this might bias the result, more comprehensive data on wages could put our findings on a firmer footing.

Finally, it should be noted that our method is context-dependent, which is both a strength and a weakness. On the one hand, our data-driven approach frees us from preconceived notions about how skills are related. On the other hand, our latent skill classification is based on and shaped by the particular pool of job postings under consideration. In particular, our skill profiles are defined at the job level, which complicates the export of our measures to other databases and their comparison with more traditional measure of skill or task content. This problem is common to other studies that take a bottom-up approach to skills (e.g., Anderson, 2017; Stephany and Teutloff, 2024). Furthermore, our high-dimensional measure of skill profiles may require too many degrees of freedom for standard sample sizes. Thus, the next step in extending this exploratory research is to understand how best to reduce the complexity of skill profiles and identify more parsimonious measures for social science research.

6 Discussion and conclusion

6.1 Key substantive findings

In this paper, we have argued that despite the widespread use of occupational classifications for understanding socioeconomic inequalities, we still know very little about *why* occupations matter and *what* they proxy. A common – though not always explicit – assumption is that occupations group jobs that require their workers to use similar skills. However, a growing literature has highlighted substantial heterogeneity *within* occupational categories (e.g., Autor and Handel, 2013; Martin-Caughey, 2021). Drawing on a representative sample of 600,000 online job postings in the United Kingdom, we used computational text analysis on detailed job skill requirements to directly measure the skill content of jobs and to assess how this varies within and across occupations. We found no evidence that occupations are bundles of skills: even detailed occupational categories do not consist of homogeneous or distinct skill profiles. Moreover, our novel and detailed measure of skill content explains substantial wage differences across jobs, *complementing* rather than *substituting* for the effect of occupational categories.

Our data-driven and bottom-up approach to defining skills sheds new light on the nature of skills in the workplace. We identified latent skill categories that capture the diversity and interdependence of skills at the job level (Anderson, 2017; Stephany and Teutloff, 2024) and analyzed how jobs fall within this high dimensional skill space. Previous research has often reduced the skill content of jobs to the presence or absence of a few predetermined and isolated skills (see Liu and Grusky, 2013). We showed that most jobs require different sets of skills, all of which mix specialized knowledge and know-how with general skills to varying degrees. In fact, only half of the jobs in our sample rely primarily on a single latent skill category, highlighting the need to consider not only differences in skill content but also *whether* and *how* jobs combine different skill sets.

These findings provide compelling empirical support for our conceptualization of *skill profiles*. Rather than reducing the granularity of skills to a few isolated factors, our notion

of skill profile emphasizes the interdependence and complementarity of the skills used by workers. Topic modeling proved to be a crucial tool to identify skill profiles, as it does not assume a fixed categorization, but instead directly models the context dependency of skills and their combinations at the job level. While several studies have argued that it is important to look at combinations of skills (Poletaev and Robinson, 2008; Geel and Backes-Gellner, 2011), one of the major contributions of this article is that we provide an empirical measure that is fully consistent with its theoretical definition. Not only does our approach allow for an unprecedented level of detail in the description of job skill content, it also strengthens the link between theory and empirical evidence (Lundberg et al., 2021).

Based on our direct and comprehensive measure of skill content, we showed that occupations cannot be adequately described as isolated bundles of skills. We compared the skill profiles of occupations at different levels of the occupational classification and showed that occupational categories are neither distinct nor internally consistent clusters of skills. While a few detailed occupational categories exhibit unique and coherent skill profiles (e.g., care occupations, or construction work), most of the observed skill similarity occurs between, rather than within, occupational boundaries. This again highlights the added value of using a bottom-up approach to describe the labor market, as it does not make prior assumptions about which skills are relevant to differentiate jobs. We thus contribute to opening the "black box" of occupations (Haupt and Ebner, 2020, 34), challenging implicit or taken-for-granted hierarchies between jobs (Braverman, 1998).

Our approach to skills has provided more insights besides its descriptive value. We showed that the skill profiles of jobs are an important predictor of the wages offered in jobs, contradicting those who hypothesized that job-level work conditions have little explanatory power (Weeden and Grusky, 2005, 192). Indeed, controlling for the skill profiles of jobs and their 1-digit occupational codes predicts wages at least as well as their 4-digit codes alone. Furthermore, we demonstrated that the skill profiles of jobs add *new* rather than *redundant* information to occupational categories. This is further evidence that there is limited overlap between skills and occupational categories (Jencks et al., 1988; Sakamoto

and Wang, 2020). Whereas others have often limited themselves to this observation (e.g., Cassidy, 2017; Freeman et al., 2020), we showed that occupations remain an important predictor of wages even when differences in skill content across jobs are comprehensively captured. These novel findings complicate the current understanding of *why* occupations matter in the labor market, and outline the contours of a new research agenda in which both the distinct roles of skills and occupations are fully recognized and conceptualized. Occupations and skills cannot be simply used as interchangeable concepts.

6.2 Implications and agenda for future research

Measuring skill content beyond occupations

Our findings question how occupational classifications are measured and how researchers use them (Sakamoto and Wang, 2020; Martin-Caughey, 2021). While occupations are explicitly defined to classify the type of work that people do, in practice this classification relies on a highly truncated representation of the skill content of jobs. For example, we noted the coexistence of jobs with different levels of seniority within the same occupational category, leading to significant differences in skill content. This is consistent with the recent study by Cassidy (2017, 406), who observed that the "hierarchical level has a major impact on [the] reported task usages [of workers]". Occupational classifications tend to ignore the organizational setting in which skills are used (Tomaskovic-Devey and Avent-Holt, 2019; Jencks et al., 1988, 1328), neglecting that skills – and their returns – are highly context-dependent (Wilmers, 2020; Stephany and Teutloff, 2024).

Furthermore, occupations are defined by both the type of work performed by workers and their skill level, but the latter cannot be proxied by skill alone. As Hodge (1981, 397) rightly puts it, "Knowing the type of work [people] do and nothing more does not suffice to rank them.". Statistical organizations typically proxy skill levels with the amount of formal and informal education required to perform the jobs (Office for National Statistics, 2010; International Labour Office, 2012). We show that this has repercussions for their

skill composition: it tends to draw occupational boundaries between jobs that are actually similar in terms of the skills required, such as the elementary cleaning occupations and the housekeeping occupations, which are included in different major groups. This has important implications for scholars who use occupations as a proxy for skills: occupations capture not only horizontal differences between jobs but also some underlying educational hierarchy.

Thus, occupations are far from being neutral analytical categories, despite their widespread use. Occupational classifications remain prominent in the study of income inequalities and inter- and intragenerational mobility (Kalleberg and Mouw, 2018; Leicht, 2020; Barone et al., 2022) and they are also widely used in economics, particularly to study the effects of technological change (Freeman et al., 2020). Our results show that occupations cannot simply be used as proxies for skills, as is often the case in economic research (Cassidy, 2017). Even when controlling for 4-digit occupations, adding the skill profiles of jobs improves the predictive power of our wage model by 16%. While this percentage might seem small, an influential literature that relies on cruder measures of task content, such as the abstract/manual/routine triptych, tends to find much smaller effects of task content net of occupational categories (Autor and Handel, 2013; Cassidy, 2017). This suggests that important information on variation in skill requirements is lost. Our bottom-up approach allows future research to pay greater attention to what workers precisely do in their jobs, looking at the full spectrum of skills and their specific combination. This is critical for advancing our understanding of wage inequality among workers, as workers' productivity and comparative advantage depend on the complementarity of their skills (Anderson, 2017; Stephany and Teutloff, 2024) and the exclusivity of their skill profile (Wilmers, 2020).

Our findings are also important for research on the effects of digitalization and robotization. This literature would benefit from a more holistic understanding of skills, that recognizes their interdependence at the job level. Our approach shows that there is considerable heterogeneity in the nature of the skills requirements of jobs: specialized knowledge is often coupled with general skills and ICT tools of varying complexity. As shown by Den-

gler and Matthes (2018), overlooking this heterogeneity leads to exaggerating the adverse effects of automation on occupations: while some tasks are easily automated, most jobs include tasks that cannot be performed by machines. Future research could leverage our methodological and conceptual skill profile approach to further explore the hybrid nature of the skill composition of jobs, for example, by more finely distinguishing between different types of interpersonal and intrapersonal soft skills. Moreover, while a temporal study was beyond the scope of this study, an analysis of the evolution of skill profiles over time would help identify which skill combinations are increasingly valued in the labor market and how workers can adapt their skill profiles in times of changing labor markets. This would help understand how recent labor market changes are disrupting the skill structure of jobs, beyond the question of which skills are gaining or losing traction.

New perspectives for the study of labor market mobility

Our study opens new perspectives for the literature on labor market mobility. Our bottom-up approach allows us to visualize what a purely skill-based labor market would look like – without boundaries imposed by educational requirements and other institutional barriers. An important mechanism in understanding (a lack of) mobility between jobs is found in the (in)transferability of workers' skills (Poletaev and Robinson, 2008; Gathmann and Schönberg, 2010; Hanushek et al., 2017, e.g.,). Our mapping of skills on occupations suggests that there is in fact substantial overlap in the skill content of occupations. We found that many occupations require similar skills, even though they are far apart in occupational hierarchies. In contrast to conventional ideas, our findings indicate that upward mobility for workers in low-wage occupations would require only limited reskilling (See also DeMaria et al. (2020)). While this might sound striking, this finding is consistent with the permeability between occupations that workers themselves perceive (Hénaut et al., 2023). Especially at the bottom of the occupational structure, workers are found to "stretch status", that is, to identify with occupations of different status that involve similar tasks (Hénaut et al., 2023).

This suggests that there is much more potential for career mobility than we usually think. However, these skill-based transitions may be hindered in practice (DeMaria et al., 2020). Workers who have the right skill profile for a job may not have the certified credentials to access it (Lancee and Bol, 2017). This draws attention to the potential barriers and institutional mechanisms that limit the occupational mobility of workers regardless of, or beyond, the matching of their skill profiles (Weeden and Grusky, 2005; Bol et al., 2019).

More research is needed to assess the relative importance of these mechanisms and understand how formal qualifications may enhance or undermine the transferability of skills (Lancee and Bol, 2017; Nedelkoska and Neffke, 2019). And although there have been recent improvements in the development of multidimensional indicators of skill mismatches, existing measures typically focus on general skills to the detriment of job-specific knowledge (e.g., Guvenen et al., 2020; Lise and Postel-Vinay, 2020). Our skills-based approach makes it possible to identify potential labor market transitions at a more fine-grained level, and to comprehensively analyze the horizontal transferability of skills. We believe this to be an important and promising research agenda, and future research could for example compare our occupational mapping based on skill proximity with the mobility flows of workers from one occupation to another (Cheng and Park, 2020).

For a sociological approach to skills and occupations

Our finding that skills and occupations play distinct roles invites scholars to reconsider their respective contributions to labor market inequality. First, we confirm the relevance of the skills used at work, also known as "task-based human capital" (Gathmann and Schönberg, 2010; Schulz et al., 2023). Previous research has often linked differences in workers' career paths to the different skills they possess (see Hanushek et al., 2017), and primarily to the ones that they acquire through formal training (DiPrete et al., 2017). Our research suggests that another part of the story takes place in the workplace, where workers' skills take shape in a given hierarchical labor process (Avent-Holt et al., 2020). While a few studies have sought to conceptualize the allocation of work tasks into jobs

and their labor market returns (Wilmers, 2020), a sociological model that disentangles the effects of school-based and task-based human capital remains to be developed (Haupt and Ebner, 2020; Schulz et al., 2023).

One of the reasons for an absence of such a model can be found in data limitations: most existing data only contain a fixed set of often general skills (creative skills, cognitive skills, physical skills). An approach where different types of skills are distinguished requires detailed data on the skill profiles of both workers and jobs. This information is increasingly available through large N online labor market data, but these data sources typically represent very specific segments of the labor market (Anderson, 2017; Stephany and Teutloff, 2024).

The present study shows that online job postings are an underappreciated source of data to comprehensively map the skill structure of the labor market and enrich the more traditional employee surveys. Indeed, our comprehensive measure of the skill content of jobs could be combined with employee data at the job title level in order to further analyze the articulation between the skills that workers possess and the skills they use on the job.

While our findings provide avenues for deepening our understanding of the role skills play in labor market inequality, our study also gives further evidence that occupations matter even when the skill composition of jobs is held constant and measured comprehensively. Occupations are important for inequality not simply because of the tasks and skills workers perform in them. This echoes the question of Martin-Caughey (2021): "What's in an occupation?"

Our findings lend credence to theories that see occupations as much more than bundles of skills. Indeed, occupations are entities embedded in institutions and societies' cultures (Haupt and Ebner, 2020, 24-47). As Weeden and Grusky (2005, 153) argue, homogeneity may arise not per se because workers share the same working conditions, but because "employers (and, to some extent, workers) fashion jobs that correspond with ideal-typical occupational templates". According to this perspective, occupational boundaries are produced by institutional mechanisms and cultural perceptions that make some jobs more

worthy and valuable than others. While our study does not provide direct evidence of these mechanisms, it is consistent with the idea that skills alone cannot explain why some jobs pay more than others. It also highlights a new data source for studying the barriers that occupations create in the labor market, as job postings contain much more information than just skill requirements (e.g., educational requirements, licenses, etc.).

In sum, rather than opposing skill- and occupation-based perspectives, it would be interesting to explore how they are intertwined, and in particular how real or perceived differences in skills can be used to legitimize socio-economic inequalities between occupational groups (Braverman, 1998).

Investigating the role of institutional factors

It is important to note that our study focuses on the UK labor market, which calls for further research in other national contexts. First, we assessed the skill content of occupations as measured by the UK Standard Occupational Classification (SOC). Although this classification follows the same logic as the widely used International Standard Classification of Occupations (ISCO), there are subtle differences, which might affect findings with different occupational classifications.

Second, the skill formation system in the UK is characterized by a relatively open market economy, combined with an educational system that offers very little vocational and occupation-specific training (Hall and Soskice, 2001a; Bol and Van De Werfhorst, 2013). The generalist educational system of the UK is not strongly tied to the labor market, and in contrast to many continental European countries (Germany, the Netherlands), many jobs do not have specific educational requirements (Bol and Weeden, 2015).

Compared to other countries, the weak linkage between specific educational credentials and specific occupations may exacerbate the within-occupation variation in the skill requirements. While it is likely that our results will be different in countries with stronger linkage between education and the labor market, we argue that our findings should be relatively similar in the US, which – similar to the UK – lacks a strong vocational edu-

cational system and is characterized by diffuse pathways to employment (DiPrete et al., 2017). An interesting avenue for future research is to explicitly study these contextual factors: to what extent does the content of occupations depend on the institutional setup of a country's skill formation system?

In conclusion, this study opens up promising avenues for further research. It has shed a different light on occupational classifications, exploring new and "more complex ways in which two occupations can be 'close to' or 'distant from' one another" (Cheng and Park, 2020, 620). We have shown that when it comes to skills, occupational categories tend to harbor a great deal of heterogeneity. Should we then consider occupational analysis as an old VW Beetle from the 1960s and trade it in for a new model (Leicht, 2020)? Before we make a hasty decision, let's first open the hood.

Notes

1. Lightcast is a merger of Emsi and Burning Glass Technologies. For more information, visit https://lightcast.io/about/data [accessed on 21/09/2023].

2. This taxonomy is updated every two weeks and is freely available via an Application Programming Interface (API): https://lightcast.io/open-skills/access [accessed on 21/09/2023].

3. Lightcast uses unique but standardized skill requirements in order to group synonyms. For example, "communication skills" and "ability to communicate" are coded in the same skill requirement.

4. The most represented job titles among the job postings for which Lightcast did not identify any skill requirements are, in descending order of importance: Laborer, Care Assistant, Support Worker, Warehouse Operative, and Cleaner. A visual inspection of the raw text of these job postings suggests that job descriptions are typically very short or unspecific. When they do include some detail, job postings either emphasize what workers are expected to do (e.g., "sort through bins full of metal cables"), or personality traits they should have (e.g., "you'll need to be caring, compassionate and empathetic"), leaving aside identifiable skill requirements.

5. The skill requirements excluded from the random sample tend to be so specific or technical that they only appear in a minority of job postings. 75% of them appeared in fewer than 8 job postings and none appeared in more than a hundred job postings in 2019.

6. Using employment figures to correct the non-representativeness of job posting data has its limitations, as the two data sources measure different quantities: while job postings data measure the *flows* of vacancies, employment data inform about their *stock* at a given time. This is particularly problematic when job posting data are used as a proxy for the demand side. However, the purpose of our study is not to identify aggregate skill dynamics. Moreover, as Cammeraat and Squicciarini (2021, 17) show, using employment data is the most appropriate option in the absence of commensurable data between job postings and openings.

7. We used ONS employment figures from the Annual population survey (APS), which are available via Nomis on https://www.nomisweb.co.uk/datasets/aps168. Data were selected for the United Kingdom for the period January 2019 to December 2019.

8. This means that the biterm approach models the generation of word co-occurrence patterns in the entire corpus rather than in individual documents. This results in a global topic distribution of the corpus.

However, Yan et al. (2013) show that the topic distribution of each individual document can be derived and estimated in a satisfactory way.

9. We did not utilize external databases such as the O*NET because our goal is precisely to avoid a top-down approach to skill classification.

10. We presented the construction of the empirical distributions and the calculation of the distance between them as two separate steps but in practice, the MMD distance performs both steps simultaneously.

11. For a technical discussion of the MMD distance and parameter, see section 2.1.2 in the Supplementary Materials.

12. We also analyzed the silhouette width of occupational categories to assess whether they are consistent clusters in terms of their skill composition. Yet, it is important to note that the silhouette width is biased toward convex or spherical clusters (Rousseeuw, 1987, 55), making it a conservative measure for assessing the cluster quality of occupations. In other words, occupations may have a poor silhouette width not because they are poor skill clusters, but because they are non-spherical skill clusters.

13. The analysis using the logged maximum hourly wage yields very similar results and is available in the Supplementary Materials, Figure S20.

14. An interactive visualization of the LDA biterm topic model is available on https://surfdrive.surf.nl/files/index.php/s/T0uNAWrbuJkkIyl. The skill requirements listed in Table 4 can be retrieved with $\lambda = 0.7$. See Sievert and Shirley (2014) for an introduction to the interface.

15. We gave ChatGPT the following instructions: "I would like you to find good names for the following latent topics, based on the top-100 word probabilities. The names should be precise but not too long (no more than 5 words)".

References

Acemoglu, D. and Autor, D. (2010). Skills, Tasks and Technologies: Implications for Employment and Earnings. Technical Report w16082, National Bureau of Economic Research, Cambridge, MA.

Alabdulkareem, A., Frank, M. R., Sun, L., AlShebli, B., Hidalgo, C., and Rahwan, I. (2018). Unpacking the polarization of workplace skills. *Science Advances*, 4(7):eaao6030.

Anderson, K. A. (2017). Skill networks and measures of complex human capital. *Proceedings of the National Academy of Sciences*, 114(48):12720–12724.

Autor, D. H. and Handel, M. J. (2013). Putting Tasks to the Test: Human Capital, Job Tasks, and Wages. *Journal of Labor Economics*, 31(S1):S59–S96.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333.

Avent-Holt, D., Henriksen, L. F., Hägglund, A. E., Jung, J., Kodama, N., Melzer, S. M., Mun, E., Rainey, A., and Tomaskovic-Devey, D. (2020). Occupations, workplaces or jobs?: An exploration of stratification contexts using administrative data. *Research in Social Stratification and Mobility*, 70:100456.

Backhaus, K. B. (2004). An Exploration of Corporate Recruitment Descriptions on Monster.com. *The Journal of Business Communication (1973)*, 41(2):115–136. Publisher: SAGE Publications.

Barone, C., Hertel, F. R., and Smallenbroek, O. (2022). The rise of income and the demise of class and social status? A systematic review of measures of socio-economic position in stratification research. *Research in Social Stratification and Mobility*, 78:100678.

Blair, P. Q. and Deming, D. J. (2020). Structural Increases in Demand for Skill after the Great Recession. *AEA Papers and Proceedings*, 110:362–365.

Blau, P. and Duncan, O. D. (1967). *The American Occupational Structure*. John Wiley & Sons, New-York.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Bol, T., Ciocca Eller, C., Van De Werfhorst, H. G., and DiPrete, T. A. (2019). School-to-Work Linkages, Educational Mismatches, and Labor Market Outcomes. *American Sociological Review*, 84(2):275–307.

Bol, T. and Van De Werfhorst, H. G. (2013). Educational Systems and the Trade-Off between Labor Market Allocation and Equality of Educational Opportunity. *Comparative Education Review*, 57(2):285–308.

Bol, T. and Weeden, K. A. (2015). Occupational Closure and Wage Inequality in Germany and the United Kingdom. *European Sociological Review*, 31(3):354–369.

Braverman, H. (1998). *Labor and monopoly capital: the degradation of work in the twentieth century*. Monthly Review Press, New York, 25th anniversary ed edition.

Buchmann, M., Buchs, H., and Gnehm, A.-S. (2020). Occupational Inequality in Wage Returns to Employer Demand for Types of Information and Communications Technology (ICT) Skills: 1991–2017. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 72(S1):455–482.

Burning Glass Technologies (2015). The Human Factor. The Hard Time Employers Have Finding Soft Skills. Technical report, Burning Glass Technologies, Boston.

Böhm, M. J., Von Gaudecker, H.-M., and Schran, F. (2024). Occupation Growth, Skill Prices, and Wage Inequality. *Journal of Labor Economics*, 42(1):201–243.

Caines, C., Hoffmann, F., and Kambourov, G. (2017). Complex-task biased technological change and the labor market. *Review of Economic Dynamics*, 25:298–319.

Cammeraat, E. and Squicciarini, M. (2021). Burning Glass Technologies' data use in policy-relevant analysis: An occupation-level assessment. OECD Science, Technology and Industry Working Papers 2021/05, Organisation for Economic Co-operation and Development. Series: OECD Science, Technology and Industry Working Papers Volume: 2021/05.

Caplow, T. (1954). *The Sociology of Work*. University of Minnesota Press., Minneapolis.

Carnevale, A. P., Jayasundera, T., and Repnikov, D. (2014). Understanding online job ads data. A technical report. Technical report, Center on Education and the Workforce, Georgetown University.

Cassidy, H. (2017). Task Variation Within Occupations. *Industrial Relations: A Journal of Economy and Society*, 56(3):393–410.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: how humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 288–296, Vancouver, British Columbia, Canada. Curran Associates Inc.

Charles, M. and Grusky, D. B. (2004). *Occupational ghettos: the worldwide segregation of women and men*. Studies in social inequality. Stanford Univ. Press, Stanford, Calif.

Cheng, S. and Park, B. (2020). Flows and Boundaries: A Network Approach to Studying Occupational Mobility in the Labor Market. *American Journal of Sociology*, 126(3):577–631.

Cortes, G. M. (2016). Where Have the Middle-Wage Workers Gone? A Study of Polarization Using Panel Data. *Journal of Labor Economics*, 34(1):63–105.

Damelang, A., Haupt, A., and Abraham, M. (2018). Economic consequences of occupational deregulation: Natural experiment in the German crafts. *Acta Sociologica*, 61(1):34–49.

DeMaria, K., Fee, K., and Wardrip, K. (2020). Exploring A Skills-Based Approach To Occupational Mobility. Community Affairs Discussion Paper 89004, Federal Reserve Bank of Philadelphia.

Deming, D. and Kahn, L. B. (2018). Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. *Journal of Labor Economics*, 36(S1):S337–S369.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market*. *The Quarterly Journal of Economics*, 132(4):1593–1640.

Dengler, K. and Matthes, B. (2018). The impacts of digital transformation on the labour market: Substitution potentials of occupations in Germany. *Technological Forecasting and Social Change*, 137:304–316.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6):570–606.

DiPrete, T. A., Bol, T., Eller, C. C., and Van De Werfhorst, H. G. (2017). School-to-Work Linkages in the United States, Germany, and France. *American Journal of Sociology*, 122(6):1869–1938.

Djumalieva, J. and Sleeman, C. (2018). An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts. In Larsen, C., Rand, S., Schmid, A., and Dean, A., editors, *Developing Skills in a Changing World of Work*, pages 425–454. Rainer Hampp Verlag.

Durkheim, E. (1893). *The division of labor in society*. Free Press, New-York, 1933 edition.

Erikson, R. and Goldthorpe, J. H. (1992). *The constant flux: a study of class mobility in industrial societies*. Clarendon Press ; Oxford University Press, Oxford [England] : New York.

Fernández-Macías, E. and Bisello, M. (2020). A Taxonomy of Tasks for Assessing the Impact of New Technologies on Work. Technical Report 2020/04, European Commission, Joint Research Centre (JRC), Seville.

Fernández-Macías, E. and Hurley, J. (2016). Routine-biased technical change and job polarization in Europe. *Socio-Economic Review*, page mww016.

Freeman, R. B., Ganguli, I., and Handel, M. J. (2020). Within-Occupation Changes Dominate Changes in What Workers Do: A Shift-Share Decomposition, 2005–2015. *AEA Papers and Proceedings*, 110:394–399.

Ganzeboom, H. B., De Graaf, P. M., and Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21(1):1–56.

Ganzeboom, H. B. and Treiman, D. J. (1996). Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research*, 25:201–239.

Gathmann, C. and Schönberg, U. (2010). How General Is Human Capital? A Task-Based Approach. *Journal of Labor Economics*, 28(1):1–49.

Geel, R. and Backes-Gellner, U. (2011). Occupational mobility within and between skill clusters: an empirical analysis based on the skill-weights approach. *Empirical Research in Vocational Education and Training*, 3(1):21–38.

Goos, M., Manning, A., and Salomons, A. (2009). Job Polarization in Europe. *American Economic Review*, 99(2):58–63.

Goos, M., Manning, A., and Salomons, A. (2014). Explaining Job Polarization: Routine-Biased Technological Change and Offshoring. *American Economic Review*, 104(8):2509–2526.

Greenacre, M. (2021). Compositional Data Analysis. *Annual Review of Statistics and Its Application*, 8(1):271–299.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773. ISBN: 1532-4435 Publisher: JMLR. org.

Guvenen, F., Kuruscu, B., Tanaka, S., and Wiczer, D. (2020). Multidimensional Skill Mismatch. *American Economic Journal: Macroeconomics*, 12(1):210–244.

Hall, P. A. and Soskice, D. W. (2001a). An Introduction to Varieties of Capitalism. In Hall, P. A. and Soskice, D., editors, *Varieties of Capitalism*, pages 1–68. Oxford University PressOxford, 1 edition.

Hall, P. A. and Soskice, D. W. (2001b). *Varieties of capitalism: the institutional foundations of comparative advantage*. Oxford University Press, Oxford.

Hamann, J. and Beljean, S. (2021). Career gatekeeping in cultural fields. *American Journal of Cultural Sociology*, 9(1):43–69.

Hanushek, E. A., Schwerdt, G., Woessmann, L., and Zhang, L. (2017). General Education, Vocational Education, and Labor-Market Outcomes over the Lifecycle. *Journal of Human Resources*, 52(1):48–87.

Haslberger, M. (2021). Routine-biased technological change does not always lead to polarisation: Evidence from 10 OECD countries, 1995–2013. *Research in Social Stratification and Mobility*, 74:100623.

Haslberger, M. (2022). Rethinking the measurement of occupational task content. *The Economic and Labour Relations Review*, 33(1):178–199.

Haupt, A. (2023). Who Profits from Occupational Licensing? *American Sociological Review*, 88(6):1104–1130.

Haupt, A. and Ebner, C. (2020). Occupations and Inequality: Theoretical Perspectives and Mechanisms. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 72(S1):19–40.

Hauser, R. M. and Warren, J. R. (1997). 4. Socioeconomic Indexes for Occupations: A Review, Update, and Critique. *Sociological Methodology*, 27(1):177–298.

Hershbein, B. and Kahn, L. B. (2018). Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings. *American Economic Review*, 108(7):1737–1772.

Hodge, R. W. (1981). The measurement of occupational status. *Social Science Research*, 10(4):396–415.

Hurrell, S. A., Scholarios, D., and Thompson, P. (2013). More than a 'humpty dumpty' term: Strengthening the conceptualization of soft skills. *Economic and Industrial Democracy*, 34(1):161–182.

Hénaut, L., Lena, J. C., and Accominotti, F. (2023). Polyoccupationalism: Expertise Stretch and Status Stretch in the Postindustrial Era. *American Sociological Review*, 88(5):872–900.

International Labour Office (2012). *International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables*. International Labour Office.

Janietz, C. and Bol, T. (2020). Occupations, organizations, and the structure of wage inequality in the Netherlands. *Research in Social Stratification and Mobility*, 70:100468.

Jencks, C., Perman, L., and Rainwater, L. (1988). What Is a Good Job? A New Measure of Labor-Market Success. *American Journal of Sociology*, 93(6):1322–1357. Publisher: University of Chicago Press.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage Inequality and the Rise in Returns to Skill. *Journal of Political Economy*, 101(3):410–442.

Kalleberg, A. L. (2009). Precarious Work, Insecure Workers: Employment Relations in Transition. *American Sociological Review*, 74(1):1–22.

Kalleberg, A. L. and Mouw, T. (2018). Occupations, Organizations, and Intragenerational Career Mobility. *Annual Review of Sociology*, 44(1):283–303.

Kim, C., Kim, J., and Ban, M. (2020). Do you know what you do for a living? Occupational coding mismatches between coders in the Korean General Social Survey. *Research in Social Stratification and Mobility*, 70:100467.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949.

Lancee, B. and Bol, T. (2017). The Transferability of Skills and Degrees: Why the Place of Education Affects Immigrant Earnings. *Social Forces*, 96(2):691–716.

Lazear, E. (2009). Firm-Specific Human Capital: A Skill-Weights Approach. *Journal of Political Economy*, 117(5):914–940.

le Grand, C. and Tåhlin, M. (2013). Class, Occupation, Wages, and Skills: The Iron Law of Labor Market Inequality. In Elisabeth Birkelund, G., editor, *Comparative Social Research*, volume 30, pages 3–46. Emerald Group Publishing Limited.

Le Vrang, M., Papantoniou, A., Pauwels, E., Fannes, P., Vandensteen, D., and De Smedt, J. (2014). ESCO: Boosting Job Matching in Europe with Semantic Interoperability. *Computer*, 47(10):57–64.

Leicht, K. T. (2020). Occupations and Inequalities in the 21st Century: What's in your Wallet? *Research in Social Stratification and Mobility*, 70:100550.

Lise, J. and Postel-Vinay, F. (2020). Multidimensional Skills, Sorting, and Human Capital Accumulation. *American Economic Review*, 110(8):2328–2376.

Liu, Y. and Grusky, D. B. (2013). The Payoff to Skill in the Third Industrial Revolution. *American Journal of Sociology*, 118(5):1330–1374.

Lundberg, I., Johnson, R., and Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, 86(3):532–565.

Macanovic, A. (2022). Text mining for social science – The state and the future of computational text analysis in sociology. *Social Science Research*, 108:102784.

Martin-Caughey, A. (2021). What's in an Occupation? Investigating Within-Occupation Variation and Gender Segregation Using Job Titles and Task Descriptions. *American Sociological Review*, 86(5):960–999.

Modestino, A. S., Shoag, D., and Ballance, J. (2020). Upskilling: Do Employers Demand Greater Skill When Workers Are Plentiful? *The Review of Economics and Statistics*, 102(4):793–805.

Mouw, T. and Kalleberg, A. L. (2010). Occupations and the Structure of Wage Inequality in the United States, 1980s to 2000s. *American Sociological Review*, 75(3):402–431.

Mouw, T., Kalleberg, A. L., and Schultz, M. A. (2024). "Stepping-Stone" versus "Dead-End" Jobs: Occupational Structure, Work Experience, and Mobility Out of Low-Wage Jobs. *American Sociological Review*, 89(2):298–345.

Nedelkoska, L. and Neffke, F. (2019). Skill Mismatch and Skill Transferability: Review of Concepts and Measurements. *Papers in Evolutionary Economic Geography (PEEG)*, 1921.

Nelson, L. K. (2021). Cycles of Conflict, a Century of Continuity: The Impact of Persistent Place-Based Political Logics on Social Movement Strategy. *American Journal of Sociology*, 127(1):1–59.

Oesch, D. and Piccitto, G. (2019). The Polarization Myth: Occupational Upgrading in Germany, Spain, Sweden, and the UK, 1992–2015. *Work and Occupations*, 46(4):441–469.

Office for National Statistics (2010). *Standard occupational classification 2010*. Palgrave Macmillan, Basingstoke, Hampshire. OCLC: 691102853.

Parkin, F. (1971). *Class inequality and political order: social stratification in capitalist and Communist societies*. MacGibbon & Kee, London.

Poletaev, M. and Robinson, C. (2008). Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984–2000. *Journal of Labor Economics*, 26(3):387–420.

Rohrbach-Schmidt, D. (2019). Putting Tasks to the Test: The Case of Germany. *Social Inclusion*, 7(3):122–135.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai China. ACM.

Sacchi, S., Kriesi, I., and Buchmann, M. (2016). Occupational mobility chains and the role of job opportunities for upward, lateral and downward mobility in Switzerland. *Research in Social Stratification and Mobility*, 44:10–21.

Sakamoto, A. and Wang, S. X. (2020). The declining significance of occupation in research on intergenerational mobility. *Research in Social Stratification and Mobility*, 70:100521.

Salvisberg, A. and Sacchi, S. (2014). Labour Market Prospects of Swiss Career Entrants after Completion of Vocational Education and Training. *European Societies*, 16(2):255–274.

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

Schulz, W., Solga, H., and Pollak, R. (2023). Vocational education, tertiary education, and skill use across career stages. *European Sociological Review*, page jcac074.

Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sostero, M. and Fernández-Macías, E. (2021). The Professional Lens: What Online Job Advertisements Can Say about Occupational Task Profiles. Technical Report JRC125917, European Commission, Seville.

Spenner, K. I. (1990). Skill: Meanings, Methods, and Measures. *Work and Occupations*, 17(4):399–421.

Stephany, F. and Teutloff, O. (2024). What is the price of a skill? The value of complementarity. *Research Policy*, 53(1):104898.

Tippins, N. T. and Hilton, M. L. (2010). *A database for a changing economy: review of the Occupational Information Network (O*NET)*. National Academies Press, Washington, D.C. OCLC: 650842713.

Tomaskovic-Devey, D. and Avent-Holt, D. (2019). *Relational Inequalities: An Organizational Approach*. Oxford University Press, 1 edition.

Treiman, D. J. (1977). *Occupational Prestige in Comparative Perspective*. Elsevier.

Weber, M. (1979). *Economy and society: an outline of interpretive sociology*. University of California Press, Berkeley London.

Weeden, K. (2002). Why Do Some Occupations Pay More than Others? Social Closure and Earnings Inequality in the United States. *American Journal of Sociology*, 108(1):55–101.

Weeden, K. and Grusky, D. (2005). The Case for a New Class Map. *American Journal of Sociology*, 111(1):141–212.

Weeden, K. A., Kim, Y.-M., Di Carlo, M., and Grusky, D. B. (2007). Social Class and Earnings Inequality. *American Behavioral Scientist*, 50(5):702–736.

Williams, M. (2013). Occupations and British Wage Inequality, 1970s-2000s. *European Sociological Review*, 29(4):841–857.

Wilmers, N. (2020). Job Turf or Variety: Task Structure as a Source of Organizational Inequality. *Administrative Science Quarterly*, 65(4):1018–1057.

Yamaguchi, S. (2012). Tasks and Heterogeneous Human Capital. *Journal of Labor Economics*, 30(1):1–53.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 1445–1456, Rio de Janeiro, Brazil. ACM Press.