DISCUSSION PAPER SERIES

DP20618
(v. 2)

WAGE MARKDOWNS AND THE SORTING OF WORKERS TO FIRMS WHEN SKILLS ARE BUNDLED

Philippe Choné, Francis Kramarz and Oskar Nordström Skans

LABOUR ECONOMICS

ISSN 0265-8003

WAGE MARKDOWNS AND THE SORTING OF WORKERS TO FIRMS WHEN SKILLS ARE BUNDLED

Philippe Choné, Francis Kramarz and Oskar Nordström Skans

Discussion Paper DP20618
First Published 04 September 2025
This Revision 28 September 2025

Centre for Economic Policy Research
187 boulevard Saint-Germain, 75007 Paris, France
2 Coldbath Square, London EC1R 5HL
Tel: +44 (0)20 7183 8801
www.cepr.org

WAGE MARKDOWNS AND THE SORTING OF WORKERS TO FIRMS WHEN SKILLS ARE BUNDLED

Abstract

We study a competitive labor market where heterogeneous workers supply multidimensional skills to heterogeneous firms. Firms produce output by aggregating employees' skill bundles, using them as inputs in a concave production function. A single friction—the bundling of workers' skills—generates an equilibrium with workers-to-firms sorting based on comparative advantage, where each firm has a unique optimal size, and where the wage schedule is convex. Skill prices vary across firms despite competitive wage-setting, with a markdown for workers endowed with balanced skills. We illustrate the empirical relevance of key assumptions and predictions using matched worker-firm data on workers' skill profiles.

JEL Classification: D20, D40, D51, J20, J24, J30

Keywords: N/A

Philippe Choné - chone@ensae.fr
CREST, ENSAE, Institut Polytechnique de Paris and CEPR

Francis Kramarz - kramarz.francis@gmail.com
Innovation Lab, College de France and Uppsala University, Sweden and CEPR

Oskar Nordström Skans - Oskar.Nordstrom_Skans@nek.uu.se
Uppsala University, Sweden

Acknowledgements
We are grateful to Ariel Burstein, Marco Cuturi, Jan Eeckhout, Xavier d'Haultfoeuille, Alfred Galichon, Nathael Gozlan, Basile Grassi, Gene Grossman, Philipp Kircher, Samuel Kortum, Thomas Le Barbanchon, Ilse Lindenlaub, Simon Mongey, Giuseppe Moscarini, Marco Ottaviani, Francois-Pierre Paty, Torsten Persson, Jean-Marc Robin, Jean-Charles Rochet, Bernard Salanie, David Schoenholzer, Bob Topel for extremely helpful conversations, remarks, and suggestions as well those of participants at multiple seminars and conferences. We thank Gaspard Chone-Ducasse and Pierre Krakovitch for excellent research assistance. Chone gratefully acknowledges support from the French National Research Agency (ANR), (Labex ECODEC No. ANR-11-LABEX-0047) and Kramarz from the European Research Council (ERC), advanced grant FIRMNET (project ID 741467).

Wage Markdowns and the Sorting of Workers to Firms when Skills are Bundled *

Philippe Choné[†], Francis Kramarz[‡], and Oskar Nordström Skans[◇]

[†] *CREST, ENSAE, Institut Polytechnique de Paris, France*

[‡] *Innovation Lab, Collège de France and Uppsala University, Sweden*

[◇] *Uppsala University, Sweden*

September 28, 2025

Abstract

We study a competitive labor market where heterogeneous workers supply multidimensional skills to heterogeneous firms. Firms produce output by aggregating employees' skill bundles, using them as inputs in a concave production function. A single friction—the bundling of workers' skills—generates an equilibrium with workers-to-firms sorting based on comparative advantage, where each firm has a unique optimal size and where the wage schedule is convex. Skill prices vary across firms despite competitive wage-setting, with a markdown for workers endowed with balanced skills. We illustrate the empirical relevance of key assumptions and predictions using matched worker-firm data on workers' skill profiles.

JEL Codes: D20, D40, D51, J20, J24, J30
Keywords: bundling; multidimensional skills; workers-to-firms matching; sorting; heterogeneous firms

*We are grateful to John Abowd, Ariel Burstein, David Card, Marco Cuturi, Jan Eeckhout, Xavier d'Haultfoeuille, Alfred Galichon, Nathael Gozlan, Basile Grassi, Gene Grossman, Philipp Kircher, Patrick Kline, Samuel Kortum, Thomas Le Barbanchon, Ilse Lindenlaub, Simon Mongey, Giuseppe Moscarini, Marco Ottaviani, François-Pierre Paty, Torsten Persson, Jean-Marc Robin, Jean-Charles Rochet, Bernard Salanié, David Schönholzer, Bob Topel for extremely helpful conversations, remarks, and suggestions as well those of participants at multiple seminars and conferences. We thank Gaspard Choné-Ducasse and Pierre Krakovitch for excellent research assistance. Choné gratefully acknowledges support from the French National Research Agency (ANR), (Labex ECODEC No. ANR-11-LABEX-0047), Kramarz and Skans from the European Research Council (ERC), advanced grant FIRMNET (project ID 741467) and Skans from Swedish Research Council 2018-04581.

1 Introduction

Wage dispersion having a large between-firm component, high-wage workers being employed by high-wage firms, and workers being sorted across firms according to their comparative advantage,[1] all suggest that the matching between workers and their employing firms is a crucial determinant of the aggregate wage distribution. Furthermore, an additional layer of firm-level forces related to firms' market power, allowing firms to mark down wages below the value of workers' marginal products, has been shown to affect the wage structure and several other related labor market outcomes.[2]

Because of these findings, it has become central to construct a framework at the intersection of Macroeconomics, Labor Economics, and Industrial Organization that can be used to explore how the technological and competitive environment affect wage structures and workers-to-firms sorting patterns. Indeed, general equilibrium models where a discrete number of firms exert power on labor markets (as well as on product markets) have recently been proposed and, in some cases, estimated.[3] However, the origin of the workers-to-firms sorting and its associated impact on the wage structure tend to be missing from most of these models, despite its central role in the empirical literature mentioned above (and also discussed in Kline, 2024).[4]

These recent studies have all adopted models that reject the assumption of perfect competition in order to explain the role of firms on the labor market. By contrast, we offer in this paper a complement to this view, showing that many patterns related to sorting and wage inequality can be explained in a world of perfect competition as long as a) heterogeneous workers have multi-dimensional skills they cannot unpack, preventing the emergence of markets for stand-alone skills and b) firms with heterogeneous input weights aggregate the multi-dimensional skills into productive inputs. Our set-up comprises a continuum of heterogeneous workers and a continuum of heterogeneous firms. On the supply side, each worker needs to sell her *entire bundle* of skills to one employing firm. On the demand side, firms produce output by *aggregating the skill-bundles of their entire workforce*, thereby giving substance to firms beyond being a mere collection of one-person jobs. The resulting general equilibrium yields the optimal wage schedule, a unique firm-size, and the worker-to-firm sorting.

[1] See e.g. Song, Price, Guvenen, Bloom, and von Wachter (2019), Fredriksson, Hensvik, and Skans (2018), and Lise and Postel-Vinay (2020) for relevant evidence.

[2] See De Loecker, Eeckhout, and Unger (2020), Yeh, Macaluso, and Hershbein (2022), and Autor, Dorn, Katz, Patterson, and Van Reenen (2020).

[3] See in particular Berger, Herkenhoff, and Mongey, 2022, Deb, Eeckhout, Patel, and Warren, 2024, Wong, 2023, Haanwinckel, 2025, Freund, 2023 ...) as well as the illuminating discussions in Van Reenen (2024) and Violante (2024).

[4] See the estimated models in Lise and Postel-Vinay (2020), Lamadon, Mogstad, and Setzler (2022), Freund (2023), or Haanwinckel (2025) however.

In our set-up—denoting i the set of workers sorting into firm j— workers' skills s are aggregated into K inputs, $\int_i s_{i,1}, \ldots, \int_i s_{i,K}$, which enter into firm j's concave production function, weighted by heterogeneous factors $\alpha_{j,k}$. Hence, firm's j production is $F(\int_i s_{i,1}, \ldots, \int_i s_{i,K}; \{\alpha_{j,k}\})$ rather than an aggregation of each job's production – $\int_i F(s_{i,1}, \ldots, s_{i,K}; \{\alpha_{j,k}\})$. The assumption that workers need firms to make their multidimensional skills productive—and the ensuing absence of markets for stand-alone skills—defines bundling and the firm's substance.

We show that the introduction of this single, little studied, friction – the bundling of workers' skills in a setup with firms rather than jobs – provides theoretical results that are directly relevant for understanding many empirical regularities on firms and labor markets. Although our contribution is inherently theoretical, we illustrate the relevance of our key assumptions and findings using Swedish register data on workers multidimensional skills linked to firm-level production data. These data clearly support a set of key motivating facts for our approach. Crucially, workers have heterogeneous bundles of skills and sort on comparative advantage across firms with heterogeneous skills' demand. These data are also shown to be consistent with a set of key theoretical predictions from our model.

The general equilibrium (GE) solution of our matching problem allows us to characterize, first, the **equilibrium sorting** of workers-to-firms, a matching that maximizes total production and, second, the **equilibrium wage schedule** that decentralizes the previous optimum. The workers-to-firms sorting will be shown to generate between and within-firms wage inequality at the equilibrium. In contrast to job-level models, we can also derive each firm's equilibrium size, defined as a a skill-weighted count of workers. The bundling friction creates a fundamental *asymmetry* between workers endowed with a balanced set of skills (Generalists) and workers with an unbalanced set of skills (Specialists). Indeed, firms can directly replace the skills of a Generalist using a diverse team of Specialists. But firms cannot replace a Specialist using a subset of a Generalist worker's skills, without leaving the Generalist's remaining skills unused. As a consequence, Specialists have the option to compete with Generalists while being rewarded for their entire skill set, but the reverse is not true. The equilibrium wage schedule is therefore convex, and features a wage "markdown" for Generalists as their skills have (weakly) lower prices than those of Specialists, even within an otherwise perfectly competitive framework.

Sorting: Firms, and their production functions, are central to our framework. We allow firms' technology profiles to be heterogeneous in terms of total factor productivity and the weight of each input in production. At equilibrium, firms hire their *unique* lowest cost mix of skills. Firms aggregate these skills into tasks, used as inputs in their concave production function. The firm-level labor demand is shown to be associated with a unique profile of skills, corresponding to a unique skill ratio in a two-skills world. Depending on aggregate supply and demand, this skill profile is obtained by either hiring multiple workers with exactly the right skill ratio, or by *bunching* together a combination of more specialized workers. When supply comprises mostly workers with diverse skill sets (Generalists), all workers in the firm will be endowed with their employing firm's optimal mix. The match between each worker's comparative advantage on one side and each firm's comparative advantage on the other drives sorting.

Firm-size: At equilibrium, each firm will have an optimal and unique size, which will be defined by a combination of a skill profile and the number of workers. The fact that size – defined as the firm-aggregated vector of skills – is *unique* at any competitive equilibrium is an important contribution to the literature, distinguishing our analysis from a large fraction of the macro-labor literature where jobs are center-stage. We also show that high TFP firms will employ a high-quality labor force. Hence, for a given level of employment, high-productivity firms employ high-quality individual workers.

The equilibrium wage schedule: The wage schedule which decentralizes the optimum of the economy faced with a bundling friction is shown to be a convex function of workers' skills. The implicit price of each skill depends on the worker's employing firm. Importantly, *the law of one price does not hold for skills*, even though wage setting is fully competitive since skills cannot be traded on stand-alone markets. Inequality in our model has a strong between-firms component because the convexity is "indexed" by firms' optimal skill profiles, defined as the skill ratio in the two-skill case.

To understand why the wage schedule is convex, note first that there is no unemployment in our competitive labor market. Hence, a generalist worker endowed with a balanced skill vector must be employed, and her wage cannot exceed that of an equivalent combination of specialists that provide the same set of skills. But specialists are also demanded by firms that primarily use their specialist skills. This asymmetry, where specialists can be combined to replace generalists, whereas the converse is not true, generates convexity. In equilibrium, a generalist will be paid less than or equal to the "equivalent" combination of pure specialists, creating a potential wage *markdown* where the price of generalist skills are lower than the price for specialists skills.[5] This prediction is supported by our multi-dimensional skills data; our estimates suggest that skill prices are 20 percent lower in firms that use a mixed set of skills.

Because the workers-to-firms equilibrium sorting is indexed by the firms' optimal skill profile, markdowns and sorting are directly connected: they are the two faces of the same equilibrium property. Firms demanding a mixed set of skills pay lower skill

[5] Admittedly a very different markdown mechanism than those emphasized in the recent literature, see Kline (2025).

prices than other firms, resulting in a polarized wage structure—skill prices are higher the further away we are from the mixed skill segment.

In our economy, workers with the same skill profile (skill ratio in the two-dimensional case) are allowed to have heterogeneous qualities, just as firms are heterogeneous in terms of total factor productivity. Worker quality reflects the intensity of skills, with a given profile (ratio). We show that the wage schedule is not only convex, but also a homogenous function of *degree one* in the quality. Hence, workers within the same firm can have heterogeneous qualities even when the equilibrium prescribes that they share the same profile (i.e. even without "bunching" of different types of specialists).

The respective supply of Generalist workers and Specialist workers facing firms' demand for each type of workers directly affects the nature of the hiring equilibrium which comes in two forms: Pure Sorting versus "Bunching". This leads us to perform comparative statics exercises as well as simulations by varying the relative prevalence of Generalists and Specialists to illustrate how the market equilibrium changes. We also check that our models' predictions – markdowns are larger when Generalists are plentiful – are supported by our skills data.

A worker's individual log-wage is the sum of a worker effect (the log of quality) and a "worker-to-firm sorting" effect which captures the firm-specific skill profile (and is also the negative of the log markdown at that firm). This log-wage decomposition into an individual quality component and a worker-to-firm sorting effect shares similarities with the AKM log-wage decomposition (Abowd, Kramarz, and Margolis, 1999) into a person and a firm effect. However, by contrast, this worker-to-firms sorting effects arises in a competitive world without firm-specific wage effects and without mobility whereas traditional firm effects, as defined by AKM and Kline (2024), require firm-to-firm mobility to be identified.

Importantly, when production functions are *non-homothetic*, a systematic relationship between firm productivity and skill profiles emerges. This relationship is supported by our data: more productive firms hire workers with more cognitive skill profiles, and pay higher skill prices for cognitive skills. This suggests that workers with more cognitive than non-cognitive skills have a comparative advantage in more productive firms, a pattern which provides a link between sorting on comparative advantage and positive assortative matching, phenomena that most previous studies have analyzed in isolation. In addition, with non-homothetic production functions, the worker-to-firm sorting effect depends on total factor productivity, and the firm's equilibrium labor share decreases with the firm's productivity.[6]

Finally, we highlight the differences of our firm-centered approach with classic workers-to-jobs matching. We show that the latter generically yields a strictly lower

[6]See Simonovska (2015) and Autor, Dorn, Katz, Patterson, and Van Reenen (2020).

optimal output in the economy than workers-to-firms matching. The two notions of matching yield the same output only in the unlikely situation where firms and jobs coincide at the optimum.

Related Literature: Our contribution lies at the intersection of Labor, Macro, and IO. Despite being motivated by a large body of empirical research and despite the suggestive evidence presented just above and along the way, this article has a very strong theoretical backbone. Hence, we start by discussing papers which examine theories connected to our approach, and mention their empirical contributions when present.

Bundling Multidimensional Skills: A limited number of articles has studied skills bundling, the first being Heckman and Scheinkman (1987). Their focus was on the law of one price per skill with firms (or rather sectors) producing from the aggregate of these skills bundles, showing that, indeed, the law did not hold in some circumstances. By contrast, Lindenlaub (2017) focuses on sorting and provides a full characterization of positive assortative matching, PAM, or its negative counterpart, NAM, in a multidimensional framework with jobs but no firms, hence with no aggregation of skills entering a firm-level production function (and, hence, no firm size). Edmond and Mongey (2022) also examine bundling using a model with two tasks (occupations), two skills, and heterogeneous workers. Their interest lies in the sorting of workers into the two occupations, within a macroeconomic perspective (see also Hernnäs, 2021). Even though they do not consider Bundling per se, Lise and Postel-Vinay (2020) consider multi-dimensional skills and how they matter for labor market analysis.

Comparative Advantage: Comparative advantage in sorting patterns play an important role in our contribution. We are not the first to connect bundling with comparative advantage. Indeed, this link has been studied in an International Trade context. Ohnsorge and Trefler (2007) show that international differences in the distribution of workers' skill bundles, such as Japan's abundance of workers with a modest mix of both quantitative and teamwork skills, have important implications for international trade, industrial structure, and domestic income distribution (also in a model with jobs and no firms). Costinot and Vogel (2010), have firms and production but no bundling. Their high-skill workers have a comparative advantage in tasks with high-skill intensity inducing sorting between skills and tasks.

Giving Firms Substance: Our research is inspired by a recent and important contribution, Eeckhout and Kircher (2018). Workers in their approach have one dimension of skills. However, to obtain firms that are more than a collection of jobs, Eeckhout and Kircher (2018) separate workers' quality from their quantity, allowing them to study rich sorting patterns. Garicano and Rossi-Hansberg (2006)' seminal contribution shares our focus on firm-level production. Firms are also central in Haanwinckel (2025), Freund (2023) and Boerma, Tsyvinski, and Zimin (2021). In the first two articles, sorting plays a central role. The latter two have firms of different productivity, but in stark contrast with our article, firm's size is exogenous (set at two).[7]

From Hedonic Models to Optimal Transport: Following Rosen (1974), hedonic models applied to product markets focus on the matching of consumers and goods (demand), with goods and firms (supply). Matching models of the labor market are somewhat similar: workers' skills and tasks lie on the supply side when and tasks and firms lie on the demand side. Our model, however, departs from an hedonic model of the labor market in two important dimensions. First, *tasks are not directly observed* by the researcher.[8] Hence, we model tasks as an unobserved function of skills. Second, the production process considered involves the aggregation of employees' skills within firms, whereas Rosen (1974) explicitly rules out such an aggregation. Chiappori, McCann, and Nesheim (2010) expresses hedonic models as optimal transport (OT) problems (see also Galichon, 2018). Our analysis contributes to the literature that studies many-to-one matching with transferable utility. Because skills are aggregated within firms, and because firms choose their size, we rely on a very recent OT method (weak optimal transport with unnormalized kernels, WOTUK) introduced by Choné, Gozlan, and Kramarz (2023), itself an extension of the so-called (weak optimal transport, WOT) introduced by Gozlan, Roberto, Samson, and Tetali (2017).[9] We depart from this strand by working with a continuum of workers and a continuum of firms (the so-called "large firms" literature). Using the literature on multidimensional OT, Chiappori, McCann, and Pass (2016) connect their work to the multidimensional screening literature (see Rochet and Choné, 1998 in the monopoly context) where Bunching plays an important role. In the present paper, we find something akin to bunching in a competitive environment with multidimensional types where firms and workers have the same dimension of heterogeneity. We discuss connections between this literature and our contribution in what follows.

Overview: We derive how wages, skill prices, and worker-to-firm sorting are jointly determined in equilibrium on a labor market where the only friction is the inability of workers to sell their multidimensional skills as separate inputs to different firms. To

[7]Firms' production functions matter – but not sorting – in recent GE models of monopsonistic labor markets (Berger, Herkenhoff, and Mongey, 2022 or Deb, Eeckhout, Patel, and Warren, 2024).

[8]Indeed, we are not aware of any data source that would offer a comprehensive picture: workers' exact skills, the exact tasks each worker performs, together with the worker's employing firm. Often occupations are used as a proxy even though the tasks performed by the worker in her employing firm are virtually never measured. However, Bittarello, Kramarz, and Maitre (forthcoming) show the extent of dispersion in tasks within occupations.

[9]Appendix F presents the exact connections between competitive matching as examined here and optimal transport theory. Matching in labor markets was also examined in its discrete (game-theoretic) version (see Crawford, 1991, Kelso and Crawford, 1982, Hatfield and Milgrom, 2005, and, more recently, Pycia, 2012 and Pycia and Yenmez, 2019).

discuss the interplay between workers-to-firms sorting and wages, Section 2 presents key motivating facts and the model setup. Section 3 presents aggregate sorting as well as the structure of the equilibrium wages and the associated Generalists' markdowns. Section 4 details workers-to-firms sorting patterns, distinguishing two types of equilibria according to whether generalist workers face a wage markdown and whether firms hire workers with different skill profiles. Section 5 shows that log wages can be decomposed into an individual quality component and a "worker- to-firm sorting effect", relates the latter effect to the wage markdown, and proves that these quantities depend on the employing firm's total factor productivity when the production function is non-homothetic. Section 6 concludes. All proofs are relegated to the Appendices.

2 A Labor Market with Bundled Skills: Suggestive Evidence and Model

In Subsection 2.1, we present motivating facts on skills bundling using Swedish data sources on workers' multivariate skills and their employing firms. Then, in Subsection 2.2, informed by our motivating facts, we introduce our theoretical model of the labor market where firms use their workers aggregated skills as inputs of their heterogeneous production functions.

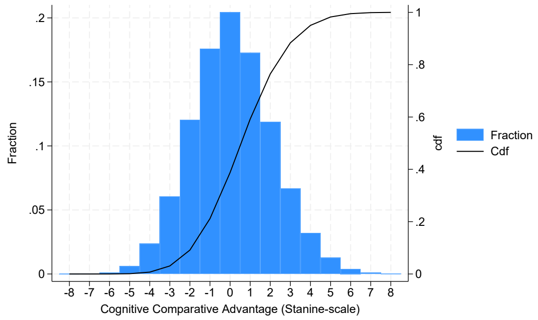
2.1 Suggestive Evidence

We start by presenting three key empirical regularities to motivate our theoretical setup: *i*) workers have multidimensional bundles of skills; *ii*) these skills are unevenly sorted across firms; *iii*) skill sorting across firms is related to firm-level fundamentals. We focus on the key economic insights in the main text, and leave a more detailed description of data and methods for Figures' notes and the Data Appendix H.

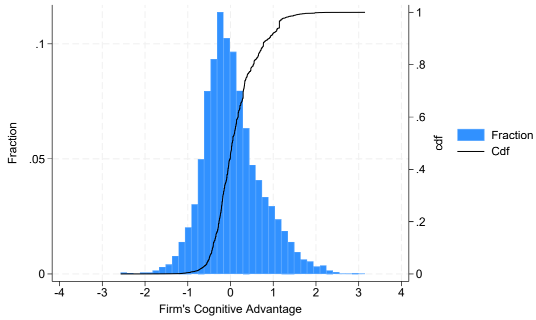
All empirical exercises in this paper use Swedish register data on skills assessed as part of the military conscription performed around age 18 for almost all Swedish males across three decades. Data are linked to wages, occupations, and employing firms' financial accounts. We use data on a two-dimensional skill bundle—Cognitive and Non-Cognitive skills—where each skill is measured on an integer *Stanine* scale approximating a normal distribution through 9 discrete bins (one step = 0.5 standard deviations).[10] Scores are normalized w.r.t. economy-wide peers (separately by skill) and the difference between a worker's Cognitive and Non-Cognitive scores thus captures his

[10] In contrast to percentile ranked bins that generate a uniform distribution, Stanine bins approximate a normal distribution and bins are therefore of unequal size, each capturing $4, 7, 12, 17, 20, 17, 12, 7, 4$ percent of the sample as we move from 1 to 9.

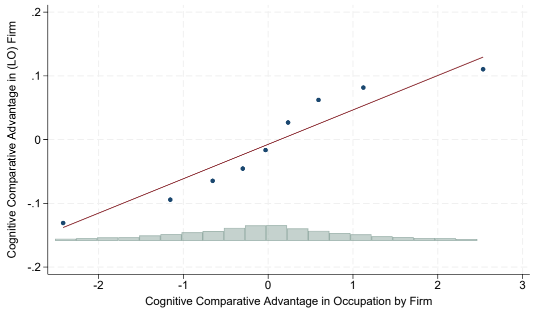
comparative advantage relative to these peers. Figure 1(a) shows that about half of all workers have a comparative advantage in one of the skill dimensions of at least 1 standard deviation (2 Stanine steps).



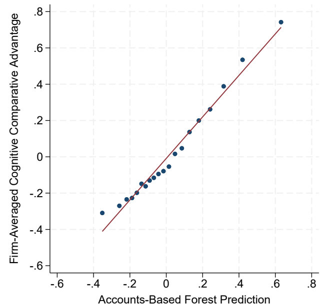
(a) Worker-level comparative advantage



(b) Firm-averaged comparative advantage



(c) Leave-one-out associations across occupations, within firms



(d) Predicted vs. actual firm-averaged comparative advantage

Figure 1: Cognitive comparative advantage of workers and firms.
Data sources: Swedish skills data measured on 1-9 Stanine scale for male workers during 2002 – 13. Comparative advantage (CA) is the skill difference along the Stanine scale. Panel a) Worker-level CA distribution. Panel b) CA averaged at the firm level, cut at $[-4, 4]$. Panel c) CA first de-meaned by occupation x year and then averaged by occupation within firm. The panel shows leave-one-out (LO) associations across occupations within firm. Panel d) Random Forest prediction related to actual outcomes. Forest is estimated using 25 variables from the firms economic accounts using scikit-learn accessed via Stata using default parameter settings. The Panel shows out-of-fold predictions, averaged from five firm-clustered data folds. Further details on data and methods can be found in Appendix H.

Figure 1(b) relates workers' comparative advantage to their sorting patterns across firms. The figure shows the distribution of *firm-averaged* comparative advantages, and the dispersion illustrates that firms are heterogeneous in their use of cognitive vs. non-cognitive skills. The dispersion reflects *heterogeneous firm-level labor demand* and not random noise—as Figure 1(c) shows, firms with an unusually high cognitive advantage in one occupation also have an unusually high cognitive advantage in their other occupations. Hence, firms are systematically heterogeneous in their use of cognitive vs non-cognitive skills in ways that go beyond occupational boundaries.

The firm-averaged comparative advantage is *predictable from firm-side fundamentals* as reflected in the firms' financial accounts. In Figure 1(d) we show that a Random Forest model can be trained to predict (out-of-sample) the firm-averaged comparative advantages using data from their financial accounts. The prediction output generates a *continuous* distribution of firm-averaged comparative advantages, consistent with our modeling approach where a continuum of heterogeneous production functions plays a central role.

2.2 Model

In the labor market we study, a continuum of heterogeneous workers supplies multidimensional skills to a continuum of heterogeneous firms, in the spirit of the empirical patterns illustrated above. The firms produce output by aggregating their employees' skills into tasks, used as inputs of a concave production function. We examine how the introduction of a single friction – the bundling of workers' skills – affects the equilibrium outcomes in this otherwise perfectly competitive world. We first describe production. Then, we define the matching concept, the assignment of workers to firms, and the notion of competitive equilibrium adopted to our setting.

Production: The firm's production process involves k intermediary inputs. Firms first transform a worker's skills into a worker's task. Then, the firm aggregates these workers' tasks. These aggregates constitute the intermediary inputs used to produce final output. Firms are heterogeneous in their production technologies. More precisely, denoting by $T = (T_1, \ldots, T_k)$ the aggregate vector of tasks produced by its employees, a firm of type ϕ produces final output $F(T; \phi)$, with F being nonnegative, continuous in (T, ϕ), concave in T, differentiable in T with partial derivatives continuous in (T, ϕ). We furthermore assume that the marginal productivities $\partial F/\partial T_j(T; \phi)$ are continuous in ϕ. Firms' types are distributed according to a probability measure $H^f(\mathrm{d}\phi)$ on a compact support $\mathcal{X}^f \subset \mathbb{R}^k_+$.

Of particular interest to us are production functions of the form $F(T; \phi) = zF(T; \alpha)$ with the firms' types $\phi = (\alpha, z)$ having two components: z reflects total factor productivity and α reflects the relative importance of each task in the production process. We assume that the worker and firm heterogeneities have the same dimension, hence α lies in a space of dimension $k - 1$.

Workers, Skills, and Tasks: Each worker is endowed with a set of skills, $\mathbf{s} = (s_1, \ldots, s_k)$, $k \geq 2$. These vectors, $\mathbf{s}$, belong to a compact subset $\mathcal{X}^s$ of $\mathbb{R}^k_+$ and are distributed according to a probability measure $H^w(\mathrm{d}s)$ in the population of workers. We assume that there is no mass of workers with zero skills, $H^w(\{0\}) = 0$.

In most of the paper, we assume that a worker with skill set s produces task $t = s$, hence a worker with twice the skill-level in a given dimension can produce twice as many tasks of that type in a given day.[11] In other words, we equate skills with tasks, even though skills are worker-side characteristics and tasks are firm side attributes.[12] We consider more general skills-to-tasks relationships of the form $t = g(s)$ in Appendix B.

Following Acemoglu and Autor (2011), our baseline specification assumes that firms aggregate tasks performed by their employees in a linear way. Since an employee with skill vector s produces task t, the amount of task j produced in firms of type ϕ is

$$\mathbf{T}_j^{\mathrm{D}}(\phi) = \int s_j \, n^{\mathrm{D}}(\mathrm{d}s; \phi), \tag{1}$$

where $n^{\mathrm{D}}(\mathrm{d}s; \phi)$ is a non-negative finite measure on $\mathcal{X}^s$ that represents the number of workers with skills s hired by a firm of type ϕ. We discuss alternative rich aggregation schemes in Appendix B.

Remark: The standard formulation of production functions with heterogeneous workers assumes that each worker has a unique skill type (say high or low educated) such that a pure type-specific count of workers can be fed into a production function as separate inputs. Our production functions instead allow output to depend on the within-firm aggregate skill vector, $\int s \, n^{\mathrm{D}}(\mathrm{d}s; \phi)$. This is the simplest and most natural extension of usual production functions when incorporating the bundling friction and, as shown below, it will bring useful structure to the problem.[13] We discuss extensions to broader classes of production functions in the Conclusion. When the production function displays increasing marginal productivities of aggregate skill types,[14] the marginal productivity of a worker in one skill increases with her co-workers' other skills. Hence complementarities across workers result from complementarities across skill types.

Assignment: An assignment of workers to firms is a family of a non-negative finite measures $n^{\mathrm{D}}(\mathrm{d}s; \phi)$ on $\mathcal{X}^s$. To allow for varying firms' sizes, we do not use probability measures as is common in the literature. Instead, the measure of workers employed by a firm of type ϕ, denoted $N(\phi) = n^{\mathrm{D}}(\mathcal{X}; \phi)$ is endogenously determined. As a

[11] Conceptually, in relation to our empirical two-dimensional application with Cognitive and Non-Cognitive skills measured on a scale of 1 to 9, a worker with a (4,2) skill set may be considered as able to perform four Cognitive and two Non-Cognitive tasks during a work day.

[12] Choné, Gozlan, and Kramarz (2024) allow labor supply to respond to wages, thus endogenizing the relationship between the workers' skills and the tasks they perform. The skills-equated-to-tasks situation obtains as a special case of their model where tasks are perfect complements in the workers' disutility cost of work. Most of our results (e.g., generalist markdowns, sorting, uniqueness of firm size) go through in the endogeneous labor supply context.

[13] See also the taxonomy of optimal transport problems in Appendix F.

[14] Formally: $\partial^2 F / \partial T_j \partial T_k > 0$ for $j \neq k$. In this case, the aggregate skills are gross-complements, i.e., the demand for one skill is non-increasing in the prices of the other skills, see Theorem 4.D.3 in Takayama (1985).

consequence, the aggregate distribution $n^{\mathrm{D}}(\mathrm{d}s;\phi)$ is also endogenously determined in equilibrium.

An assignment n^{D} clears the labor market if

$$\int n^{\mathrm{D}}(\mathrm{d}s;\phi)\, H^f(\mathrm{d}\phi) = H^w(\mathrm{d}s) \tag{2}$$

for all skill vectors $s \in \mathcal{X}^s$The market-clearing assignment quantify the number of workers with skill s hired by firms of each type ϕ. In short form, we write the market clearing equation (2) as $n^{\mathrm{D}} H^f = H^w$. Integrating this equation with respect to s shows that, for any market-clearing assignment n^{D}, the expected number of employees over all firms is one (without loss of generality):

$$\int N(\phi) H^f(\mathrm{d}\phi) = 1. \tag{3}$$

It follows that the (endogenously) modified distribution of firms' types $\tilde{H}^f(\mathrm{d}\phi) = N(\phi)H^f(\mathrm{d}\phi)$ is a probability measure that is absolutely continuous with respect to the original distribution $H^f(\mathrm{d}\phi)$. In Appendix F, we formally define the matching between workers' skills s and firms' technologies ϕ as a "transport plan" and connect our framework to recent developments in optimal transport theory.

A market-clearing assignment n^{D} is *optimal* if it maximizes total output in the economy, i.e., if it solves

$$\mathcal{J}^b(H^f, H^w) \overset{\mathrm{d}}{\equiv} \sup_{n^{\mathrm{D}} \mid n^{\mathrm{D}} H^f = H^w} \int F\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi);\phi\right) H^f(\mathrm{d}\phi). \tag{4}$$

We show in Lemma G.1 that $\mathcal{J}^b(H^f, H^w) < \infty$. As our production function F is nonlinear in the firm-aggregate vectors of tasks T, the total output in the economy is a nonlinear function of the assignment n^{D}.[15]

Competitive Equilibrium: Under bundling, a worker's set of skills cannot be unpacked, hence firms must purchase her entire skill package $s = (s_1, \ldots, s_k)$. The market wage of a worker with skill s is denoted by $w(s)$. The wage schedule $w(.)$ is therefore a map: $\mathcal{X}^s \to \mathbb{R}_+$. We rule out agency problems and information asymmetries: a firm that hires a worker with skill s obtains the vector of tasks s in return for the paid wage $w(s)$. Given a wage schedule $w(.)$, the demand for skill is the assignment

[15]By contrast, if firms' production instead had been the sum of each of their employees' production, then total output $\iint F(s;\phi)\, n^{\mathrm{D}}(\mathrm{d}s;\phi) H^f(\mathrm{d}\phi)$ would have been linear in n^{D}, see Appendix G.

$n^{\mathrm{D}}(\mathrm{d}s;\phi)$ on $\mathcal{X}^s$ that maximizes the firms' profit:

$$\Pi(\phi;w)=\max_{\nu\in\mathcal{M}(\mathbb{R}^k_+)} F\left(\int s\,\nu(\mathrm{d}s);\phi\right)-\int w(s)\,\nu(\mathrm{d}s), \tag{5}$$

where $\mathcal{M}(\mathbb{R}^k_+)$ is the set of all positive measures on $\mathbb{R}^k_+$. A ***competitive equilibrium*** *is a pair (n^{D},w) composed of a wage schedule w and a market-clearing assignment n^{D} such that n^{D} reflects the demand for skills under the wage w, i.e., $n^{\mathrm{D}}(\mathrm{d}s;\phi)$ solves (5) for all firms' types ϕ.*

3 Aggregate Outcomes

This Section is organized in order to give us the tools necessary to examine both between-firms and within-firms inequality in labor markets faced with a bundling friction.

In Subsection 3.1, we show the existence of competitive equilibria and derive properties of the wage schedule. In Subsection 3.2 we then show that each firm's size is uniquely defined in equilibrium. Then, in Subsection 3.3, we present the structure of aggregate sorting (which will be contrasted later with individual sorting). In Subsection 3.4, we present two properties of the equilibrium wage schedule – one-homogeneity and convexity – and how the latter generates wage markdowns for "Generalist" workers endowed with a balanced set of skills. Finally, in Subsection 3.5, we provide an empirical estimation of the Generalists' Markdown.

3.1 Competitive Equilibria with a Convex Wage Schedule

Competitive equilibria will be shown to exist under the following assumption:

Assumption 1. *For all $T\in\mathbb{R}^d_+$ and all $\phi\in\mathcal{X}^f$, $F(\mu T;\phi)/\mu$ tends to 0 as $\mu\to+\infty$.*

Assumption 1 holds for homogenous production functions with decreasing returns to scale, as is the case in our leading example (11). It also holds for the non-homothetic production function used in Appendix E.2.

The bundling environment is characterized by missing markets. Firms cannot purchase some amount of skills, separately for each skill type $j=1,\dots,k$. In particular the notion of optimality refers to the "Primal" Problem (4), which includes the constraints that only workers can be hired and that only skill-*vectors* can be traded.

Proposition 1 (Existence of equilibria with a convex wage schedule)**.** *Under Assumption 1 there exist competitive equilibria (n^{D},w), with the wage schedule being convex* *and homogenous of degree one. Any equilibrium assignment of workers to firms n^{D} is optimal.*

To prove the existence of competitive equilibria, we prove the existence of optimal market clearing-assignments and show that they can be decentralized by convex and 1-homogeneous wage schedules, see Appendix A.1. In the following paragraphs, we present an intuitive proof.

Convexity and 1-Homogeneity of the Wage Schedule: The convexity and the homogeneity of the wage schedule are direct consequences of the linear aggregation of workers' skills into k inputs used by to produce, as shown in equation (1). Fulfillment of these two conditions guarantee the absence of arbitrage opportunities for firms. Indeed, when these properties do not hold, firms can reduce their wage bill by replacing some workers with combinations of workers yielding the same aggregate skills.

Suppose for instance that there exist worker types s, s', and s'' such that $s'' = \nu s + (1-\nu)s'$ with $0 < \nu < 1$, $w(s) = w(s') = 1$, and $w(s'') > 1$. Then, no firm would want to hire type-s'' workers because a combination of type-s and type-s' workers would deliver the same amount of intermediary inputs in return for a lower wage bill. Specifically, diminishing demand $n^{\mathrm{D}}(s'', \phi)$ by ε and increasing $n^{\mathrm{D}}(s, \phi)$ by $\nu\varepsilon$ and $n^{\mathrm{D}}(s', \phi)$ by $(1-\nu)\varepsilon$ leaves the firm-aggregated vector of skills (hence the firm-level vector of tasks) unchanged and reduces the wage bill.

Similarly, the 1-homogeneity property directly stems from the linear aggregation of skills in the production function. Consider two workers with proportional skills s and μs for some $\mu > 0$. These workers have the same relative skill endowments but differ in their overall quality, embodied by the multiplicative factor μ. Assume, by contradiction, that $w(\mu s) < \mu w(s)$. Then no firm would hire worker type s as diminishing $N(s; \phi)$ by ε and increasing $N(\mu s; \phi)$ by ε/μ leaves the firm aggregate of skills unchanged while reducing the wage bill. It follows that the demand for worker s is zero, a contradiction. The reverse inequality, $w(\mu s) > \mu w(s)$, is ruled out by the same argument.

3.2 Firm Size is Uniquely Defined

We show that the firm's aggregated skill vector $\mathbf{T}^{\mathrm{D}}(\phi)$ introduced in (1) is uniquely defined and introduce a well-defined notion of firm's size. Hereafter, we denote by Φ^{+}_{nd} the set of all convex, positively one-homogenous, and non-decreasing functions $w : \mathbb{R}^k_+ \to \mathbb{R}_+$.

Proposition 2 (Uniqueness of wage schedule and firm-aggregated skill vector)**.** *Suppose Assumption 1 holds and assume furthermore that $F(T; \phi)$ is* strictly *concave in T. Then the firm-aggregated skill vector $\mathbf{T}^{\mathrm{D}}(\phi) = \int s\, n^{\mathrm{D}}(\mathrm{d}s; \phi)$ is unique among all competitive* *equilibria* (n^{D}, w), *and is the solution of*

$$\Pi(\phi; w) = \max_T F(T; \phi) - w(T), \tag{6}$$

where w *is any equilibrium wage schedule that is convex and homogenous of degree one. Furthermore, the wage schedule* $w \in \Phi^+_{\mathrm{nd}}$ *is uniquely defined on the support of* $\mathbf{T}^{\mathrm{D}}$*, i.e.,* $w(\mathbf{T}^{\mathrm{D}})$ *is unique.*

Since F is concave and w is convex, the above problem is well-posed, with a unique solution characterized by

$$F_j(\mathbf{T}^{\mathrm{D}}(\phi); \phi) = w_j(\mathbf{T}^{\mathrm{D}}(\phi)). \tag{7}$$

At any competitive equilibrium, the marginal productivity of each skill equals its marginal price. When the wage schedule is locally linear, i.e., is of the form $\sum \bar{w}_j s_j$, we are back to $F_j(\mathbf{T}^{\mathrm{D}}(\phi); \phi) = \bar{w}_j$, i.e., the marginal productivity of each skill equals its unique price. More generally, the productivity of skill j equals its *implicit* price in the neighborhood of the aggregate skill T, i.e., the partial derivative $w_j = \partial w / \partial s_j$ evaluated at that point. Figure 2 shows the tangency of the firm's production isoquant and the iso-wage surface.

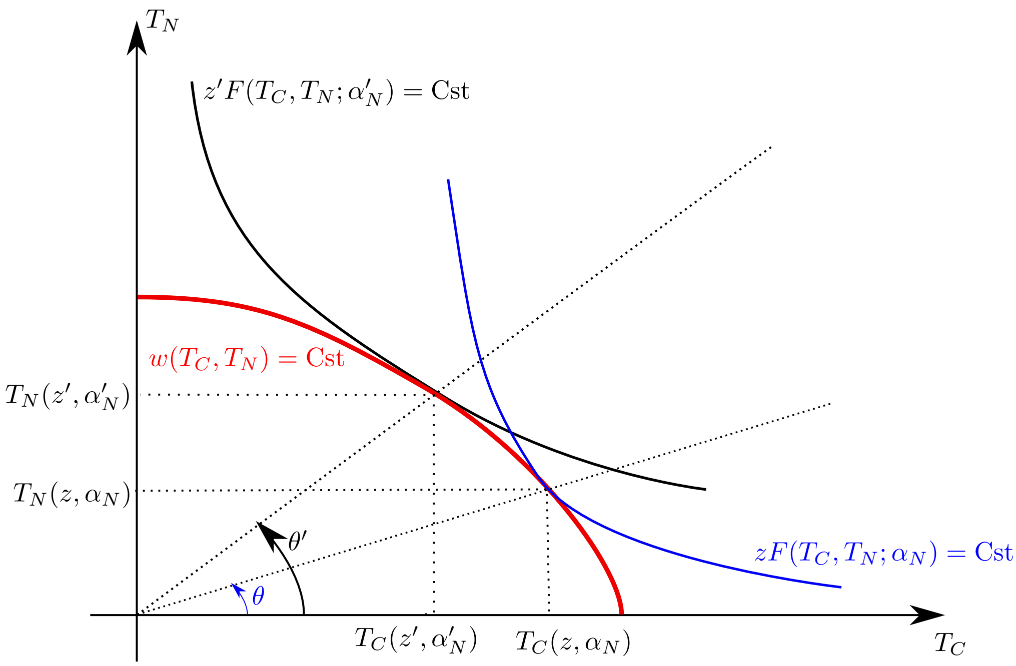


Figure 2: Matching in the horizontal (skill profile) dimension: Firm (α'_N, z') is more intensive in skill N than firm (α_N, z), with $\alpha'_N > \alpha_N$

The Substance of Firms: Our framework distinguishes itself from the job-level approaches dominating the labor literature because it gives substance to firms in equilibrium.[16] We allow firms to transform their workers' skills into tasks which are aggregated and then used as inputs in firm-specific production functions. As a consequence, a firm's size is shown to be a uniquely-defined concept. This stands in contrast with frameworks where firms are redundant because total production is its sum across jobs.

We define the *size* $\Lambda^{\mathrm{D}}(\phi)$ of firm type ϕ by valuing the skills it uses at the prices assigned to pure specialist workers. These specialists only have one skill and their wage $w(e_i)$, $e_i = (0, \ldots, 1, \ldots, 0)$, with 1 in the ith coordinate, thus reflects the price of skill i in a setting without bundling constraints. Thus, the *size* of firm type ϕ is:

$$\Lambda^{\mathrm{D}}(\phi) = \sum_i w(e_i)\mathbf{T}_i^{\mathrm{D}}(\phi). \tag{8}$$

Note that $w(e_i)$ is uniquely defined in equilibrium according to Proposition (2).[17]

Proposition 3. *Consider a production function of the form $zF(T_C, T_N; \alpha_N)$. Then the size of the firm, $\Lambda^{\mathrm{D}}(z, \alpha_N)$, increases with the firm's total factor productivity z.*

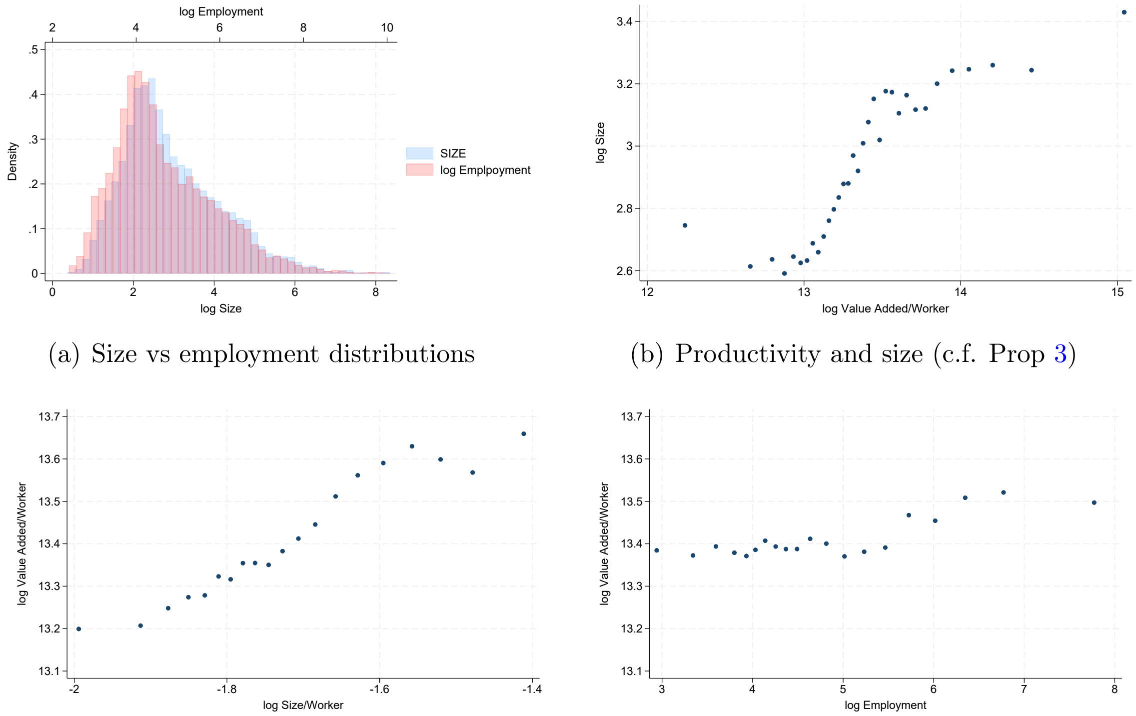
Size of Firms: An empirical illustration The measure of firm *size* differs from traditional concepts based on the number of workers. To illustrate its empirical content, we need to characterize firms' use of skills and weight these by the skill prices of pure specialists. As an empirical approximation, we use our Swedish skill data and estimate within-firm returns to Cognitive (resp. non-Cognitive) skills in firms that are dominated by Cognitive (resp. non-Cognitive) workers.[18] Next, we compute the empirical counterpart of *size* by weighting the skill sum of each firm by these wage estimates.[19] The ensuing log *size* distribution is more symmetric than the distribution of the log number of employees as shown in Figure 3(a). Proposition 3 shows that *size* should grow with total factor productivity and, as shown in Figure (3(b)), we find a clear association between *size* and measured labor productivity in our data. In the final two panels, we put labor productivity on the y-axis and show separately how it relates to *size* and log number of employees, holding the other factor (i.e. either *size* or employment) fixed, using the approach of Cattaneo, Crump, Farrell, and Feng (2024). As is evident, *size* is positively related to labor productivity, conditional on the log number of employees (Figure 3(c)), whereas the converse is not true, suggesting that the size concept indeed has clear economic meaning.

[16] See, however, Eeckhout and Kircher, 2018 for a firm-level exception.

[17] If e_i does not belong to the support $\mathcal{X}^s$ of the skill distribution H^w, i.e., if no worker in the economy is a pure specialist, we extend the wage schedule on $\mathbb{R}_+^k$ using Equation (33). See also Appendix A.2.

[18] Details on the exact procedure are presented below.

[19] As we do not have skills for all workers (but, given our sample restrictions, for at least 10, and 1/4, of workers in each firm) we use the average skill of observed workers and multiply it by the full



(a) Size vs employment distributions (b) Productivity and size (c.f. Prop 3)

(c) Productivity and size, cond. on employment (d) Productivity and employment, cond. on size

Figure 3: Size, Employment, and Value-added per Worker

3.3 Aggregate Sorting

We now study how the firm-aggregated skills vector $\mathbf{T}^{\mathrm{D}}(\phi)$ varies with the firm's type ϕ. We conduct the analysis under the assumption that the production function can be written as $F(T,\phi) = zF(T_C, T_N; \alpha_N)$, with $\phi = (\alpha, z)$, where z and α_N reflect total factor productivity and the relative importance of skill N in the production process.

The *firm-aggregated skill profile* $\boldsymbol{\theta}^{\mathrm{D}}(\phi)$ is skill use, normalized by firm *size*:

$$\boldsymbol{\theta}^{\mathrm{D}}(\phi) = \mathbf{T}^{\mathrm{D}}(\phi)/\Lambda^{\mathrm{D}}(\phi). \tag{9}$$

To simplify exposition, we concentrate on the two-skill case, and denote these skills by C and N to concur with our Cognitive vs. Non-cognitive skill data. In this case, any aggregate skill profile $\boldsymbol{\theta}^{\mathrm{D}}$ can equivalently be represented by the angle $\theta^{\mathrm{D}}(\phi)$ such that $\tan\theta^{\mathrm{D}} = T_N/T_C$. The aggregate workers-to-firms matching can thus be represented as $\mathbf{T}^{\mathrm{D}}(\phi) = (\Lambda^{\mathrm{D}}(z,\alpha_N), \theta^{\mathrm{D}}(z,\alpha_N))$. We already know from Proposition 3 that the firm's size Λ^{D} increases with total factor productivity z. We now examine whether the skill profile θ^{D} increases with the technological parameter α_N. In the proposition below, we

number of workers, thus implicitly imputing the skills for remaining workers to match those of included workers.

rely on the definition adopted by Lindenlaub (2017) to generalize the notion of positive assortative matching in multidimensional settings.[20]

Proposition 4. *The firm-aggregated matching function* $(\Lambda^{\mathrm{D}}(z,\alpha_N), \theta^{\mathrm{D}}(z,\alpha_N))$ *exhibits positive assortative matching if and only if*

$$\frac{\partial F_C}{\partial \alpha_N}\frac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - \frac{\partial F_N}{\partial \alpha_N}\frac{\partial F_C}{\partial \Lambda^{\mathrm{D}}} \geq 0. \tag{10}$$

Positive assortative matching holds in particular if the production functions have homothetic isoquants and F_N/F_C *increases with* α_N*.*

Hence, when (10) holds, the aggregate skill profile $\theta^{\mathrm{D}}(\phi) = \arctan \mathbf{T}_N^{\mathrm{D}}/\mathbf{T}_C^{\mathrm{D}}$ increases with the importance of non-cognitive skills in the firm production function, α_N. Moreover, the aggregate sorting exhibits a PAM property à la Lindenlaub (2017) when production functions are homothetic.[21] While Lindenlaub (2017) considers the matching between individual workers' and jobs' characteristics, the PAM property in our context applies to firms' *aggregates* rather than to individual workers' characteristics. At the individual level, sorting patterns may be blurred by bunching, as we discuss in Section 4.3.

Two-skill CES example Our leading example of homothetic production function exhibits constant elasticity of substitution and decreasing returns to scale:

$$zF(T;\alpha) = (z/\eta)\,\left[\sum_{j=1}^{k} \alpha_j T_j^{\rho}\right]^{\eta/\rho}, \tag{11}$$

with $\sum_{j=1}^{k} \alpha_j = 1$, $0 < \eta < 1$, and $\rho < 1$. When $\rho < \eta$, the production function displays increasing marginal productivities of aggregate skill types, i.e., $\partial^2 F/\partial T_j \partial T_k > 0$ for all $j \neq k$. The marginal productivity of a worker in one skill increases with her co-workers' other skills (recall Footnote 14).

With two skills, we get:

$$zF(T_C, T_N; \alpha_N) = \frac{z}{\eta}\left[\alpha_C T_C^{\rho} + \alpha_N T_N^{\rho}\right]^{\eta/\rho}, \tag{12}$$

with $\alpha_C + \alpha_N = 1$. In this case, the general workers-to-firms matching condition (7) simplifies to:

$$\left[\tan \theta^{\mathrm{D}}(\alpha_N)\right]^{1-\rho} = \frac{\alpha_N}{1-\alpha_N}\,\frac{w_C\left(\theta^{\mathrm{D}}(\alpha_N)\right)}{w_N\left(\theta^{\mathrm{D}}(\alpha_N)\right)}. \tag{13}$$

[20] According to this definition, the sorting exhibits positive assortative matching (PAM) if the Jacobian $D_{(z,\alpha_N)}(\Lambda^{\mathrm{D}}, \theta^{\mathrm{D}})$ is a P-matrix, i.e., all the principal minors of the Jacobian are positive. The PAM property implies that Λ^{D} and θ^{D} increase with z and α_N respectively.

[21] We examine sorting patterns under non-homotheticity in Subsection 5.2.

The increasing function $\theta^{\mathrm{D}}(\alpha_N)$ implicitly defined by (13) represents the matching between workers and firms. The relative skill endowment in non-cognitive skills of the workers, $\theta^{\mathrm{D}}(\alpha_N)$, increases with the demand intensity in that skill, α_N, as illustrated in Figure 2. Firm size $\Lambda^{\mathrm{D}}(\alpha_N; z)$ is given by (A.10) in the Appendix.

3.4 Convex Wages Imply a Generalist Markdown

As when discussing the substance of firms above, we can consider the skill vector of pure specialists that only are endowed with one skill i with a price equal to the wage $w(e_i)$. Because w is convex and homogenous of degree one, it is sub-additive, which implies that:

$$w(\mathbf{s}) = w\left(\sum_{i=1}^{k} s_i e_i\right) \leq \sum_{i=1}^{k} w(e_i s_i) = \sum_{i=1}^{k} w(e_i) s_i. \tag{14}$$

We denote the marginal return to skill i by $w_i(\mathbf{s}) = \partial w / \partial s_i$, which we alternatively call the implicit price of skill i. Because w is 1-homogenous, a worker with skill vector $\mathbf{s}$ is paid $w(\mathbf{s}) = \sum_i w_i(\mathbf{s}) s_i$, with the implicit price $w_i(\mathbf{s}) = w_i(\boldsymbol{\theta})$ being a function of the worker' skill profile $\boldsymbol{\theta}$. For pure specialists, we have $w(e_i) = w_i(e_i)$.

Skill Profiles and the Generalists' Markdown We measure the overall quality of a worker with skill vector $\mathbf{s}$ by valuing her skills at the prices of pure specialist workers:

$$\lambda(\mathbf{s}) = \sum_i w(e_i) s_i, \tag{15}$$

where $w(e_i)$ is uniquely defined in equilibrium (see Proposition 2). In what follows, we will refer to worker's quality as the *vertical dimension* of a worker's heterogeneity. To describe the *horizontal dimension* of a worker's heterogeneity, we normalize each worker's skill vector by her quality, thus obtaining her skill "profile" $\boldsymbol{\theta} = \mathbf{s}/\lambda(s)$. Hence, by definition, skill profiles have quality equal to 1.

Figure 4 shows the case with two skills $i \in \{C, N\}$. The wages of the pure cognitive and non-cognitive specialists are denoted by $p_C = w(1, 0)$ and $p_N = w(0, 1)$, respectively. The bold dashed line represents the skill vectors of quality 1, i.e. the set of $\mathbf{s}$ such that $\lambda(\mathbf{s}) = p_C s_C + p_N s_N = 1$. The generalist worker $\mathbf{G} = (G_C, G_N)$ is located above the bold dashed line and, thus, has quality $\lambda(\mathbf{G}) > 1$. By construction (see just above), her skill profile $\boldsymbol{\theta}_G$ has quality $\lambda(\boldsymbol{\theta}_G) = 1$.

The sub-additivity of the wage schedule translates into a generalist markdown:

$$m(\boldsymbol{\theta}) = \frac{\lambda(\mathbf{s})}{w(\mathbf{s})} = \frac{\sum_i w(e_i) s_i}{w(\mathbf{s})} \geq 1, \tag{16}$$

which depends only on the worker's skill profile. Generalist workers are paid (weakly) less than the equivalent combination of pure specialists. For instance, the generalist worker $\mathbf{G} = (G_C, G_N)$ represented on Figure 4 is paid $w(\mathbf{G}) = w_C G_C + w_N G_N = 1$ whereas she would be paid $\lambda(\mathbf{G}) = p_C G_C + p_N G_N > 1$ should her skills be compensated on the basis of the specialists' prices p_C and p_N. Thus, the generalist markdown is equal to $m(\mathbf{G}) = m(\boldsymbol{\theta}_G) = p_C G_C + p_N G_N > 1$.

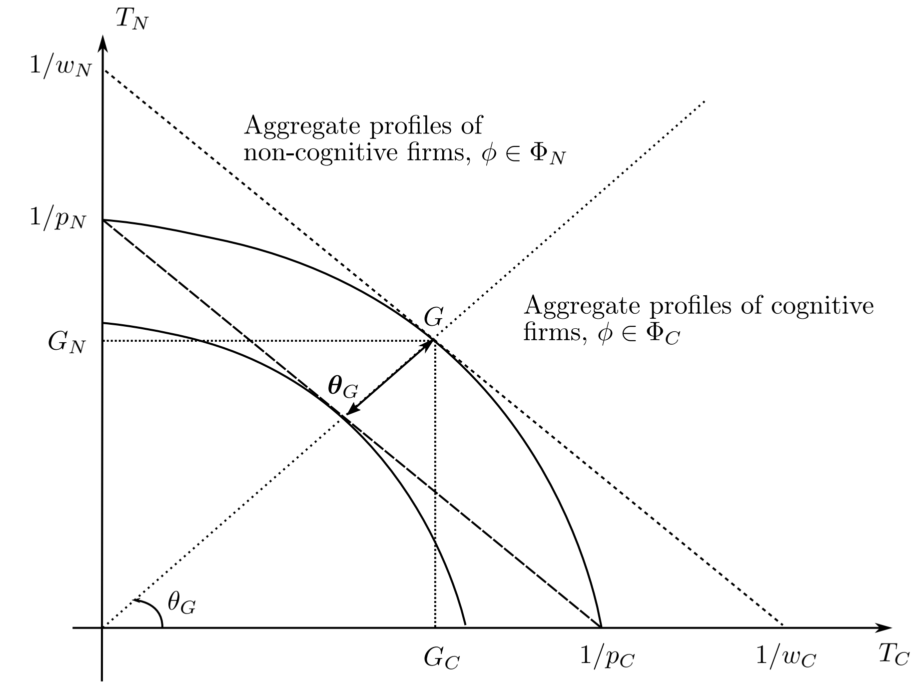


Figure 4: The (solid) isowage lines are homothetic because the wage schedule is homogenous. The bold dashed line represents the set of workers of quality 1. The generalist worker $\mathbf{G} = (G_C, G_N)$ is paid $w(\mathbf{G}) = 1$. Her skill profile is $\boldsymbol{\theta}_G$. The generalist markdown is $m(\mathbf{G}) = p_C G_C + p_N G_N > 1$.

By construction, the minimal value of the generalist markdown is attained for pure specialist workers. The next result examines where the markdown attains its maximum.

Lemma 1. *The maximal value of the generalist markdown is attained for skill profiles such that all implicit prices $w_i(\boldsymbol{\theta})$ are proportional to the pure specialist's prices $w(e_i)$.*

The above result allows to us define the "most generalist" workers in the economy, i.e. those with the highest generalist markdown. This is the case for the worker with type $\mathbf{G}$ on Figure 4. For this worker, the marginal prices w_C and w_N are proportional to p_C and p_N. Geometrically, the tangent to the isowage line at point $\mathbf{G}$ is parallel to the bold dashed line that represents the workers of quality 1.

By convexity of the wage schedule, the generalist markdown increases (i.e., the wage decreases) as $\boldsymbol{\theta}$ moves from the cognitive specialist's profile $(1/p_C, 0)$ to the most generalist skill profile $\boldsymbol{\theta}_G$ and decreases as $\boldsymbol{\theta}$ moves from $\boldsymbol{\theta}_G$ to the non-cognitive specialist's profile $(0, 1/p_N)$.

3.5 Cross-Sectional Estimates of the Generalists' Markdown

The wage markdown implies that skill prices are lower in firms that purchase a balanced set of skills relative to firms with a more specialized skill demand. To take this prediction to the data, we need to estimate the wage returns to skills separately for firms with different skill-demands. Thus, we first aggregate firms into three groups indexed by τ: Generalist firms, C-Specialized firms, and N-Specialized firms, defined as follows: A firm j is C-specialized (resp. N-specialized) if the firm-averaged difference between cognitive and non-cognitive skills in j is greater (lower) than 0.5 on the Stanine scale (thus, 1/4 of the individual-level standard deviation). Otherwise, the firm is labeled as a Generalist firm, which by definition demands a balanced set of skills. Endowed with this definition, around half of all workers are employed in Generalist firms, whereas the rest is split evenly across C/N-specialized firms.

We use the following notation (local to this exercise): For $\tau \in \{C, G, N\}$, the dummy variable D_j^τ is one if firm j is of type τ, zero otherwise. We estimate the marginal returns of cognitive skills and non-cognitive skills for each type of firm, resulting in 6 separate parameters denoted by ϕ_C^τ and ϕ_N^τ respectively. The model includes firm fixed effects to account for unmodelled differences across firms and the unknown constants. Hence, the estimated model has the following form:

$$\ln W_i = \psi_{j(i)} + \sum_{\tau \in \{C,G,N\}} (\, \phi_C^\tau \, C_i + \phi_G^\tau \, N_i \,) \, D_{j(i)}^\tau + X_i \beta + \epsilon_i, \tag{17}$$

where $\psi_{j(i)}$ denotes the firm fixed-effect and where X include controls for 3-digit occupation-specific year-effects as well as a third order polynomial in age. In variations, we let the firm-effects vary across time and interact occupations with firms without altering the estimates of interest in any meaningful way.

Importantly, the sum $\phi_C^G + \phi_N^G$ measures wage returns to $C + N$-skills in generalist firms. The corresponding returns to skills in the C and N specialized firms, in which there is no markdown by construction, are ϕ_C^C and ϕ_N^N respectively. The absolute generalist markdown is the difference between $\phi_C^C + \phi_N^N$ and $\phi_C^G + \phi_N^G$. Our estimate of interest, the empirical counterpart of (16), is the relative markdown:

$$m = \frac{\phi_C^C + \phi_N^N}{\phi_C^G + \phi_N^G}. \tag{18}$$

Table 1: Markdown Regression Results

	Model Variations			Sample Variations	
	(1)	(2)	(3)	(4)	(5)
Relative Markdown	1.19	1.19	1.22	1.27	1.43
Absolute Markdown	0.006***	0.006***	0.005***	0.008**	0.008***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
Combined Returns					
...in Generalist Firms	0.030***	0.030***	0.024***	0.029***	0.019***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
...in Specialized Firms	0.036***	0.035***	0.030***	0.037***	0.027***
	(0.002)	(0.002)	(0.001)	(0.003)	(0.002)
Observations	3,435,998	3,435,998	3,357,226	1,260,748	1,837,650
Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes	Yes
Occupation by Year	Yes	Yes	Yes	Yes	Yes
Firm by Year	No	Yes	Yes	No	No
Occupation-Firm-Year	No	No	Yes	No	No
Sample	Base	Base	Base	Mid-Skilled	High School

Notes: Data for 2002–2013. Including firm-years with complete data for at least 25% of workers and at least 10 observations. Test cohorts 1970–1996. Skills measured on a 1–9 Stanine scale. Threshold for generalist firms: < 0.5 difference in average skills. Outcome is log hourly wage. Mid-Skilled workers have skill levels of 4 to 6 in both dimensions. High School sample includes workers with a High School education. Standard errors in parentheses. * $p< 0.1$, ** $p< 0.05$, *** $p< 0.01$.

Note that identification does not rely on mobility. Consistent with the theory, we estimate the markdown using cross-sectional variations in skill prices across firm types, while these prices are estimated using within-firm variation.[22]

The generalist markdown is greater than one if skill prices are lower for workers in generalist firms when compared to the combined marginal price of skills for workers employed in the specialized firms. Our estimates are presented in Table 1. The relative markdown is stable at 20% across specifications that vary the fixed effects configurations by controlling for firm-by year or firm-by-occupation year fixed effects (first three columns of the Table). A possible concern is that the generalist-specialist dichotomy is harder to detect empirically for workers with very low or very high overall skill levels but as shown in the last two columns, the estimated markdown increases when focusing on workers at the middle of the skill distribution.

4 Worker-Level Outcomes

To describe the way individual workers and their individual profiles match with firms, we contrast two possible firm-specific hiring policies at equilibrium: hiring a unique worker's skill profile vs hiring a diversity of workers' skill profiles. In Subsection 4.1, we relate aggregate sorting to individual profiles. Then, in Subsection 4.2, we examine pure sorting i.e. when all workers hired at a firm share the same profile, a profile also shared with the firm's and the implied Generalists' markdown. By contrast, in Subsection 4.3, we examine the associated diversity of skill profiles hired within a firm when Bunching emerges with the associated locally linear wage schedule. Finally, in Subsection 4.4, we show how the supply of Generalists versus that of Specialists determines when Pure Sorting emerges and when Bunching takes place (using both theory and some supportive empirical evidence).

4.1 Two Types of Local Equilibria

The results presented below detail the qualitative properties of the matching between firms and *individual* workers. The results mathematically formalize the intuition that, *locally in the skill space*, either the generalists' supply dominates or the specialists' supply dominates. In the former case (studied in Subsection 4.2), we get linear wages, bunching, and no local generalist markdown. In the latter (studied in Subsection 4.3), we get strictly convex wage schedules,[23] no bunching, and a positive wage markdown.

[22]We do include firm fixed effects, but as we are not including individual fixed effects, our estimates are identified even without mobility.

[23]More precisely, strictly convex iso-wage curves (recall the wage schedule is one-homogeneous.

This dichotomy arises because of the competitive asymmetry already mentioned: specialists can compete with generalists on their turf, but the reverse is not true because of the bundling friction — generalists cannot split themselves across markets.

We start by characterizing how the firm-aggregated skill maps relate to individual profiles.

From aggregate to individual matching: The aggregation of skills performed by firms, induced by their labor demand $\mathbf{T}^{\mathrm{D}}(\phi)$ (the aggregate sorting map), "transforms" the measure of skills which becomes different from the one available originally in the economy. The firm's demand is more "generalist" than the measure of skills available in the economy, H^w: the firm's has an excess local demand for generalists when the local supply has an excess of specialists.

To make this statement precise, we introduce the probability measure $\mathbf{T}^{\mathrm{D}}H^f$ on the skill space as[24]

$$\mathbf{T}^{\mathrm{D}}H^f(S) = \Pr(\mathbf{T}^{\mathrm{D}}(\phi) \in S) = \int \mathbb{1}_{\mathbf{T}^{\mathrm{D}}(\phi)\in S} H^f(\mathrm{d}\phi),$$

for any set of skills S. In words, $\mathbf{T}^{\mathrm{D}}H^f(S)$ is the probability that a firm chooses its *aggregated* skill vector in S. To define how generalist a distribution of skills is, we rely on the positively homogenous convex (phc) order introduced by Choné, Gozlan, and Kramarz (2023).

Given two positive measures $\mu_0(\mathrm{d}s)$ and $\mu_1(\mathrm{d}s)$ in the space of skills, we say that μ_0 is dominated by μ_1 in the phc order if and only if $\mu_1 h \geq \mu_0 h$ for all positively 1-homogenous convex functions h. Intuitively, this happens when the measure of workers' skills is more generalist under μ_0 than under μ_1. We apply this definition in the lemma presented just below to $\mu_0 = \mathbf{T}^{\mathrm{D}}H^f$, firm's demand, and $\mu_1 = H^w$, workers' supply.

Lemma 2. *For any market-clearing assignment n^{D}, the distribution of skill induced by the firm-aggregated skill profile $\mathbf{T}^{\mathrm{D}}(\phi) = \int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)$ satisfies*

$$\mathbf{T}^{\mathrm{D}}H^f \leq_{\mathrm{phc}} H^w. \tag{19}$$

Conversely, for any given map $T : \Phi \to \mathbb{R}^k_+$ such that $TH^f \leq_{\mathrm{phc}} H^w$, there exists a market-clearing assignment n^{D} such that $T(\phi) = \int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)$.

In other words, the ordering $T_\# H^f \leq_{\mathrm{phc}} H^w$ is not only necessary but also sufficient for T being generated by a skill-aggregation process. As already mentioned but explained in greater detail in what follows, the ordering (19) expresses that there is a local excess supply of specialist workers and a local excess demand for generalist ones.

[24] This measure may also be denoted as $\mathbf{T}^{\mathrm{D}}_\# H^f(S)$, where $\#$ is the "push-forward" operator.

Characterization of the optimal wage schedule The equilibrium wage schedule will allow us to match the two distributions: the demand for skills on one side and the supply for skills on the other in the most efficient way (i.e. by maximizing total production in the economy).

We thus start by providing additional characterizations of the wage schedule, our solution to the dual problem (see Appendix A.1):

$$\mathcal{J}^b(H^f, H^w) = \inf_{w \in \Phi^+_{\mathrm{nd}}} \int \bar{\Pi}(\phi; w) H^f(\mathrm{d}\phi) + \int w(s) H^w(\mathrm{d}s), \tag{20}$$

where Φ^+_{nd} denotes the set of all convex, positively one-homogenous, and non-decreasing functions $w : \mathbb{R}^k_+ \to \mathbb{R}_+$, and where $\bar{\Pi}(\phi; w)$ is equal to:

$$\bar{\Pi}(\phi; w) = \max_{T \in \mathbb{R}^k_+} F(T; \phi) - w(T). \tag{21}$$

We now offer a characterization of the competitive equilibrium that involves the firm-aggregate skill demand $\mathbf{T}^{\mathrm{D}}$ and the wage schedule w:

Proposition 5. *The equilibrium wage schedule $w \in \Phi^+$ is characterized by the existence of a family of positive measures $(Q^t(\mathrm{d}s))$ such that*

(i) $t = \int s\, Q^t(\mathrm{d}s)$ for almost all t;

(ii) $H^w = Q\mathbf{T}^{\mathrm{D}} H^f$;

(iii) w is linear on the support of Q^t for $\mathbf{T}^{\mathrm{D}} H^f$-almost all t,

where $\mathbf{T}^{\mathrm{D}}(\phi)$ is the unique maximizer in Problem (21).

Workers-to-Firms Sorting: The family of positive measures (Q^t) allows to match the (uniquely defined) aggregate demand for skills, $\mathbf{T}^{\mathrm{D}} H^f$, with the distribution of workers' skill sets in the economy, H^w.[25] If the two distributions $\mathbf{T}^{\mathrm{D}} H^f$ and H^w **are equally generalist**, i.e., if $\mathbf{T}^{\mathrm{D}} H^f =_{\mathrm{phc}} H^w$, this can be done at the level of skill profiles, i.e., a firm with a particular need for profile $\boldsymbol{\theta}^{\mathrm{D}}$ meets that need by employing workers that all have individual profile $\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{D}}$. Mathematically, this occurs when the support of Q^t is included in the half-line $\mathbb{R}_+ t$, i.e., in the set of skill vectors s having the same skill profile as t, namely .

If on the contrary the demand for skills $\mathbf{T}^{\mathrm{D}} H^f$ **is more generalist** than the supply H^w ($\mathbf{T}^{\mathrm{D}} H^f <_{\mathrm{phc}} H^w$), then for some firm-type ϕ workers employed by ϕ-firm have different profiles, i.e., the support of Q^t is larger than $\mathbb{R}_+ t$ for $t = \mathbf{T}^{\mathrm{D}}(\phi)$. In this configuration, complementarities across workers within the firm materialize: the productivity

[25]Equation (A.13) in the Appendix connects Q^t to the assignment n^{D}.

of workers endowed with (mostly) one skill and deprived of the other skills is enhanced by the presence of co-workers endowed with the other, complementary, skills.

The Shape of the Wage Schedule: Also key in the above characterization of the equilibrium: the wage schedule is linear in skills on the support of Q^t for all $t = \mathbf{T}^{\mathrm{D}}(\phi)$. Because w is homogenous of degree 1, we already know that w is linear in the vertical dimension, $w(\lambda\mathbf{s}) = \lambda w(\mathbf{s})$. Yet, when there is within-firm heterogeneity in skill profiles, the wage must also be linear in this horizontal dimension.

4.2 Equilibria with Pure Sorting and Positive Markdowns

We now examine how a firm of type ϕ's achieves the aggregated skill vector $\mathbf{T}^{\mathrm{D}}(\phi)$ in the most economical way:

$$w(\mathbf{T}^{\mathrm{D}}(\phi)) = \inf\left\{\int w(\mathbf{s})\, n^{\mathrm{D}}(\mathrm{d}s) : n^{\mathrm{D}} \in \mathcal{M}(\mathcal{X}^s), \int \mathbf{s}\, n^{\mathrm{D}}(\mathrm{d}\mathbf{s}) = \mathbf{T}^{\mathrm{D}}(\phi)\right\}, \tag{22}$$

where $\mathcal{M}(\mathcal{X}^s)$ is the set of all positive measure on $\mathcal{X}^s$ and w is convex and homogenous of degree one.

In this Subsection, we focus on equilibria with pure (horizontal) sorting where the iso-wage surfaces are strictly concave. In this case, the minimization of the wage bill at a given aggregate skill in (22) imposes that firms with type ϕ hires only workers with skill profile $\boldsymbol{\theta}^{\mathrm{D}}(\phi) = \mathbf{T}^{\mathrm{D}}(\phi)/\Lambda^{\mathrm{D}}(\phi)$, see Figure 2. With two tasks ($k = 2$), the equilibrium workers-to-firms matching and wage schedule are then linked through an ordinary differential equation. To show this, we assume as in Proposition 4 that the production function is homogenous of degree $\eta < 1$ and F_N/F_C increases with α_N. As above, we represent the firm-aggregated skill vector $\mathbf{T}^{\mathrm{D}}$ as $(\Lambda^{\mathrm{D}}, \boldsymbol{\theta}^{\mathrm{D}})$, with Λ^{D} being the firm's size and $\boldsymbol{\theta}^{\mathrm{D}}$ its aggregate skill profile. The equilibrium condition (7) yields in this context

$$\frac{F_N\left(\boldsymbol{\theta}^{\mathrm{D}}(\alpha_N);\alpha_N\right)}{F_C\left(\boldsymbol{\theta}^{\mathrm{D}}(\alpha_N);\alpha_N\right)} = \frac{w_N\left(\boldsymbol{\theta}^{\mathrm{D}}(\alpha_N)\right)}{w_C\left(\boldsymbol{\theta}^{\mathrm{D}}(\alpha_N)\right)}, \tag{23}$$

which implicitly defines an increasing matching map $\theta^{\mathrm{D}}(\alpha_N)$. We can write the equilibrium condition for any α_N as

$$\int_0^{\theta^{\mathrm{D}}(\alpha_N)} \Lambda^{\mathrm{S}}(\theta) H^w(\mathrm{d}\theta) = \int_0^{\alpha_N} Z^f(\alpha)\Lambda^{\mathrm{D}}(\alpha_N;1) H^f(\mathrm{d}\alpha), \tag{24}$$

where $\Lambda^{\mathrm{S}}(\theta) = \int_z \lambda H^w(\mathrm{d}\lambda|\theta)$ and $Z^f(\alpha) = \int_z z^{1/(1-\eta)} H^f(\mathrm{d}z|\alpha)$ are exogenous quantities that depend on the primitive distributions H^f and H^w. The left-hand side of (24) represents the total quality of workers with skill profile below $\theta^{\mathrm{D}}(\alpha_N)$. The right-hand side represents the total quality of workers employed by firms with technological parameter below α_N. Differentiating with respect to α_N yields the ordinary differential equation for the matching map $\theta^{\mathrm{D}}(\alpha_N)$:

$$\Lambda^{\mathrm{S}}(\theta^{\mathrm{D}}(\alpha_N))\, h^w(\theta^{\mathrm{D}}(\alpha_N))\, \frac{\mathrm{d}\theta^{\mathrm{D}}}{\mathrm{d}\alpha_N} = Z^f(\alpha_N) h^f(\alpha_N) \Lambda^{\mathrm{D}}(\alpha_N, 1), \tag{25}$$

where the firm's size $\Lambda^{\mathrm{D}}(\alpha_N, 1)$ is given by (A.10) with $z = 1$, and $h^f(\alpha_N)$ and $h^w(\theta)$ are the pdf of α_N and θ.

4.3 Equilibria with Bunching and Locally Linear Wages

We now turn to situations in which different firm-types hire workers with similar skill-types. We refer to this phenomenon as bunching. When the wage schedule w is locally linear, the minimization of the wage bill, see equation (22), is compatible with a firm hiring employees with different skill profiles:

$$\int w(s)\ n^{\mathrm{D}}(\mathrm{d}s; \phi) = w\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s; \phi) \right) = w(\mathbf{T}^{\mathrm{D}}(\phi)).$$

Figure 5 illustrates a case where w is linear on the non-degenerate cone lying between the half-lines (OA) and (OB). To minimize the firm's wage bill, the support of the assignment measure $n^{\mathrm{D}}(\mathrm{d}s; \phi)$ must be included in the cone (AOB). Because the wage schedule w is linear on that cone, we have For instance, firms with type ϕ on Figure 5, rather than picking employees with skills proportional to $\boldsymbol{\theta}^{\mathrm{D}}(\phi)$, i.e., along the half-line $[OM)$, can use skills located in the entire cone AOB. According to equation (19), the distribution of workers' skills H^w lies closer to the boundary of the cone than the demand distribution $\mathbf{T}^{\mathrm{D}}_{\#} H^f$. On Figures 5 and 6, the supply of skills is more concentrated along the half-lines OA and OB, while the demand is more concentrated in the interior of the cone.

Bunching in the horizontal dimension leads to many-to-many matching as illustrated on Figure 6. Firms with different types hire workers with the same skill profile, and workers with the same type may be employed by firms with different technologies. For instance, firms ϕ and ϕ' on the figure, which have different technological intensities in skill α, both hire workers with skills in the cone (AOB). In the extreme case where workers' skill are located only along the two rays (OA) and (OB), firms F and F' both hire workers with skill profiles A and B, but in different proportions to achieve their aggregate demand.

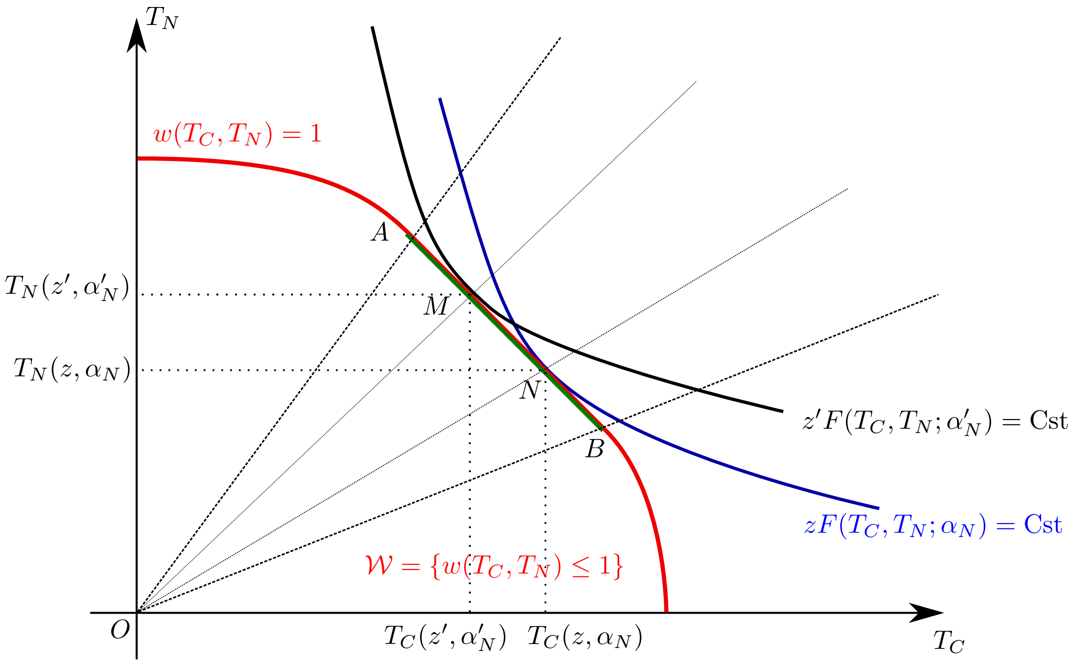


Figure 5: Matching is not pure. Firms $\phi = (\alpha_N, z)$ and $\phi' = (\alpha'_N, z')$, pick their employees in the cone (AOB). Firm ϕ' is more intensive in skill N: $\alpha'_N > \alpha_N$ and $\theta^{\mathrm{D}}(\alpha'_N) > \theta^{\mathrm{D}}(\alpha_N)$.

4.4 The Determinants of Pure Sorting vs Bunching

In this Subsection, we first illustrate our previous results in a simplified economy. To do so, we show how the shape of the equilibrium wage schedule and the degree of within-firm heterogeneity in skill profiles depends upon the fundamentals of this simplified economy. We also explain how the wage schedule flattens when the availability of specialist workers (relative to generalists) increases. Then, we present more general results – using comparative statics techniques – allowing us to compare economies with more Generalists with economies having a larger supply of Specialists. We also present supportive empirical evidence of our comparative statics results.

A simple economy with two tasks and three skill profiles

The distribution of workers' types has three skill profiles $\theta_a < \theta_b < \theta_c$, where $\tan\theta_j = s_{Nj}/s_{Cj}$ is the endowment of workers $j \in \{a, b, c\}$ in skill N relative to skill C, i.e., their comparative advantage in the non-cognitive skill. As shown on Figure 7, profile b is the Generalist profile, while profiles a and c correspond to cognitive specialists and non-cognitive specialists respectively.

We start by constructing an initial balanced equilibrium where the supply and demand is exactly aligned so that there is no bunching, even though the price schedule is linear. To do so, we pick any positive numbers w_C and w_N and construct distributions H^w and H^f for which the linear wage schedule $w(s_C, s_N) = w_C s_C + w_N s_N$ prevails

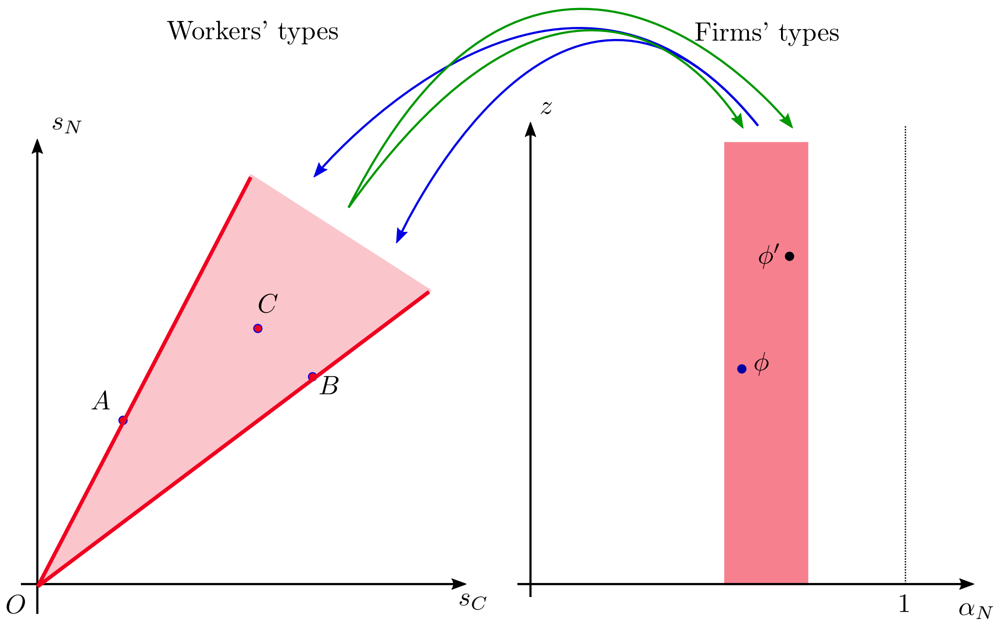


Figure 6: Sorting with bunching: Within-firm heterogeneity in skill profiles

in equilibrium. To achieve that, we choose three values for the technological intensities in skill N, α_{Nj}, $j \in \{a, b, c\}$, such that

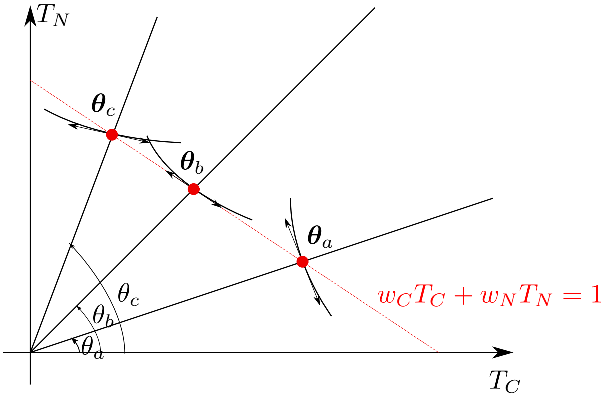
$$\frac{1-\alpha_{Nc}}{\alpha_{Nc}} (\tan\theta_c)^{1-\rho} < \frac{w_C}{w_N} = \frac{1-\alpha_{Nb}}{\alpha_{Nb}} (\tan\theta_b)^{1-\rho} < \frac{1-\alpha_{Na}}{\alpha_{Na}} (\tan\theta_a)^{1-\rho}.$$

Firms with technical intensity α_{Nj} hire workers with profile $\theta_j = \theta^{\mathrm{D}}(\alpha_{Nj})$. Firms with type α_a would prefer workers endowed with more skill C relative to skill N, but no such workers are available in the economy. In this discrete setting, the equilibrium is achieved separately on each half-line $\theta_j\mathbb{R}_+$, $j \in \{a, b, c\}$, separately. In this discrete context, Equation (25) takes the form

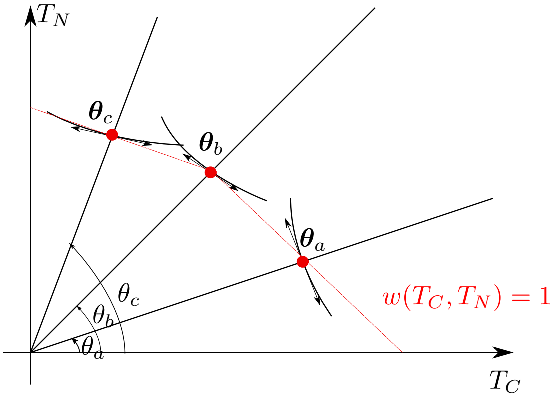
$$\Lambda^{\mathrm{S}}(\theta_j)\, h^w(\theta_j) = Z^f(\alpha_{Nj}) h^f(\alpha_{Nj}) \Lambda^{\mathrm{D}}(\alpha_{Nj}, 1).$$

We choose $\Lambda^{\mathrm{S}}(\theta_j)\, h^w(\theta_j)$ and $Z^f(\alpha_{Nj}) h^f(\alpha_{Nj})$ so that the above equation holds for all $j \in \{a, b, c\}$, i.e. so that Figure 7(a) represents the equilibrium configuration.

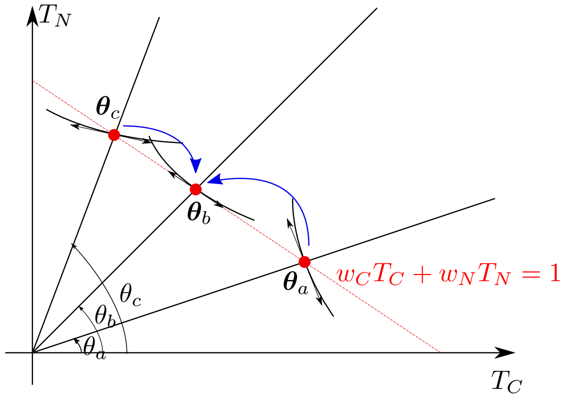
Next, starting from this initial equilibrium, we alter the distribution of skills in the economy in two opposed directions. First, we slightly increase the relative number of generalists in the economy. Specifically, we slightly increase the (quality-adjusted) number of generalist workers , $\Lambda^{\mathrm{S}}(\theta_b)\, h^w(\theta_b)$. To equalize the demand and the supply of



(a) Linear wage schedule



(b) More generalists make the schedule nonlinear



(c) Less generalists and more specialists create bunching

Figure 7: Equilibrium with three relative skill endowments in the economy

generalists, we need to reduce their wage. The equilibrium configuration is modified as shown on Figure 7(b). The wages of the two specialist types a and c remain unchanged, as well as the behavior of α_{Na}-firms and α_{Nc}-firms. The wage schedule, however, becomes non-linear, with generalist workers being subject to a wage markdown.

Second, we decrease the relative number of generalists, again starting from the initial equilibrium of Figure 7(a), and show that bunching emerges. Specifically, we reduce $\Lambda^{\mathrm{S}}(\theta_b)\, h^w(\theta_b)$ by $\nu_b > 0$ and simultaneously increase the number of specialists $\Lambda^{\mathrm{S}}(\theta_a)\, h^w(\theta_a)$ and $\Lambda^{\mathrm{S}}(\theta_c)\, h^w(\theta_c)$ by ν_a and ν_c respectively, where $\nu_a > 0$ and $\nu_c > 0$ are such that $\nu_b \boldsymbol{\theta}_b = \nu_a \boldsymbol{\theta}_a + \nu_c \boldsymbol{\theta}_c$. Figure 7(c) shows the new equilibrium configuration. Firms with technical intensities α_{Na} and α_{Nc} do not change their behavior. Firms with intensity α_{Nb} keep the same aggregate skill $\mathbf{T}^{\mathrm{D}}(\phi)$ but obtain this using a different composition of workers. They hire all workers with relative skill endowment θ_b, but also some workers of type θ_a and θ_c workers, specifically ν_a and ν_c efficiency units, respectively. Hence in equilibrium α_{Na}-firms and α_{Nb}-firms both hire some θ_a workers, while α_{Nb}-firms and α_{Nc}-firms both hire some θ_c workers. In the extreme case where

$\nu_b = \Lambda^{\mathrm{S}}(\theta_b)\, h^w(\theta_b)$, there are no more θ_b-workers in the economy, and firms with intensity α_{Nb} achieve their optimal aggregate skill θ_b by mixing θ_a and θ_c workers.

The above example illustrates our use of the term *bunching*. Because there is always perfect separation in terms of the firm's aggregate skill mix – θ^{D} always increases with α_N – there is no bunching of the sort studied in goods consumption: here there is full sorting. By contrast in our setting, firms with different skill intensities (i.e. different α_N's) display "bunching" at equilibrium when hiring identical workers to "construct" their optimal mix of skills, $\boldsymbol{\theta}^{\mathrm{D}}(\alpha_N)$.

Some Comparative Statics: We now extend the previous 3-types example to general environments with continuous or discrete distribution of types. We want to understand how equilibrium outcomes change when the distribution of skills in the economy becomes more "specialist" (richer in Specialists). To do so, we consider two distributions of workers' skills H_0^w and H_1^w such that H_1^w is more "specialist" than H_0^w.

Lemma 3. *If $H_0^w \leq_{\mathrm{phc}} H_1^w$, the equilibrium output is larger in an economy with for a more "specialist" distribution:*

$$\mathcal{J}^b(H^f, H_0^w) \leq \mathcal{J}^b(H^f, H_1^w), \tag{26}$$

where the output is given by (4).

In the limiting case where under H_1^w all workers in the economy are pure specialists, i.e., $H_1^w = \sum_j \left(\int s_j H_0^w(\mathrm{d}s)\right) \delta_{e_j}$, the bundling friction disappears and we get full efficiency. Pure specialist workers endowed with λ efficiency units of skill j are paid $w(\lambda e_j) = \lambda p_j$, with $p_j = w(e_j)$ being the price of skill j. The wage schedule is fully linear: $w(\mathbf{s}) = \sum_j p_j s_j$. All firms get their preferred skill profile $\boldsymbol{\theta}^{\mathrm{D}}$ by mixing pure specialists and the first-order conditions (7) are simply

$$F_j(\mathbf{T}^{\mathrm{D}}(\phi); \phi) = p_j.$$

In other words, the law of one price prevails, and hence the marginal productivities of skills are equalized across firms.

The next proposition shows that the wage schedule gets flatter as we introduce more specialist workers (relative to generalists) in the economy and get from H_0^w to the corresponding pure specialist distribution $H_1^w = \sum_j \left(\int s_j H_0^w(\mathrm{d}s)\right) \delta_{e_j}$. To move continuously move from H_0^w to H_1^w, we define, for $0 \leq \tau \leq 1$,

$$H_\tau^w = (1-\tau) H_0^w + \tau H_1^w,$$

which increases with τ in the sense of the positively homogenous convex order.

Proposition 6. *As the supply of pure specialist workers in the economy increases, the weighted sum of the generalist markdown*

$$\int [m_\tau(t) - 1]\, w_\tau(t) H_0^w(\mathrm{d}t)$$

decreases with τ*.*

In Appendix E.1, we numerically illustrate how the relative prevalence of specialist workers affects the generalist markdown. We contrast two polar cases where the skill supply comprises essentially generalist workers (scenario A) or essentially specialist workers (scenario B). In Scenario A, the wage schedule is nonlinear, there is a positive markdown, and specialist firms have limited size because specialist workers are expensive. In Scenario B, the wage schedule is linear (zero markdown), the law of one price prevails, specialist firms have access to skills that are more aligned with their technologies and are larger in size compared to Scenario A.[26]

An empirical illustration: To empirically illustrate these results, we define markets as combination of a location and an occupation. Next we split these markets according to the way Generalist firms achieve a balanced set of skills on those markets. In the first group, we focus on markets where Generalist firms are more likely to employ a mix of specialized workers, indicating that Generalists are relatively scarce, leading to more frequent Bunching. In a the second group, we focus on markets where Generalist firms instead primarily employ Generalist workers, i.e workers who each has a balanced set of skills indicating that the supply of Generalists is sufficient to satisfy the demand without bunching. We then estimate the wage demand separately for these two types of markets, see the first two columns of Table 2, with table notes for details. Apart from the sample split, the estimation strategy follows Table 1 closely, using all Specialized firms to estimate counterfactual skill prices.

Consistent with the theory, we find that the Markdown is lower on those markets where Generalist firms are more likely to employ a mix of specialists, i.e. where there is more "Bunching" and thus where the wage schedule is more likely to be linear. By contrast, the markdown is larger on those markets where Generalist firms mostly employ Generalist workers.

[26] When proving Proposition 6 in the Appendix, we show more generally that the wage schedule flattens (i.e., becomes less convex) when moving towards a more specialist distribution (not necessarily the *pure* specialist distribution examined above).

Table 2: Markdown Regression Results by Market Bunching and Productivity

	Sample Variations			
	Bunching	No Bunching	Low Productivity	High Productivity
Relative Markdown	1.16	1.42	1.50	1.07
Absolute Markdown	0.005**	0.010***	0.012***	0.002
	(0.002)	(0.002)	(0.002)	(0.002)
Combined Returns				
...in Generalist Firms	0.030***	0.025***	0.023***	0.032***
	(0.001)	(0.001)	(0.001)	(0.001)
...in Specialist Firms	0.035***	0.036***	0.035***	0.034***
	(0.002)	(0.001)	(0.002)	(0.002)
Observations	2,732,560	1,616,917	1,970,144	1,801,351
Firm Fixed Effects	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes
Occupation by Year	Yes	Yes	Yes	Yes
Firm by Year	No	No	No	No
Occupation-Firm-Year	No	No	No	No
Sample	Bunching	No Bunching	Low Productivity	High Productivity

Notes: Data for 2002–2013. Including firm-years with complete data for at least 25% of workers and at least 10 observations. Test cohorts 1970–1996. Skills measured on a 1–9 Stanine scale. Threshold for generalist firms: < 0.5 difference in average skills. Outcome is log hourly wage. No Bunching markets uses only generalist firms in the top quartile of markets, where markets (defined by the intersection of 3-digit ISCO occupation and NUTS-3 region) are ranked by the share of employees with a gap of at most one Stanine step between cognitive and non-cognitive skills within generalist firms. Firm productivity is split at the firm-level median. Standard errors in parentheses. * $p< 0.1$, ** $p< 0.05$, *** $p< 0.01$.

5 From Workers-to-Firms Sorting to Wage Inequality

In this Section, we first show that, in competitive labor markets with bundling, the log wages can be decomposed into an individual quality component and a "worker-to-firm sorting effect". Then, worker-to-firm sorting effects and wage markdowns are shown to depend on the employing firm's total factor productivity when the production function is non-homothetic. Finally, we examine how our firm-centered framework differs from the classic job-centered matching model and as well as usual versions of the Roy Model, and explain how our framework generates both between- and within-firms wage inequality.

5.1 Workers-to-Firms Sorting Effects

To discuss how our approach relates to the matched employer-employee data literature, see Abowd, Kramarz, and Margolis (1999) and Kline (2024), we start from the generalist markdown defined in (16), $m(\boldsymbol{\theta}) = \lambda / w(\mathbf{s})$. This markdown can be rewritten in log-form as

$$\ln w(\mathbf{s}) = \ln \lambda - \ln m(\boldsymbol{\theta}), \tag{27}$$

where $\boldsymbol{\theta}$ is the worker' individual skill profile and λ is her overall quality, i.e., the sum of her skills valued using the specialists' prices.

In the absence of bunching (see Subsection 4.2), pure sorting in the horizontal dimension implies that the individual worker's profile $\boldsymbol{\theta}$ coincides with her employing firm's aggregate profile $\boldsymbol{\theta}^{\mathrm{D}}(\alpha_N, z)$. Hence, equation

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{D}}(\alpha_N, z) \tag{28}$$

characterizes the workers-to-firm sorting where multiple workers sharing the same $\boldsymbol{\theta}$ are employed in the same firm (α_N, z). The generalist log markdown, $\ln m(\boldsymbol{\theta}^{\mathrm{D}}(\alpha, z))$, is thus (the opposite of) a workers-to-firms sorting effect.

By contrast, in the presence of bunching (recall Subsection 4.3), workers employed at the same firm may have different skill profiles $\boldsymbol{\theta}$s and equation (28) does not hold. As a result, the type of the employing firm no longer appears in the markdown equation (27). However, when bunching is limited, within-firm heterogeneity in profiles is restricted. In this case, individual profiles $\boldsymbol{\theta}$s remain close to the firm-aggregated profiles $\boldsymbol{\theta}^{\mathrm{D}}(\alpha, z)$, and hence the equation

$$\ln w(\mathbf{s}) \approx \ln \lambda - \ln m(\boldsymbol{\theta}^{\mathrm{D}}(\alpha, z)), \tag{29}$$

where (α, z) is the type of a firm that employs workers with skill vector $\mathbf{s}$, remains approximately true.[27]

The Nature of Workers-to-Firms Sorting Effects:

Property (27) is reminiscent of the additive decomposition of the log-wage into a person and a firm effect, contained in Abowd, Kramarz, and Margolis (1999) (AKM, hereafter). The first component of (27), captured by worker's quality (skills evaluated using the equilibrium prices of specialists), does not need much explanation. To understand what is captured within the second component requires a precise discussion. In the pure sorting world (no bunching), multiple workers $\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{D}}(\alpha_N, z)$, with potentially different qualities but an identical profile, enter into *one firm* (α_N, z). Hence, our " Workers-to-Firms Sorting Effect" $-\ln m(\boldsymbol{\theta}^{\mathrm{D}}(\alpha, z))$ is firm-specific and does not reflect a "match effect" where the person and the firm are interacted. Can we then talk about a "Firm-Effect"? A Firm-Effect is a component of pay that workers lose when leaving a firm and gain when entering another one. Identification of the Firm-Effects (or AKM firm-effects) can thus only derive from workers' firm-to-firm mobility. As was made clear from the outset, our framework does not generate any mobility as workers' sorting across firms is immediate and permanent. As a consequence, our Workers-to-Firm Sorting Effects should *not* be thought of as Firm-Effects. Indeed, as we have demonstrated above, markdowns – hence workers-to-firm sorting effects – are indeed identified in the cross-section *without firm-to-firm mobility* within our framework.

As explained by Kline (2024) pp.156-158, "any cross-sectional distribution of wages could be driven by worker sorting rather than firm heterogeneity". This observation suggests different empirical strategies to connect our model to estimated firm wage effects. In particular, a sequence of successive static equilibria with changing supply and/or demand conditions will alter the sorting patterns and thus generate firm-to-firm worker mobility (see Appendix C), potentially allowing us to identify firm wage effects.

5.2 When the Production Function is Non-Homothetic

In the following, we examine the impact of non-homotheticity on the aggregate sorting map $\boldsymbol{\theta}^{\mathrm{D}}(\alpha_N, z)$, on the generalist markdown $m(\theta)$, and on the Workers-to-Firms Sorting

[27]Notice that the markdown $m(\theta)$ generally varies over linear parts of the wage schedule, such as the facet (AB) on Figure 5. Denote by w_i^{AB} the constant implicit price of skill i on the facet. The wage $\sum_i w_i^{AB} s_i$ is linear over the facet. Then the markdown

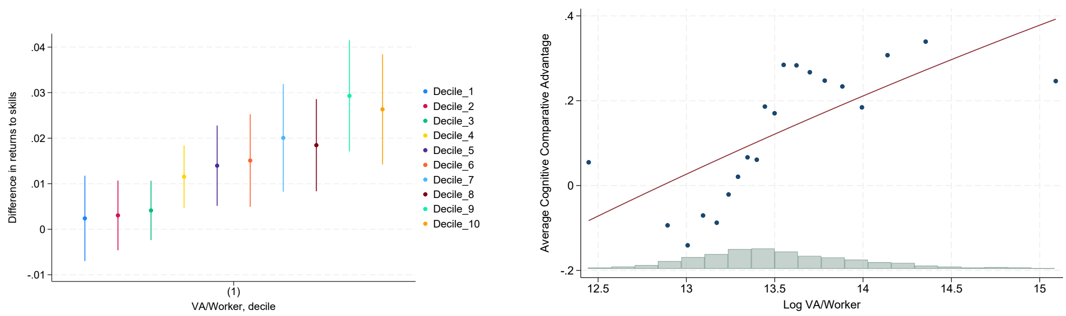
$$m(\mathbf{s}) = \frac{\sum_i w(e_i) s_i}{\sum_i w_i^{AB} s_i}$$

is constant on the facet only if the specialist prices $w(e_i)$ and the implicit prices on the facet w_i^{AB} are proportional.

effects $-\ln m(\theta)$.[28] We have seen that the firm-aggregated skill vector $\mathbf{T}^{\mathrm{D}}$ can be written as $\Lambda^{\mathrm{D}}\boldsymbol{\theta}^{\mathrm{D}}$, where Λ^{D} is the firms' size and $\boldsymbol{\theta}^{\mathrm{D}}$ is its aggregate profile. With two skills, the latter can be represented as $\theta^{\mathrm{D}}(\phi) = \arctan \mathbf{T}^{\mathrm{D}}_N/\mathbf{T}^{\mathrm{D}}_C$. We study how the sorting θ^{D} and the markdown $m(\theta^{\mathrm{D}})$ depend on total factor productivity z.

Lemma 4. *If the production function F is homothetic, the aggregate skill profile θ^{D} does not depend on z. If on the contrary the marginal rate of technical substitution $F_C/F_N(T;\phi)$ increases with the firms' size Λ, the aggregate skill profile θ^{D} decreases with z, i.e., more productive firms are more cognitive.*

Non-Homothetic Sorting: Suggestive Evidence: To assess the presence of no-homotheticity, we rely on Lemma 4. Ideally, we would use a measure of firm-level TFP (z) and test if there is an empirical relationship to the skill profiles (θ) of the same firms. Unfortunately, true tfp is unobserved. However, a number of observable variables are known to be related to z, the closest being labor productivity. We therefore illustrate how sorting and returns to cognitive and non-cognitive skills differ across the labor productivity distribution. The results shown in Figure 8(a) and Figure 1(d), interpreted through the lens of Lemma 4, suggest that sorting is indeed non-homothetic - more productive firms rely more on workers with stronger cognitive abilities.[29]



(a) Differential returns in skills F_C/F_N increases with z

(b) Firm-averaged comparative advantage and labor productivity

Figure 8: Sorting is non-homothetic

Generalists' Markdowns with Non-Homothetic Production Functions: We now examine the generalist markdown $m(\boldsymbol{\theta}^{\mathrm{D}}(\phi))$ associated with the average skill profile of a type-ϕ firm. We want to understand how the firm's markdown depends on its total factor productivity z. We are particularly interested in the non-homothetic case where the productivity ratio F_C/F_N depends on the firm's size.

[28] Recall that in the absence of bunching, $\theta = \theta^{\mathrm{D}}(\alpha_N, z)$, see (28).

[29] We have verified that we get similar results if we relate sorting and skill prices to other (theory-consistent) correlates with z, most notably number workers and the log-level of production.

Hereafter, we say that a firm of type ϕ is cognitive if its aggregate skill profile $\boldsymbol{\theta}^{\mathrm{D}}(\phi)$ is below that of the most generalist profile, $\theta_G = \arctan \boldsymbol{\theta}_G$, and that the firm is non-cognitive otherwise, see Figure 4.[30] Let Φ_C (resp. Φ_N) be the set of types of cognitive (resp. non-cognitive) firms.

Proposition 7. *Suppose that for all firms ϕ the marginal rate of technical substitution $F_C/F_N(T;\phi)$ increases with the firms' size Λ. Then, for cognitive firms $\phi \in \Phi_C$ the generalist markdown $m(\boldsymbol{\theta}^{\mathrm{D}}(\phi))$ decreases with the firm's total factor productivity z, while it increases with z for non-cognitive firms.*

As a general rule, the variations of generalist markdowns with TFP depends on fine details of the economy (technologies and skills distributions). This is confirmed by the numerical simulations presented in Appendix E.2. In three different configurations, we represent the opposite of the log markdown $-\ln m(\theta^{\mathrm{D}}(\alpha_N, z))$ or "worker-to-firm sorting effect" as a function of the technological parameter α_N for two different values of the TFP z. This quantity, which indicates how well generalist workers are paid compared to the equivalent mix of specialists, can increase or decrease with z, or even be non-monotonic in z, see Figure E.5.

An empirical illustration: To provide suggestive evidence on this last point, we split the data according to the median of firm-level labor productivity and derive separate markdown estimates. Results are presented in the last two columns of Table 2. The resulting estimates show that the wage markdown tends to be smaller in more productive firms. Hence, more productive generalist firms seem to pay higher wages. The evidence shown on Figure 8 and the proof of Proposition 7 jointly suggest that the workers employed in more productive generalist firms may have more cognitive skill profiles, making them closer to specialist workers and hence displaying a lower generalist markdown (a higher wage).

5.3 Workers-to-Firms Matching vs. Workers-to-Jobs Matching

As this study is among the first to consider workers-to-firms matching, it is important to highlight the differences with the classic workers-to-jobs matching. To describe the latter, we adopt the framework of Lindenlaub (2017) and introduce the classic objective

$$\mathcal{T}_F(H^f, \gamma) = \max_{\pi \in \mathcal{P}(\gamma, H^f)} \int F(s;\phi)\, \pi(\mathrm{d}s, \mathrm{d}\phi), \tag{30}$$

where $\mathcal{P}(\gamma, H^f)$ denotes the set of all couplings between the distributions of workers' skills and firms' characteristics, γ and H^f. Contrary to (4), total output computed

[30] The notions of cognitive and non-cognitive skills are uniquely defined in equilibrium because the wage schedule and the firm-aggregated skill vectors $\mathbf{T}^{\mathrm{D}}(\phi)$ are unique.

according to (30) is the sum of the output produced in each job and is linear in the matching $\pi(\mathrm{d}\mathbf{s}, \mathrm{d}\phi)$ between workers and jobs.

The next proposition formalizes the intuition that workers-to-firms matching offers **more** productive flexibility than workers-to-jobs matching because it allows firms to freely combine workers to produce tasks, hence a more efficient allocation.

Proposition 8. *If n^{D} is an optimal market-clearing assignment, then the distribution of skills induced by the equilibrium firm-aggregated skill vector, $\mathbf{T}^{\mathrm{D}}H^f$, is solution to*

$$\mathcal{J}^b(H^f, H^w) = \int F(\mathbf{T}^{\mathrm{D}}(\phi); \phi) H^f(\mathrm{d}\phi) = \max_{\gamma \leq_{\mathrm{phc}} H^w} \mathcal{T}_F(H^f, \gamma). \tag{31}$$

From Lemma 2, we know that the distribution induced by the equilibrium firm-aggregated skill vector, $\mathbf{T}^{\mathrm{D}}H^f$, is equally or more generalist than the distribution of the workers' skills, H^w. According to Proposition 8, among all such distributions, $\mathbf{T}^{\mathrm{D}}H^f$ maximizes the total output that can be attained without skill aggregation, i.e., through a "workers-to-jobs" matching such as the one studied by Lindenlaub (2017).

Workers-to-jobs matching generically yields a strictly lower optimal output in the economy than workers-to-firms matching. The two notions of matching yield the same output *only if firms and jobs coincide at the optimum*, i.e., if the optimal assignment of workers to firms, $n^{\mathrm{D}}(\mathrm{d}s; \phi)$, is the mass point located at $\mathbf{T}^{\mathrm{D}}(\phi)$. This is equivalent to saying that the positive measures Q^t introduced in Proposition 5 are the mass points at t, see (A.13). This certainly cannot happen when $\mathbf{T}^{\mathrm{D}}H^f$ is more generalist than H^w because in this case the support of Q^t is larger than the half-line $\mathbb{R}^+ t$: firms mix workers with different profiles (bunching). Even when $\mathbf{T}^{\mathrm{D}}H^f$ and H^w are equally generalist (no bunching), firms generally need to employ workers with different qualities to achieve their optimal size, and hence the support of Q^t is not a singleton. Hence, workers-to-firms matching is *more efficient* than workers-to-jobs matching.

5.4 Wage Inequality in an Economy with Bundling

In this Subsection, we assess the sources of wage inequality in our model and then connect our framework with Roy (1951). To this aim, we first detail components of our equilibrium (convex and homogenous) wage schedule w that are essential to measure inequality in this economy. Recall the implicit price of skill i for workers of type $\mathbf{s}$ is defined as $w_i(\mathbf{s}) = \partial w / \partial s_i$. Implicit prices are homogeneous of degree zero, and as such depend on skill profiles $\boldsymbol{\theta} = s/\lambda$ but not on workers' qualities λ. Using Euler's homogenous function theorem and the convexity of wages, we get, for all skill vectors $\mathbf{s}$

and $\mathbf{s}'$

$$w(\mathbf{s}) = \sum_{i=1}^{k} w_i(\mathbf{s})s_i \geq w(\mathbf{s}') + \sum_{i=1}^{k} w_i(\mathbf{s}')(s_i - s_i') = \sum_{i=1}^{k} w_i(\mathbf{s}')s_i. \tag{32}$$

Using the above property, we show in Appendix A.2 that the wage schedule can be represented as:

$$w(\mathbf{s}) = \max_{\phi'} \mathbf{s} \cdot \nabla w(\mathbf{T}^{\mathrm{D}}(\phi')) = s \cdot \nabla w(\mathbf{T}^{\mathrm{D}}(\phi(\mathbf{s}))), \tag{33}$$

where $\phi(\mathbf{s})$ is a type of firm that employs workers with type $\mathbf{s}$. Figure 9 shows that the iso-wage surfaces are the envelopes of their tangents, a familiar property in the non-linear pricing literature (e.g., Wilson, 1993 and Laffont and Martimort, 2009).

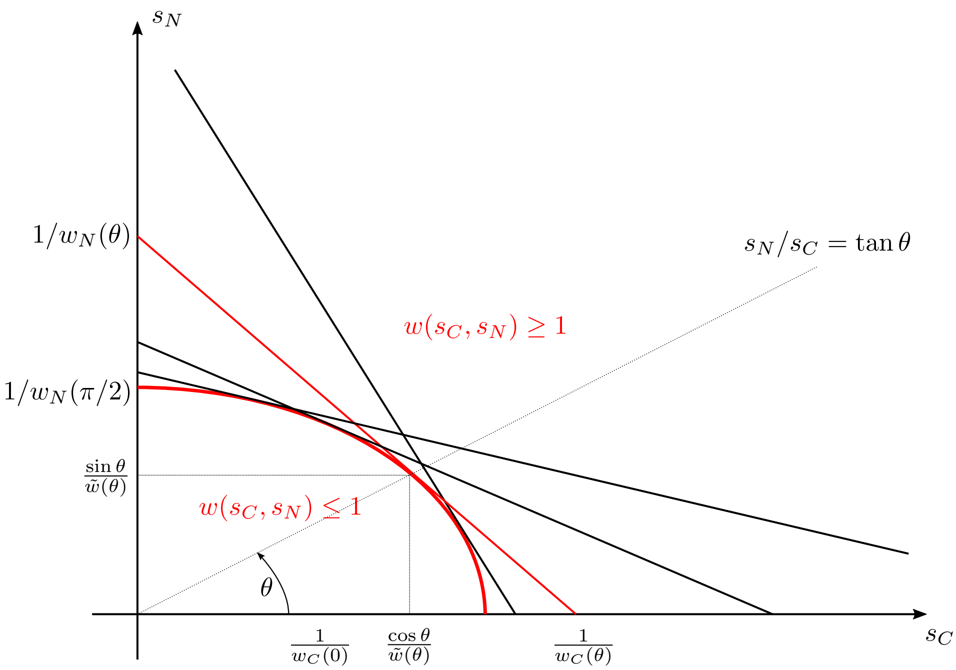


Figure 9: The set of workers paid less than one dollar is convex. The implicit prices of skills C and N for workers with skill profile θ are $w_C(\theta)$ and $w_N(\theta)$

Firms compensate their employees according to a wage schedule that is linear in skills, with the price of each skill reflecting the firm's marginal productivity. Hence, within firms, all wage differences are explained by skill differences. Recall that employees' skills may differ in the quality dimension λ and/or –if bunching prevails– in the skill profile dimension θ. Inequality from price differences occur only between firms, and then result in the generalist markdown presented in Subsection 3.4. In case of bunching, marginal returns of skills are uniform across "bunched" firms, for instance firms hiring workers with profiles in the facet (AB) on Figure 5. In this sense, there is no between-firms inequality across firms belonging to the same bunch (an edge of the wage schedule).

To summarize, because there is always workers-to-firms sorting based on firms' comparative advantage in one skill, there will always be between-firms inequality reflecting aspects of this sorting. If the sorting is pure at equilibrium, because the supply and the demand are equally generalist, there will be some – potentially less – within-firms inequality coming from the dispersion in workers' qualities within the firm, even though all such workers will share the same profile. When bunching emerges, with flat edges in the equilibrium wage schedule, the share of within-firms inequality increases caused by the variety of workers' profiles hired in each (perfectly sorted) firm and between-firms inequality takes place between edges where firms are bunched.

Roy Model: We now connect our contribution to Roy (1951). In such an assignment model, workers decide to self-select into their preferred option among the menu of linear wage schedules $\sum_{i=1}^{k} w_i(s'_i)s_i$ indexed by s' or into their preferred firm. In such a context, Equations (32) would be thought of as an incentive constraint expressing that a worker with skills s prefers the linear schedule "designed for her", i.e., chooses $s' = s$.[31] Similarly, Equation (33) would represent the choice of firms by workers. By contrast, the present paper's modeling framework involves no supply-side decisions on workers' side; the above two equations are purely demand-driven: it results from the structure of our production function, in particular from the aggregation of skills within firms.

The classical Roy literature (see for instance the presentation of Heckman and Honore, 1990) assumes two skills and two sectors, with each sector using only one skill and each skill being priced separately. Workers choosing one sector are paid for the skill valued in that sector, their other skill being left unused. The solid line in Figure 10 shows the wage isoline $w = 1$ for the Roy wage schedule $w(s_C, s_N) = \max(p_C s_C, p_N s_N)$, where p_C and p_N are the prices in the two sectors. This correspond to the special case where all firms use only one of the two skills; in our baseline example, all firms have $(\alpha_C, \alpha_N = 1)$ equal to either $(1, 0)$ or $(0, 1)$. Heckman and Scheinkman (1987), Edmond and Mongey (2022), assumes special forms for the family of linear tariffs. For instance, in the case of two skills, both of these papers assume two sectors with homogenous firms within each sector and a sector-specific wage schedule, in other words they restrict attention to two-part wage schedules. Katz and Murphy (1992) assume the law of one-price, i.e., one price per each separate skill, which corresponds to linear wage isolines.[32] Again here, our approach of the demand side in its full generality (by

[31]Our empirical results, presented above, illustrate that a worker is paid less if she deviates from $s' = s$ and in this sense is not well "matched".

[32]Linearity also obtains under full unbundling. Considering a gradual process of skill unbundling, Choné, Gozlan, and Kramarz (2024) explain how the wage schedule changes as market for stand-alone skills open.

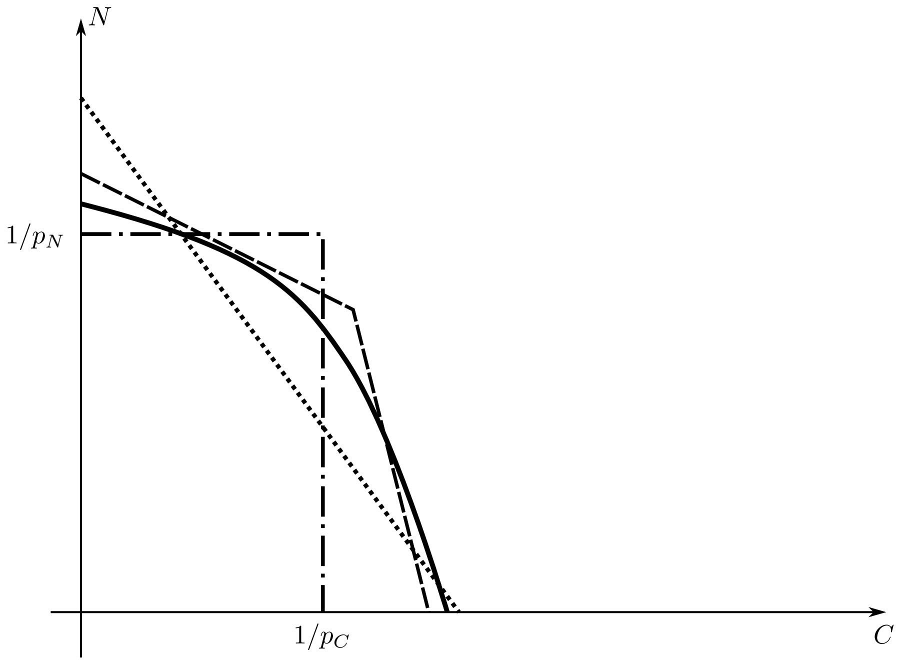


Figure 10: Shapes of isowage lines: Roy (1951): $w(s_C, s_N) = \max(p_C s_C, p_N s_N)$ (dashdotted); Edmond and Mongey (2022) or Heckman and Scheinkman (1987) (dashed); Katz and Murphy (1992) (dotted); continuum of skills (solid)

contrast with the technical restrictions present in the literature listed just above) and no role for a supply side – beyond a bundling constraint –generates very rich results on labor market inequality long thought to be impossible to obtain in a competitive framework.

6 Conclusion

This paper characterizes general equilibrium wage schedules and worker-to-firm sorting patterns in a competitive model while allowing workers to have multidimensional skills and firms to have heterogeneous demand for these skills. We highlight one fundamental supply-side friction, *skill bundling*, which prevents any individual worker from selling her skills to different firms on stand-alone markets. On the demand side, firms produce output by aggregating workers' skills within heterogeneous production functions, rather than aggregating each job's individual output. Importantly, the supply and demand for skill profiles can differ at any local point in the distribution of possible skill profiles.

Our general equilibrium solutions highlight a broad set of consequences arising from the fundamental asymmetry caused by the bundling friction: When local demand for a given mixed skill profile exceeds the supply of this particular profile, firms can rely on a combination of more specialized workers to substitute for the shortage of workers with this exact skill profile. Thus, when local demand exceeds local supply, market wages generate a flat skill-price schedule and an equilbrium with imperfect sorting of workers to firms, a case we refer to as *bunching*. If, on the other hand, the local supply exceeds the local demand for any given mixed skill profile, workers with these skills are unable to split their skills and sell them on more specialized markets. Thus, worker skill profiles are perfectly sorted across firms and the skill-price schedule becomes strictly convex, resulting in a *markdown* whereby workers with a mixed (i.e. general) skill-profile are paid less for their skills than more specialized workers.

We show that wages, in general, have a log-additive structure with a "person" effect, capturing the worker's quality, and a workers-to-firms sorting effect, arising through the interplay of local supply and demand at the relevant market associated with the employing firm's skill demand. Our equilibrium solutions characterize the wages, the sorting of workers to firms, and defines a unique size for each firm as a function of market conditions. Our framework therefore allows for a very rich modeling of firm-side heterogeneity, making it possible to trace out the impact of this heterogeneity on general equilibrium outcomes. We illustrate this by allowing firms to have non-homothetic production functions where firms' demand for skill profiles is a function of their total factor productivity, and show how this affects the general equilibrium properties of the economy.

The paper is inherently theoretical, but throughout the paper, we illustrate the empirical relevance of key assumptions and predictions using data on multidimensional skills linked to data on the productivity of employing firms. We illustrate that workers indeed sort across firms based on multidimensional skill bundles. We highlight the non-random nature of this sorting—firms that employ workers with a unusually strong comparative advantage in cognitive skills in one occupation, also do this in other occupations. Furthermore, we show that we can predict, out-of-sample, the cognitive bias of workers using firm-accounts data that do not include any direct information of worker heterogeneity, thus supporting the assumption that sorting is caused by demand side fundamentals. In relation to our theoretical results, we provide evidence suggesting that returns to skills indeed are lower in firms with a balanced skill-profile (consistent with a markdown), and that this pattern is more prevalent on markets where mix-profile (generalist) firms are less prone to bunching. Finally, we illustrate that sorting and wage patterns are consistent with non-homothetic production functions—more productive firms tend to hire more workers with a cognitive comparative advantage and pay more for these skills.

The model we provide is rich and versatile, but as presented here it is completely static and workers have no incentive to move across firms. However, this is not as limiting as it may appear, as reasonable extensions should generate reasons for firm-to-firm mobility while keeping the general principles of our framework intact. Introducing time-variation into workers' skill profiles, while preserving the total amount of skills in the economy, will generate direct motives for workers to move across firms. Although this is beyond the scope of this article, modelling this kind of time-variation would be an interesting topic for future research, in particular in light of the large empirical literature discussing the wage impact of firm mobility summarized in Kline (2024). Furthermore, because of the rich characterization of firm demand, mobility will also be induced by any shock to firms' production functions (including TFP-shocks, in the non-homothetic case) by causing a mismatch between workers' and firms' skill profiles. We provide a tentative illustration of such an extension in Appendix C, where we allow technological shocks to make cognitive firms more productive, generating firm-to-firm mobility at the cost of losing firm's fixed identity.

The model is also static in terms of the technological environment we postulate. However, the emergence of new markets (platforms) and intermediaries (temp agencies or business-service firms) may have increased the opportunities for workers to effectively sell their skills to multiple firms. Exploring how market equilibria change when markets for stand-alone skills gradually open up requires a different set of mathematical tools, and we therefore present results on this topic in a separate follow-up paper (Choné, Gozlan, and Kramarz, 2024).

Beyond mobility and intermediaries, there are at least three natural directions to expand the research presented in this paper. First, our bundling framework might apply to other literatures, mis-allocation (see Hsieh and Klenow, 2009 and Sraer and Thesmar, 2023) being one prominent example. Indeed, even with competitive benchmarks, input prices may not be uniform across firms or industries when the inputs are bundled. Second, in the same spirit, combining bundling with other frictions or wage setting mechanisms, e.g. using models of Random Search or introducing Nash Bargaining, will allow the framework to speak to a richer set of phenomena. The addition of a bargaining stage could generate an AKM equation – and a Firm-Wage effect – inducing a motive for a profit's pass-through (as surveyed in Kline, 2025). Finally, even though we provide most of our results for a very flexible set of production functions, our skill-aggregation schemes are always assuming conical aggregation of skills where production depends on the sum of skill vectors within firms. We conjecture, however, that it is possible to extend the model to also allow for *non-conical* problems, allowing the analyst to examine an even broader class of production functions.[33] Such production functions would make it possible to incorporate costs and benefits of within-firm diversity (see

[33]Useful existence and duality results pointing in this direction are provided in Choné, Gozlan, and Kramarz (2023).

e.g. Kremer, 1993, Garicano and Rossi-Hansberg, 2006) within a general equilibrium framework.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67(2), 251–333.

Acemoglu, D., and D. Autor (2011): "Skills, tasks and technologies: Implications for employment and earnings," in *Handbook of labor economics*, vol. 4, pp. 1043–1171. Elsevier.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020): "The fall of the labor share and the rise of superstar firms," *The Quarterly Journal of Economics*, 135(2), 645–709.

Berger, D., K. Herkenhoff, and S. Mongey (2022): "Labor Market Power," *American Economic Review*, 112(4), 1147–1193.

Bittarello, L., F. Kramarz, and A. Maitre (forthcoming): "The Task Content of Occupations," *Revue Economique*.

Boerma, J., A. Tsyvinski, and A. Zimin (2021): "Sorting with Team Formation," *Working Paper*.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A distributional framework for matched employer employee data," *Econometrica*, 87(3), 699–739.

Card, D., and J. E. DiNardo (2002): "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles," *The Journal of Labor Economics*, 20(4), 733–783.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024): "On Binscatter," *The American Economic Review*, 114(5), 1488–1514.

Chiappori, P.-A., R. McCann, and B. Pass (2016): "Multidimensional matching," *arXiv preprint arXiv:1604.05771*.

Chiappori, P.-A., R. J. McCann, and L. P. Nesheim (2010): "Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness," *Economic Theory*, 42(2), 317–354.

Choné, P., N. Gozlan, and F. Kramarz (2023): "Weak optimal transport with unnormalized kernels," *SIAM Journal of Mathematical Analysis*, 55(6), 6039–6092.

——— (2024): "Workers-to-Firms Matching When Skills Become Unbundled," *CEPR Discussion Paper*, DP19438.

Choné, P., and F. Kramarz (2022): "Transporting superstars," On-going work.

Costinot, A., and J. Vogel (2010): "Matching and inequality in the world economy," *Journal of Political Economy*, 118(4), 747–786.

Crawford, V. P. (1991): "Comparative statics in matching markets," *Journal of Economic Theory*, 54(2), 389–400.

De Loecker, J., J. Eeckhout, and G. Unger (2020): "The Rise of Market Power and the Macroeconomic Implications," *The Quarterly Journal of Economics*, 135(2), 561–644.

Deb, S., J. Eeckhout, A. Patel, and L. Warren (2024): "Market Power and Wage Inequality," *Econometrica*, 92(3), 603–636.

Edmond, C., and S. Mongey (2022): "Unbundling labor," Discussion paper.

Eeckhout, J., and P. Kircher (2018): "Assortative Matching With Large Firms," *Econometrica*, 86(1), 85–132.

Fredriksson, P., L. Hensvik, and O. N. Skans (2018): "Mismatch of talent: Evidence on match quality, entry wages, and job mobility," *American Economic Review*, 108(11), 3303–38.

Freund, L. B. (2023): "Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities," *Working Paper*.

Galichon, A. (2018): *Optimal transport methods in economics*. Princeton University Press.

Garicano, L., and E. Rossi-Hansberg (2006): "Organization and inequality in a knowledge economy," *The Quarterly journal of economics*, 121(4), 1383–1435.

Gozlan, N., C. Roberto, P.-M. Samson, and P. Tetali (2017): "Kantorovich duality for general transport costs and applications," *Journal of Functional Analysis*, 273(11), 3327–3405.

Haanwinckel, D. (2025): "Supply, demand, institutions, and firms: A theory of labor market sorting and the wage distribution," *American Economic Review*, Forthcoming.

Hatfield, J. W., and P. R. Milgrom (2005): "Matching with contracts," *American Economic Review*, 95(4), 913–935.

Heckman, J., and J. Scheinkman (1987): "The importance of bundling in a Gorman-Lancaster model of earnings," *The Review of Economic Studies*, 54(2), 243–255.

Heckman, J. J., and B. E. Honore (1990): "The empirical content of the Roy model," *Econometrica: Journal of the Econometric Society*, pp. 1121–1149.

Hernnäs, S. (2021): "Automation, Task Content, and Occupations when Skills and Tasks are Bundled," Discussion paper.

Hsieh, C.-T., and P. J. Klenow (2009): "Misallocation and Manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 124(4), 906–938.

Jung, J. W., I. Simonovska, and A. Weinberger (2019): "Exporter heterogeneity and price discrimination: A quantitative view," *Journal of International Economics*, 116, 103–124.

Katz, L. F., and K. M. Murphy (1992): "Changes in relative wages, 1963–1987: supply and demand factors," *The Quarterly Journal of Economics*, 107(1), 35–78.

Kelso, A. S., and V. P. Crawford (1982): "Job matching, coalition formation, and gross substitutes," *Econometrica*, pp. 1483–1504.

Kline, P. (2024): "Firm wage effects," *Handbook of Labor Economics*, 5, 115–181.

Kline, P. M. (2025): "Labor Market Monopsony: Fundamentals and Frontiers," .

Kremer, M. (1993): "The O-ring theory of economic development," *The quarterly journal of economics*, 108(3), 551–575.

Laffont, J.-J., and D. Martimort (2009): *The theory of incentives: the principal-agent model*. Princeton university press.

Lamadon, T., M. Mogstad, and B. Setzler (2022): "Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market," *The American Economic Review*, 112(1), 169–212.

Lashkari, D., A. Bauer, and J. Boussard (2024): "Information technology and returns to scale," *American Economic Review*, 114(6), 1769–1815.

Lemieux, T. (2006): "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?," *American Economic Review*, 96(3), 461–498.

Lindenlaub, I. (2017): "Sorting multidimensional types: Theory and application," *The Review of Economic Studies*, 84(2), 718–789.

Lindqvist, E., and R. Vestman (2011): "The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment," *American Economic Journal: Applied Economics*, 3(1), 101–128.

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

Mood, C., J. O. Jonsson, and E. Bihagen (2012): "Socioeconomic persistence across generations: cognitive and noncognitive processes," in *From parents to children : the intergenerational transmission of advantage*, pp. 53–84.

Ohnsorge, F., and D. Trefler (2007): "Sorting It Out: International Trade with Heterogeneous Workers," *The Journal of Political Economy*, 115(5), 868–892.

Paty, F.-P., P. Choné, and F. Kramarz (2022): "Algorithms for Weak Optimal Transport with an Application to Economics," *arXiv preprint arXiv:2205.09825*.

Pycia, M. (2012): "Stability and preference alignment in matching and coalition formation," *Econometrica*, 80(1), 323–362.

Pycia, M., and M. B. Yenmez (2019): "Matching with externalities," .

Rochet, J.-C., and P. Choné (1998): "Ironing, sweeping, and multidimensional screening," *Econometrica*, pp. 783–826.

Rosen, S. (1974): "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy*, 82(1), 34–55.

Roy, A. D. (1951): "Some thoughts on the distribution of earnings," *Oxford economic papers*, 3(2), 135–146.

Sato, R. (1977): "Homothetic and non-homothetic CES production functions," *The American Economic Review*, 67(4), 559–569.

Simonovska, I. (2015): "Income differences and prices of tradables: Insights from an online retailer," *The Review of Economic Studies*, 82(4), 1612–1656.

Skans, O. N., P. Choné, and F. Kramarz (2023): *The Economics of Creative Destruction.*chap. 9. When Workers' Skills Become Unbundled: Some Empirical Consequences for Sorting and Wages, pp. 253–278. Harvard University Press, Cambridge, Massachussetts and London, England.

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter (2019): "Firming Up Inequality," *The Quarterly Journal of Economics*, 134(1), 1–50.

Sraer, D., and D. Thesmar (2023): "How to Use Natural Experiments to Estimate Misallocation," *The American Economic Review*, 113(4), 906–938.

Takayama, A. (1985): *Mathematical economics.* Cambridge university press.

Teulings, C. N. (2005): "Comparative advantage, relative wages, and the accumulation of human capital," *Journal of Political Economy*, 113(2), 425–461.

Van Reenen, J. (2024): "A Comment on Market Power and Wage Inequality," *Econometrica*, 92(3), 643–646.

Violante, G. L. (2024): "A Comment on Market Power and Wage Inequality," *Econometrica*, 92(3), 637–641.

Wilson, R. B. (1993): *Nonlinear pricing.* Oxford University Press.

Wong, H. C. (2023): "Understanding High-Wage Firms," Discussion paper.

Yeh, C., C. Macaluso, and B. Hershbein (2022): "Monopsony in the US Labor Market," *The American Economic Review*, 112(7), 2099–2138.

A Appendix

A.1 Proof of Proposition 1

The existence of optimal market clearing assignment is based on an extension of Theorem 6.4 of Choné, Gozlan, and Kramarz (2023), see online Appendix (G). The same result also proves the duality formula 20 and the existence of a solution to the dual problem. From this, we prove the existence a competitive equilibrium and the optimality of equilibria.

Existence of a competitive equilibrium Consider an optimal market-clearing assignment n^{D}. Denote by $\mathbf{T}^{\mathrm{D}}(\phi)$ the corresponding firm-aggregated skill vector. Then, for any dual optimizer w, we have by definition of the profit function

$$F(\mathbf{T}^{\mathrm{D}}(\phi);\phi) - \int w(s)\, n^{\mathrm{D}}(\mathrm{d}s;\phi) \leq \bar{\Pi}(\phi;w). \tag{A.1}$$

Integrating with respect to $H^f(\mathrm{d}\phi)$ and using $n^{\mathrm{D}}H^f = H^w$ yields

$$\mathcal{J}^b(H^f,H^w) = \int F(\mathbf{T}^{\mathrm{D}}(\phi);\phi)H^f(\mathrm{d}\phi) \leq \int \bar{\Pi}(\phi;w)H^f(\mathrm{d}\phi) + \int w(s)H^w(\mathrm{d}s) = I^*. \tag{A.2}$$

The equality $\mathcal{J}^b(H^f,H^w) = I^*$ shows that we must have equality in (A.1) for H^f-almost every $\phi \in \Phi$, meaning that the optimal market-clearing assignment n^{D} is decentralized by the wage schedule w.

Optimality of equilibria

On the one hand, there exists a family of nonnegative finite measures $n^{\mathrm{D}}(\mathrm{d}s,\phi)$ satisfying $n^{\mathrm{D}}H^f = H^w$ that achieves the upper bound in (4), hence the existence of an optimal assignment of workers to firms. On the other hand, there exists a convex, 1-homogeneous function w that achieves the lower bound in (20).

First, consider an equilibrium (n^{D},w). Let us denote by T the firm-aggregated skill vector corresponding to the assignment n^{D}, i.e., $\mathbf{T}^{\mathrm{D}}(\phi) = \int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)$. Using the market clearing condition $n^{\mathrm{D}} H^f = H^w$, we have

$$\begin{aligned} I^* &\le \int \bar{\Pi}(\phi; w) H^f(\mathrm{d}\phi) + \int w(s) H^w(\mathrm{d}s) \\ &= \int F(\mathbf{T}^{\mathrm{D}}(\phi); \phi) H^f(\mathrm{d}\phi) - \iint w(s)\, n^{\mathrm{D}}(\mathrm{d}s; \phi) H^f(\mathrm{d}\phi) + \int w(s) H^w(\mathrm{d}s) \\ &= \int F(\mathbf{T}^{\mathrm{D}}(\phi); \phi) H^f(\mathrm{d}\phi) \le \mathcal{J}^b(H^f, H^w). \end{aligned}$$

Because $\mathcal{J}^b(H^f, H^w) = I^*$, there is equality in the first and last inequalities above, implying that w is a dual optimizer (i.e., a solution of (20)) and that the equilibrium assignment n^{D} is optimal.

Qualitative properties of the optimal wage schedule We explain now in more detail why we can restrict attention to convex, 1-homogeneous and non-decreasing wage schedules. The firms' problem (5) can be broken down into two subproblems that consist respectively in finding the firm-aggregated skill vector T and in achieving that aggregate vector in the most economical way. Formally

$$\Pi(\phi; w) = \max_{T \in \mathcal{Z}^s} F(T; \phi) - \bar{w}(T), \tag{A.3}$$

where $\mathcal{Z}^s$ is the conical hull of $\mathcal{X}^s$: $\mathcal{Z}^s = \left\{ \sum_{j=1}^k a_j s_j, a_1, \ldots a_k \in \mathbb{R}_+, s_1, \ldots, s_k \in \mathcal{X}^s \right\}$ and

$$\bar{w}(T) = \inf \left\{ \int w(s)\, n^{\mathrm{D}}(\mathrm{d}s) : n^{\mathrm{D}} \in \mathcal{M}(\mathcal{X}^s), \int s\, n^{\mathrm{D}}(\mathrm{d}s) = T \right\}. \tag{A.4}$$

It is easy to check that the function $\bar{w}$ defined in (A.4) is convex and homogenous of degree one. For any $s \in \mathcal{X}^s$, we can take the $n^{\mathrm{D}}(\mathrm{d}s)$ as the mass point at s, thus showing that $\bar{w}(s) \le w(s)$. The map $\bar{w} : \mathcal{Z}^s \to \mathbb{R}_+$ is therefore the greatest convex and homogenous function such that $\bar{w} \le w$ on $\mathcal{X}^s$.

By construction of $\bar{w}$, we have: $\Pi(\phi; w) = \bar{\Pi}(\phi; \bar{w})$. Moreover, because $\bar{w} \le w$, we have: $\int \bar{w}(x) H^w(\mathrm{d}x) \le \int w(x) H^w(\mathrm{d}x)$. It follows that if w is a dual optimizer, i.e., a solution of Problem (20), so is $\bar{w}$. Using $\bar{w}$ instead of w in (A.1) and (A.2) shows that the optimal market-clearing assignment n^{D} is decentralized by the convex and positively homogenous wage schedule $\bar{w}$.

Finally, we show that at any equilibrium the wage schedule can be assumed to nondecreasing. Starting from any convex, 1-homogeneous wage schedule p, we consider the modified schedule: $\bar{w}(t) = \inf_{t' \ge t} w(t')$. It is easy to see that $\bar{p}$ is convex, 1-homogeneous, and nondecreasing, and that $\bar{w} \le w$. Under schedule w, because the production function $F(T; \phi)$ is nondecreasing in T, firms choose $T' > T$ if T' is less costly than T. So replacing w with $\bar{w}$ does not alter the firms' demand behavior, which implies: $\Pi(\phi;\bar{w}) = \Pi(\phi;w)$. It follows that $\bar{w} \leq w$ is solution to the dual problem (20).

Proof of Lemma 1 Differentiating the generalist markdown (16) with respect to s_i yields

$$\frac{\partial}{\partial s_i} \frac{\sum w(e_i) s_i}{w(\mathbf{s})} = \frac{w(e_i) w(\mathbf{s}) - w_i(\mathbf{s}) \sum w(e_i)}{w(\mathbf{s})^2}$$

which is zero when the marginal returns to skill w_i are proportional to the pure specialist price $w(e_i)$.

A.2 Proof of Proposition 2

Consider two competitive equilibria (n^{D}, w_1) and (n^{D}, w_2). Let $\mathbf{T}_i^{\mathrm{D}} = \int x\, n_i^{\mathrm{D}}(\mathrm{d}x;\phi)$, $i = 1,2$ denote the corresponding firm-aggregated skill vectors. We know from the proof of Proposition 1 that w_1 and w_2 are dual optimizers and that for any dual optimizer w the vectors $\mathbf{T}_1^{\mathrm{D}}(\phi)$ and $\mathbf{T}_2^{\mathrm{D}}(\phi)$ are solutions to Problem (6) for H^f-almost every ϕ. Because F is strictly concave and w is convex, this problem is strictly concave, which yields $\mathbf{T}_1^{\mathrm{D}} = \mathbf{T}_2^{\mathrm{D}} \stackrel{\mathrm{d}}{=} \mathbf{T}^{\mathrm{D}}$. It follows that $\nabla F(\mathbf{T}^{\mathrm{D}};\phi) = \nabla w(\mathbf{T}^{\mathrm{D}}) = \nabla w_1(\mathbf{T}^{\mathrm{D}}) = \nabla w_2(\mathbf{T}^{\mathrm{D}})$. By homogeneity of the wage (Euler's identity), we have $w(\mathbf{T}^{\mathrm{D}}) = \mathbf{T}^{\mathrm{D}} \cdot \nabla w(\mathbf{T}^{\mathrm{D}})$. It follows that $w = w_1 = w_2$ on the image of $\mathbf{T}^{\mathrm{D}}$.

In the absence of bunching (Figure 2), all workers employed by a firm of type ϕ have the same skill profile, i.e., their skill vector is collinear to $\mathbf{T}^{\mathrm{D}}(\phi)$. By homogeneity of the wage schedule, it follows that $w(s) = s.\nabla w(\mathbf{T}^{\mathrm{D}}(\phi))$. In the presence of bunching (Figure 5), the skill vectors s of workers employed by firm ϕ belong to a linear part of the iso-wage line $w = 1$ and are paid the same implicit prices $w_i(\mathbf{T}^{\mathrm{D}}(\phi))$. So $w(s) = s.\nabla w(\mathbf{T}^{\mathrm{D}}(\phi))$ still holds if worker with skill vector s is employed by firm of type ϕ. Because w is convex and 1-homogenous, we have $w(s) \geq w(\mathbf{T}^{\mathrm{D}}(\phi')) + \nabla w(\mathbf{T}^{\mathrm{D}}(\phi')) \cdot (s - \mathbf{T}^{\mathrm{D}}(\phi')) = s \cdot \nabla w(\mathbf{T}^{\mathrm{D}}(\phi'))$ for all firm type ϕ', which yields

$$w(s) = \sup_{\phi'} s \cdot \nabla w(\mathbf{T}^{\mathrm{D}}(\phi')). \tag{A.5}$$

Because $\mathbf{T}^{\mathrm{D}}(\phi)$ and $\nabla w(\mathbf{T}^{\mathrm{D}}(\phi)) = \nabla F(\mathbf{T}^{\mathrm{D}}(\phi))$ are unique, so is the wage schedule w. $\square$

Notice that Proposition 2 holds for the CES production function (11) with complementary skills ($\rho < 0$) although it is not strictly concave (this function is zero if one skill is not used). This is because the maximizer T of $\bar{\Pi}(\phi;w)$ is unique for all non-trivial wage schedules w – the sole property of the production function used above.

A.3 Proof of Proposition 3

When there are two skills C and N, the average profile of the workers can be expressed as the angle $\theta^{\mathrm{D}} = \arctan T_N/T_C$. We change variables

$$\mathbf{T}^{\mathrm{D}} = (\mathbf{T}^{\mathrm{D}}_C, \mathbf{T}^{\mathrm{D}}_N) = \Lambda^{\mathrm{D}}(\boldsymbol{\theta}^{\mathrm{D}}_C, \boldsymbol{\theta}^{\mathrm{D}}_N) = \Lambda^{\mathrm{D}}\left(\frac{1}{p_C + tp_N}\,,\,\frac{t}{p_C + tp_N}\right), \tag{A.6}$$

with $t = \tan\theta^{\mathrm{D}}$. The first-order conditions of the firm's problem can be written

$$\begin{aligned} zF_C(T_C, T_N; \alpha_N) - w_C(T_C, T_N) &= 0 && \text{(A.7)} \\ zF_N(T_C, T_N; \alpha_N) - w_N(T_C, T_N) &= 0, && \text{(A.8)} \end{aligned}$$

Differentiating

$$\begin{pmatrix} F_C \\ F_N \end{pmatrix} + \begin{pmatrix} z\dfrac{\partial F_C}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \Lambda^{\mathrm{D}}} & z\dfrac{\partial F_C}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \theta^{\mathrm{D}}} \\ z\dfrac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \Lambda^{\mathrm{D}}} & z\dfrac{\partial F_N}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \theta^{\mathrm{D}}} \end{pmatrix} \begin{pmatrix} \dfrac{\partial \Lambda^{\mathrm{D}}}{\partial z} \\ \dfrac{\partial \theta^{\mathrm{D}}}{\partial z} \end{pmatrix} = 0$$

Hence

$$\begin{pmatrix} \dfrac{\partial \Lambda^{\mathrm{D}}}{\partial z} \\ \dfrac{\partial \theta^{\mathrm{D}}}{\partial z} \end{pmatrix} = -\frac{1}{d} \begin{pmatrix} z\dfrac{\partial F_N}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \theta^{\mathrm{D}}} & -\left(z\dfrac{\partial F_C}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \theta^{\mathrm{D}}}\right) \\ -\left(z\dfrac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \Lambda^{\mathrm{D}}}\right) & z\dfrac{\partial F_C}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \Lambda^{\mathrm{D}}} \end{pmatrix} \begin{pmatrix} F_C \\ F_N \end{pmatrix},$$

where d is the determinant of

$$\begin{pmatrix} z\dfrac{\partial F_C}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \Lambda^{\mathrm{D}}} & z\dfrac{\partial F_C}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \theta^{\mathrm{D}}} \\ z\dfrac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \Lambda^{\mathrm{D}}} & z\dfrac{\partial F_N}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \theta^{\mathrm{D}}} \end{pmatrix} = \begin{pmatrix} z\dfrac{\partial F_C}{\partial T_C} - \dfrac{\partial w_C}{\partial T_C} & z\dfrac{\partial F_C}{\partial T_N} - \dfrac{\partial w_C}{\partial T_N} \\ z\dfrac{\partial F_N}{\partial T_C} - \dfrac{\partial w_N}{\partial T_C} & z\dfrac{\partial F_N}{\partial T_N} - \dfrac{\partial w_N}{\partial T_N} \end{pmatrix} \begin{pmatrix} \dfrac{\partial T_C}{\partial \Lambda^{\mathrm{D}}} & \dfrac{\partial T_C}{\partial \theta^{\mathrm{D}}} \\ \dfrac{\partial T_N}{\partial \Lambda^{\mathrm{D}}} & \dfrac{\partial T_N}{\partial \theta^{\mathrm{D}}} \end{pmatrix}.$$

By concavity of the firm's problem, the determinant of the first matrix on the right hand side is positive, while simple computations using (A.6) show that the determinant of the second matrix is positive, which yields $d > 0$.

The derivative of total quality with respect to total factor productivity is thus

$$\frac{\partial \Lambda^{\mathrm{D}}}{\partial z} = \frac{1}{d}\left[F_N\left(z\frac{\partial F_C}{\partial \theta^{\mathrm{D}}} - \frac{\partial w_C}{\partial \theta^{\mathrm{D}}}\right) - F_C\left(z\frac{\partial F_N}{\partial \theta^{\mathrm{D}}} - \frac{\partial w_N}{\partial \theta^{\mathrm{D}}}\right)\right].$$

Consider the above bracketed terms. The first term $F_C w'_N - F_N w'_C = w_C w'_N - w_N w'_C$ is positive because the w_N/w_C increases with θ^{D} by concavity if the iso-wage curve. The second term $F_N \partial F_C/\partial\theta^{\mathrm{D}} - F_C \partial F_N/\partial\theta^{\mathrm{D}}$ is positive by convexity of the production isoquants. It follows that the bracketed terms is positive and hence that Λ^{D} increases with z.

Homogenous production functions

Corollary A.1 (Homogenous production functions and TFP)**.** *Assume furthermore that the production functions are homogenous of degree $\eta < 1$. Then the firm-aggregated intermediary input $\mathbf{T}^{\mathrm{D}}(\phi)$, the firm's wage bill, and the firm's profits are proportional to $z^{1/(1-\eta)}$, where z denotes firm's total factor productivity.*

We know that the average skill profile $\boldsymbol{\theta}^{\mathrm{D}}$ does not depend on z. The total quality of a firm ϕ's employees, $\Lambda^{\mathrm{D}}(\phi)$, is determined by maximizing its profit:

$$\Pi(\phi; w) = \max_{\Lambda} z\, F(\Lambda \boldsymbol{\theta}^{\mathrm{D}}(\alpha); \alpha) - \Lambda w(\boldsymbol{\theta}^{\mathrm{D}}(\alpha)). \tag{A.9}$$

Using that F is homogeneous of degree $\eta < 1$, we find that the total quality of workers employed by firm $\phi = (\alpha, z)$:

$$\Lambda^{\mathrm{D}}(\alpha, z) = \left[\frac{\eta\, z\, F(\boldsymbol{\theta}^{\mathrm{D}}(\alpha); \alpha)}{w(\boldsymbol{\theta}^{\mathrm{D}}(\alpha))}\right]^{\frac{1}{1-\eta}} \tag{A.10}$$

The firm's aggregate skill is $\mathbf{T}^{\mathrm{D}}(\phi) = \Lambda^{\mathrm{D}}(\alpha, z)\boldsymbol{\theta}^{\mathrm{D}}(\alpha)$. Using that F is homogenous of degree η, we can write its wage bill as

$$w(\mathbf{T}^{\mathrm{D}}(\phi)) = \Lambda^{\mathrm{D}}(\alpha, z) w(\boldsymbol{\theta}^{\mathrm{D}}(\alpha)) = \left[\eta\, z\, F\left(\frac{\boldsymbol{\theta}^{\mathrm{D}}(\alpha)}{w(\boldsymbol{\theta}^{\mathrm{D}}(\alpha))}; \alpha\right)\right]^{\frac{1}{1-\eta}}. \tag{A.11}$$

The firm's profit is

$$\begin{aligned}
\Pi(\phi; w) &= (1-\eta)\,(z\eta^{\eta})^{\frac{1}{1-\eta}} \left[F\left(\frac{\boldsymbol{\theta}^{\mathrm{D}}(\alpha)}{w(\boldsymbol{\theta}^{\mathrm{D}}(\alpha))}; \alpha\right)\right]^{\frac{1}{1-\eta}} \\
&= (1-\eta)\,(z\eta^{\eta})^{\frac{1}{1-\eta}}\, w(\boldsymbol{\theta}^{\mathrm{D}}(\alpha)) \left[\frac{F\left(\boldsymbol{\theta}^{\mathrm{D}}(\alpha); \alpha\right)}{w(\boldsymbol{\theta}^{\mathrm{D}}(\alpha))}\right]^{\frac{1}{1-\eta}}.
\end{aligned} \tag{A.12}$$

All the above quantities depend on the TFP parameter z through $z^{1/(1-\eta)}$.

A.4 Proof of Proposition 4

To prove the PAM property, we first need to show θ^{D} increases with α_N. Similar computations as above yield

$$\begin{pmatrix} \dfrac{\partial \Lambda^{\mathrm{D}}}{\partial \alpha_N} \\ \dfrac{\partial \theta^{\mathrm{D}}}{\partial \alpha_N} \end{pmatrix} = -\frac{1}{d} \begin{pmatrix} z\dfrac{\partial F_N}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \theta^{\mathrm{D}}} & -\left(z\dfrac{\partial F_C}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \theta^{\mathrm{D}}}\right) \\ -\left(z\dfrac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \Lambda^{\mathrm{D}}}\right) & z\dfrac{\partial F_C}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \Lambda^{\mathrm{D}}} \end{pmatrix} \begin{pmatrix} z\dfrac{\partial F_C}{\partial \alpha_N} \\ z\dfrac{\partial F_N}{\partial \alpha_N} \end{pmatrix},$$

Noticing that w_C and w_N are independent of Λ^{D}, we find that the derivative of the aggregate skill profile with respect to technological intensity is

$$\frac{\partial \theta^{\mathrm{D}}}{\partial \alpha_N} = \frac{z^2}{d} \left[\frac{\partial F_C}{\partial \alpha_N}\frac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - \frac{\partial F_N}{\partial \alpha_N}\frac{\partial F_C}{\partial \Lambda^{\mathrm{D}}}\right].$$

Because $d > 0$, θ^{D} increases with α_N of and only if (10) holds true. If production isoquants are homothetic, we have $F_C \partial F_N/\partial \Lambda^{\mathrm{D}} = F_N \partial F_C/\partial \Lambda^{\mathrm{D}}$ and hence $(\partial F_C/\partial \Lambda^{\mathrm{D}}, \partial F_N/\partial \Lambda^{\mathrm{D}}) = -\kappa(F_C, F_N)$ for some constant $\kappa > 0$, which, together with F_N/F_C increasing in α_N, guarantees that (10) holds.

Finally, we need to check that the determinant of the sorting matrix

$$\begin{pmatrix} \dfrac{\partial \Lambda^{\mathrm{D}}}{\partial z} & \dfrac{\partial \Lambda^{\mathrm{D}}}{\partial \alpha_N} \\ \dfrac{\partial \theta^{\mathrm{D}}}{\partial z} & \dfrac{\partial \theta^{\mathrm{D}}}{\partial \alpha_N} \end{pmatrix} = -\frac{1}{d} \begin{pmatrix} z\dfrac{\partial F_N}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \theta^{\mathrm{D}}} & -\left(z\dfrac{\partial F_C}{\partial \theta^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \theta^{\mathrm{D}}}\right) \\ -\left(z\dfrac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_N}{\partial \Lambda^{\mathrm{D}}}\right) & z\dfrac{\partial F_C}{\partial \Lambda^{\mathrm{D}}} - \dfrac{\partial w_C}{\partial \Lambda^{\mathrm{D}}} \end{pmatrix} \begin{pmatrix} F_C & z\dfrac{\partial F_C}{\partial \alpha_N} \\ F_N & z\dfrac{\partial F_N}{\partial \alpha_N} \end{pmatrix}$$

is positive. We have observed above (when proving that $d > 0$) that the determinant of the first matrix at the left-hand side is positive and the determinant of the second matrix is positive because F_N/F_C increases with α_N, which gives the desired result.

A.5 Proof of Proposition 5

Proof of Lemma 2 Consider a market-clearing assignment n^{D} such that $\mathbf{T}^{\mathrm{D}}(\phi)$ is the firm-aggregated skill vector $\mathbf{T}^{\mathrm{D}}(\phi) = \int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)$. Because any convex and positively 1-homogenous function is sub-additive, we have

$$\begin{aligned} \int h(s) T_{\#} H^f(\mathrm{d}s) &= \int h(\mathbf{T}^{\mathrm{D}}(\phi)) H^f(\mathrm{d}\phi) \\ &= \int h\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)\right) H^f(\mathrm{d}\phi) \\ &\leq \iint h(s)\, n^{\mathrm{D}}(\mathrm{d}s;\phi) H^f(\mathrm{d}\phi) = \int h(s) H^w(\mathrm{d}s), \end{aligned}$$

which proves $T_\# H^f \leq_{\text{phc}} H^w$.

The converse property follows from the new variant of Strassen Theorem established by CGK. Theorem 5.2 in their paper establishes that for any distribution γ more "generalist" than H^w in the sense that $\gamma \leq_{\text{phc}} H^w$, there exists a market-clearing assignment $Q^t(\mathrm{d}s)$ such that $Q\gamma = H^w$ and $y = \int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)$ for γ-almost every y. Applying this result to the distribution $\gamma = T_\# H^f$ yields the desired property, as well as the link between the family of positive measures (Q^t) and the workers-to-firms assignment n^{D}:

$$n^{\mathrm{D}}(\mathrm{d}s;\phi) = Q^{\mathbf{T}^{\mathrm{D}}(\phi)}(\mathrm{d}s). \tag{A.13}$$

□

Proof of Proposition 5 It is easy to see that the functional

$$S(w) = \int \bar{\Pi}(\phi; w) H^f(\mathrm{d}\phi) + \int w(s) H^w(\mathrm{d}s),$$

is convex in w.[34] Let μ denote its differential

$$< \mu, h >= \lim_{\varepsilon \to 0} \frac{S(w + \varepsilon h) - S(w)}{\varepsilon}$$

evaluated on any test function h. By the envelope theorem,

$$< \mu, h >= -\int h(\mathbf{T}^{\mathrm{D}}(\phi)) H^f(\mathrm{d}\phi) + \int h(s) H^w(\mathrm{d}s),$$

where $\mathbf{T}^{\mathrm{D}}(\phi)$ is the (unique) maximizer in Problem (21). In short,

$$\mu = -\mathbf{T}^{\mathrm{D}}_\# H^f + H^w.$$

Because the functional S is convex, a function $w \in \Phi^+$ minimizes S (i.e., is the dual optimizer) if and only if

$$< \mu, h > \geq 0 \tag{A.14}$$

for all h such that $w + h \in \Phi^+$. This implies in particular that the two following conditions:

$$< \mu, h > \geq 0 \text{ for all } h \in \Phi^+ \tag{A.15a}$$

$$< \mu, w > \geq 0. \tag{A.15b}$$

[34]This results from the fact that $\bar{\Pi}(\phi; w)$ is convex in w, for all ϕ.

Conversely, suppose that (A.15a) and (A.15b) hold and take h such that $w + h \in \Phi^+$. We have

$$< \mu, h >=< \mu, w + h > - < \mu, w >=< \mu, w + h > \geq 0,$$

which yields (A.14). So the two conditions (A.15a) and (A.15b) are equivalent to (A.14), and hence together characterize the dual optimizer. Condition (A.15a) can be rewritten as (19).

From the CGK version of Strassen Theorem, the inequality (19) in the sense of the positive homogenous convex ordering is equivalent to the existence of a family of positive measures $(Q^t(\mathrm{d}s))$ such that $t = \int s\, Q^t(\mathrm{d}s)$ for almost all t and $H^w = Q\mathbf{T}^{\mathrm{D}}_{\#}H^f$. Using the latter equality and the sub-additivity of w, $w(t) \leq \int w(s)Q^t(\mathrm{d}s)$, we get

$$\int w(t)\mathbf{T}^{\mathrm{D}}_{\#}H^f(\mathrm{d}t) \leq \int\int w(s)Q^t(\mathrm{d}s)\mathbf{T}^{\mathrm{D}}_{\#}H^f(\mathrm{d}t) = \int w(s)H^w(\mathrm{d}s).$$

Condition (A.15b) states that the above inequality is an equality, which is equivalent to w being linear on the support of Q^t for $\mathbf{T}^{\mathrm{D}}_{\#}H^f$-almost all t. Hence the desired result. □

A.6 Proof of Proposition 6

Proof of Lemma 3 This ordering can be seen in two ways. A first way to prove (26) is to write the objective as

$$\mathcal{J}^b(H^f, H^w) = \max_{\gamma \leq_{\mathrm{phc}} H^w} \max_{\pi \in \mathcal{P}(\gamma, H^f)} \int F(s; \phi)\, \pi(\mathrm{d}s, \mathrm{d}\phi), \tag{A.16}$$

where $\mathcal{P}(\gamma, H^f)$ denotes the set of all couplings between γ and H^f, see CGK. This formulation of the primal problem gives (26). The ordering (26) can also be derived from the dual version of the problem (Equation (20)) by noticing that $\int w_1(s)H^w_0(\mathrm{d}s) \leq \int w_1(s)H^w_1(\mathrm{d}s)$, where w_1 is the dual optimizer for H^w_1. An immediate consequence of (31) is that for any pair of skill distributions (H^w_0, H^w_1) such that $H^w_0 \leq_{\mathrm{phc}} H^w_1$, i.e., such that H^w_1 is "more specialist" than H^w_0, total output is greater for H^w_1 than for H^w_0.[35] □

Proof of Proposition 6 Take two distributions H^w_0 and H^w_1 such that $H^w_0 \leq_{\mathrm{phc}} H^w_1$. By CGK's version of the Strassen Theorem, we know that there exists a family of positive measures $(P^t(\mathrm{d}s))$ such that $t = \int s\, P^t(\mathrm{d}s)$ for almost all t and $H^w_1 = PH^w_0$.

[35] This is because the set $\{\gamma : \gamma \leq_{\mathrm{phc}} H^w_1\}$ is larger than the set $\{\gamma : \gamma \leq_{\mathrm{phc}} H^w_0\}$.

For any $\tau \in [0,1]$, we set $H^w_\tau = (1-\tau)H_0 + \tau H^w_1$ and

$$V(\tau) = \mathcal{J}^b(H^f, H^w_\tau).$$

The function $V(\tau)$ is the lower bound of a family of affine functions of τ and is thus concave in τ. By the envelope Theorem, its derivative is given by

$$V'(\tau) = \int w_\tau(s)(H^w_1 - H^w_0)(\mathrm{d}s),$$

where $w_\tau \in \Phi^+_{\mathrm{nd}}$ is the unique minimizer of the dual problem for the distribution H^w_τ. We can rewrite the derivative as

$$V'(\tau) = \int \left[\int w_\tau(s)P^t(\mathrm{d}s) - w_\tau(t) \right] H^w_0(\mathrm{d}t) \geq 0.$$

Because w_τ is sub-additive, the derivative $V'(\tau)$ is nonnegative, confirming (26). The above quantity is a global indicator of the non-linearity of w_τ. It is zero if and only if w_τ is linear on the support of H^w_0. Because V is concave, we know that this indicator is non-increasing in τ:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \int \left[\int w_\tau(s)P^t(\mathrm{d}s) - w_\tau(t) \right] H^w_0(\mathrm{d}t) \leq 0,$$

for all $\tau \in [0,1]$. An interesting special case is $P^t(\mathrm{d}s) = \sum_{i=1}^k \delta_{t_i e_i}$ and $H^w_1 = \sum \left(\int s_i H^w_0(\mathrm{d}s) \right) \delta_{e_i}$. In this case, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\tau} \int \left[\sum_{i=1}^k t_i w_\tau(e_i) - w_\tau(t) \right] H^w_0(\mathrm{d}t) = \\
\frac{\mathrm{d}}{\mathrm{d}\tau} \int \left[m_\tau(t) - 1 \right] w_\tau(t) H^w_0(\mathrm{d}t) \\
\frac{\mathrm{d}}{\mathrm{d}\tau} \sum_{i=1}^k w_\tau(e_i) \int t_i H^w_0(\mathrm{d}t) - \int w_\tau(t) H^w_0(\mathrm{d}t) \leq 0,
\end{aligned}$$

for all $\tau \in [0,1]$. A stronger result would be that the difference $\int w_\tau(s)P^t(\mathrm{d}s) - w_\tau(t)$ is non-increasing in τ *for each t separately*. This would express that w_τ becomes more linear around t as τ increases from zero to one.

A.7 Proof of Proposition 7

Proof of Lemma 4 From the proof of Proposition 4, we have

$$\frac{\partial \theta^{\mathrm{D}}}{\partial z} = (z/d)\left\{F_C \frac{\partial F_N}{\partial \Lambda^{\mathrm{D}}} - F_N \frac{\partial F_C}{\partial \Lambda^{\mathrm{D}}}\right\}.$$

where $d > 0$. It follows θ^{D} is independent of z when the production isoquants are homothetic and decreases with z if $\partial(F_C/F_N)/\partial\Lambda^{\mathrm{D}} > 0$. □

From Lemma 4, we know that high-z firms use relatively more cognitive skills. If the firm is cognitive ($\phi \in \Phi_C$), its aggregate skill profile gets further away from $\boldsymbol{\theta}_G$ and hence the generalist markdown goes down. By contrast, if the firm is non-cognitive, using more cognitive skills means getting closer to $\boldsymbol{\theta}_G$, hence a higher markdown.

The condition $\partial(F_C/F_N)/\partial\Lambda^{\mathrm{D}} > 0$ holds for instance for the CES function modified in the spirit of Simonovska (2015) and Jung, Simonovska, and Weinberger (2019):[36]

$$zF(T;\alpha_N) = (z/\eta)\ \left[\alpha_C(T_C + \bar{T}_C)^{\rho} + \alpha_N T_N^{\rho}\right]^{\eta/\rho} - K, \tag{A.17}$$

where $\alpha_C + \alpha_N = 1$ and $\bar{T}_C$ is a positive constant and the constant K ensures that $zF(0, \alpha_N) = 0$. Indeed here

$$\frac{F_C}{F_N} = \frac{\alpha_C}{\alpha_N}\left[\frac{T_N}{T_C + \bar{T}_C}\right]^{1-\rho}$$

and hence F_C/F_N evaluated at $\mathbf{T}^{\mathrm{D}} = \Lambda^{\mathrm{D}}\boldsymbol{\theta}^{\mathrm{D}}$ increases with Λ^{D}.

A.8 Proof of Proposition 8

The inequality $\int F(\mathbf{T}^{\mathrm{D}}(\phi);\phi)H^f(\mathrm{d}\phi) \leq \max_{\gamma \leq_{\mathrm{phc}} H^w} \mathcal{T}_F(H^f,\gamma)$ follows immediately from the fact that $\mathbf{T}^{\mathrm{D}}H^f \leq_{\mathrm{phc}} H^w$.

Conversely, consider a distribution $\gamma \leq_{\mathrm{phc}} H^w$ and a coupling π between γ and H^f. Disintegrate the coupling as $\pi = H^f P^{\phi}$. Because $\gamma \leq_{\mathrm{phc}} H^w$, there exists by CGK version of Strassen theorem a family of nonnegative measure R^t such that $\int sR^t(\mathrm{d}s) = t$ and $H^w = R^t\gamma$. Now define

$$n^{\mathrm{D}}(\mathrm{d}s;\phi) = \int_t R^t(\mathrm{d}s)P^{\phi}(\mathrm{d}t)$$

[36] See Sato (1977) for a comprehensive study of non-homothetic CES function.

The family n^{D} is a market-clearing assignment because

$$\int_\phi n^{\mathrm{D}}(\mathrm{d}s,\phi)H^f(\mathrm{d}\phi) = \int_\phi \int_t R^t(\mathrm{d}s)P^\phi(\mathrm{d}t)H^f(\mathrm{d}\phi) = \int_t R^t(\mathrm{d}s)\gamma(\mathrm{d}t) = H^w(\mathrm{d}s).$$

Also notice that

$$\int_s s\, n^{\mathrm{D}}(\mathrm{d}s;\phi) = \int_s \int_t s R^t(\mathrm{d}s)P^\phi(\mathrm{d}t) = \int_t t P^\phi(\mathrm{d}t).$$

By concavity of F, it follows that

$$\begin{aligned} \int F(t;\phi)\pi(\mathrm{d}t,\mathrm{d}\phi) &= \int F(t;\phi)P^\phi(\mathrm{d}t)H^f(\mathrm{d}\phi) \\ &\le \int F\left(\int tP^\phi(\mathrm{d}t)\right)H^f(\mathrm{d}\phi) = \int F\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)\right)H^f(\mathrm{d}\phi) \le \mathcal{J}^b(H^f,H^w) \end{aligned}$$

Maximizing over π and γ shows that $\int F(\mathbf{T}^{\mathrm{D}}(\phi);\phi)H^f(\mathrm{d}\phi) \ge \max_{\gamma \le_{\mathrm{phc}} H^w} \mathcal{T}_F(H^f,\gamma)$

B From Workers' Skills to Firms' Tasks (and Back)

We now propose extensions of our baseline model in two directions. We examine (i) how the tasks performed by workers depend on their skills; (ii) how firms aggregate the tasks performed by their employees.

Link between skills and tasks at the worker level So far, we have equated skills and tasks, i.e., we have assumed $s = t$ for all workers. For now, we restrict attention to skills-to-tasks relationships of the form $t = g(s)$, where the function g is exogenous and occupation-specific.From the above analysis, we know that *in the space of tasks* the "wage schedule" is convex and homogenous of degree one, i.e., $w^t(t_C, t_N)$ is convex and homogenous of degree one in (t_C, t_N). We also know that under the assumptions of Proposition 4, for instance with the production function (12), the firms-to-tasks matching is PAM, in the sense that there exists an increasing relationship between firms' intensities α_N and task profiles t_N/t_C. However, as already mentioned, tasks are unobserved in existing data; only workers' skills are. It is therefore important to know (i) whether the observed wage schedule $w(s) = w^t(g(s))$, i.e., the wage as a function of the observed skills, inherits the properties of w; and (ii) whether the firms-to-skills matching inherits the properties of the firms-to-tasks matching.

To answer this question, we start from the simplest and most intuitive way to characterize the above relationship between skills and tasks and assume that each task uses each of the worker's skills in fixed quantities:

$$(t_C, t_N) = g(s_C, s_N) = (d_{CC}s_C + d_{CN}s_N, d_{NC}s_C + d_{NN}s_N), \tag{B.1}$$

with $d_{ij} \geq 0$. For such a linear relationship $t = Ds$, it is straightforward to check that wages are convex and homogenous in skills. Moreover, if $\det D > 0$ there is an increasing relationship between workers' skill profiles s_N/s_C and task profiles t_N/t_C. The same is true when the skills-to-tasks relationship is homogenous of degree $\gamma > 0$, as in $(t_C, t_N) = g(s_C, s_N) = (s_C^\gamma, s_N^\gamma)$, so that in these two cases firms-to-tasks PAM translates into firms-to-skills PAM. The linear case of (B.1) is not econometrically different from $t = s$ since the skill-to-task function g and the distribution of α cannot be separately identified (tasks are unobserved in existing data). This is not so when g is, for instance, homogenous of degree γ, which would translate into a wage schedule $w(s) = w^t(g(s))$ that is homogenous of the same degree in s, something potentially observable in the data.

Aggregation of workers' tasks within firms One may consider aggregation technologies that are not additively separable in the employees' tasks. An often used aggregation scheme is CES:

$$T = \left(\left[\int t_C^\gamma \, n^{\mathrm{D}}(\mathrm{d}t) \right]^{1/\gamma}, \left[\int t_N^\gamma \, n^{\mathrm{D}}(\mathrm{d}t) \right]^{1/\gamma} \right). \tag{B.2}$$

In our leading example (11) where the production function $F(T_C, T_N)$ is itself CES, such a skills-aggregation scheme leads to a two-level nested CES. For our theoretical results to apply, we need $F(T)$ to be concave in the assignment n^{D}, i.e., we need the modified production function $\tilde{F}(T_C, T_N) = F(T_C^{1/\gamma}, T_N^{1/\gamma})$ to be concave in T, which obtains if $\gamma > \max(\rho, \eta)$.[37] We can thus allow for some degree of imperfect substitution between co-workers' tasks.

Finally, the number of skills needs not be equal to the number of tasks. Suppose two types of cognitive skills and two types of non-cognitive skills are used to produce two tasks according to

$$T = (T_1, T_2) = (\mathrm{CES}(C_1, N_1; \beta_1), \mathrm{CES}(C_2, N_2; \beta_2)).$$

Task 1 uses skills C_1 and N_1, with β_1 representing the (potentially firm-specific) technical intensity in C_1 with an equivalent formulation holding for Task 2. The final output

[37] The new production function $\tilde{F}$ is quasi-concave ($\rho/\gamma < 1$) and homogenous of degree $\eta/\gamma < 1$.

is then produced by combining the two tasks according to $F(T;\alpha)$. As above, the production function can be rewritten as $\tilde{F}(C_1;C_2,N_1,N_2;\alpha,\beta_1,\beta_2)$, where capital letters represent firm-aggregated quantities (for instance $C_1 = \int s_{C1}\, n^{\mathrm{D}}(\mathrm{d}s_{C1})$) and (α,β_1,β_2) is a *firm-specific* vector of technical parameters.

A distinctive feature of the setup examined in the present paper is that firm-specific parameters interact only with firm-aggregated quantities. Using the first-order conditions (7), aggregate sorting properties can be derived as in Proposition 4 noticing that $\tilde{F}_{C_i}/\tilde{F}_{N_i}$ increases with β_i and $\tilde{F}_{T_1}/\tilde{F}_{T_2}$ increases with α. This class of production functions strikingly differs from settings where a firm's set of technical characteristics interact with *individual* workers' characteristics, which pushes to individual rather than aggregate sorting.[38] The above formulation which connects skills and tasks, when compared with Haanwinckel (2025) or Teulings (2005), offers more between-firms heterogeneity or, when compared with Eeckhout and Kircher (2018), possesses a clear within-firm aggregation scheme.

C Technological Shocks under Bundling

In this Appendix, we study the impact of a positive productivity shock that affects firms differently, depending on their technological profile (α).

Skill-Biased Technological Change (SBTC) was seen by many as an essential driving force of labor markets transformations (wage inequality, in particular) over the 80s and the 90s, in the context of an increased supply of educated workers in the US.[39] We study this eighties/nineties style environment faced with such technical change in the following.

Workers supply their skills as a bundle, the quantity being fixed, but are faced with a SBTC shock affecting the productivity of cognitive tasks in their employing firms, in a labor market where cognitive skills are in (relative) short supply before the shock. Because of this SBTC shock, with or without an education shock, at the new equilibrium, prices (wages of each skill) change, hence the sorting of workers to firms changes, as well as the size of each firm. Following insights from Simonovska (2015),

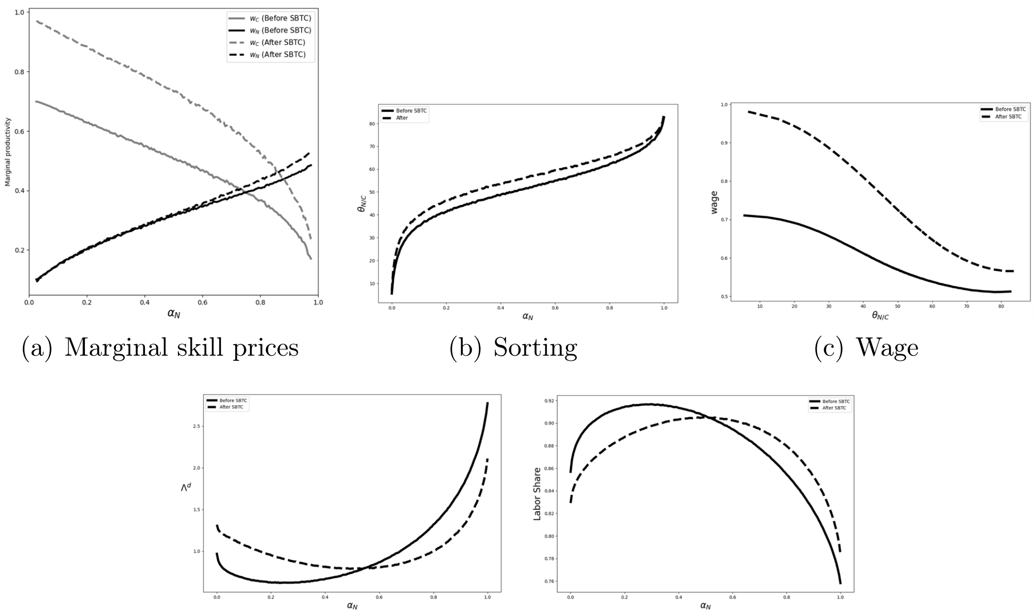
[38] Choné and Kramarz (2022) considers the case where tasks are produced by interacting individual firm and worker characteristics and are then aggregated within each firm.

[39] See Katz and Murphy (1992) is an important view on SBTC, see Card and DiNardo (2002) for a thorough discussion, but Lemieux (2006) shows the role of the secular increase in education that took place there in explaining increasing inequality without a large role for SBTC. However, Bittarello, Kramarz, and Maitre (forthcoming) show that, in France, returns to cognitive skills decreased between 1990 and 2013 *because* of the increase in the supply of college-educated workers.

the production function we use to study SBTC is non-homogenous :

$$F(T;\alpha,z) = \frac{z}{\eta}\left[\left((\alpha_N T_N^{\rho} + \alpha_C T_C^{\rho})^{1/\rho} + A\right)^{\eta} - A^{\eta}\right], \tag{C.1}$$

with $A = 1$. The two skills are complements ($\rho = -1$) and returns to scale are decreasing ($\eta = .5$). We assume that the technical parameter α_N is uniformly distributed on $[0,1]$. In this Section, cognitive skills are relatively rare: workers' skill profiles $\theta = \arctan s^N/s^C$ are distributed as a Beta(7,4) random variable on $(0,\pi/2)$. There is no heterogeneity in workers' quality ($\lambda = 1$) or in firms' total factor productivity ($z = 1$). In the following numerical simulation, we use the mirror-descent algorithms developed in Paty, Choné, and Kramarz (2022).



(a) Marginal skill prices (b) Sorting (c) Wage

(d) Firm's sizes (e) Labor share

Figure C.1: Skill-biased technical change. Cognitive firms experience a positive productivity shock

We now consider a positive productivity shock that affects cognitive firms, starting from the environment described above. The firms' total factor productivity parameter z increases from $z = 1$ to $z = 1 + (1 - \alpha_N)/2$. Purely non-cognitive firms, with $\alpha_N = 1$, are unaffected. Purely cognitive firms ($\alpha_N = 0$) have $z = 1.5$ after the shock.

The first-order effect of this SBTC shock is, obviously, an increase in the price of cognitive skills paid by all firms, in particular those most in demand for such skills (left side of Figure D.1(a)). As the price of cognitive tasks increase, firms use and demand more non-cognitive tasks, affecting sorting (Figure D.1(b)). This induced increase in demand for non-cognitive tasks in turn impacts the price of non-cognitive skills (see Figure D.1(a), right part). Wages are an outcome of both sorting and skill prices.

Indeed, Figure D.1(c) shows that following the SBTC all wages – including those of non-cognitive workers – increase because firms not only use and demand more cognitive tasks but also non-cognitive ones. Importantly, our SBTC shock increases wage inequality. Cognitive firms increase in size, as shown in Figure D.1(d) even though the price of cognitive skills increases because the productivity (direct) effect of the increases in the TFP parameter, z, dominates and outweighs the price effect. Similarly, the labor share, defined as the payroll-to-sales ratio wL/F, decreases for the most cognitive firms (Figure D.1(e)) even though the corresponding wage increases: essentially because firms produce more (are more efficient), i.e., the direct effect of the increased productivity, z, which is included within the denominator of the labor share – wL/F – dominates.

As we just saw, SBTC affects the sorting of workers to firms. Hence, to accomodate this changing sorting, workers will move from firm-to-firm. More precisely, workers with a skill profile θ will move from firms with $\alpha_{before}^{-1}(\theta)$ to firms with $\alpha_{after}^{-1}(\theta)$. Adding mobility costs will dampen the associated magnitude.[40]

A long line of research, briefly mentioned above, tries to identify how SBTC affects labor market outcomes, resulting potentially in increased wage inequality, in the context of a growing supply of college-educated workers, as happened in the US over the 80s or in France over the 90s and 2000s. In most studies, such as Katz and Murphy (1992) or Lemieux (2006), the potential role of skills-bundling was never evoked. Recently, Lindenlaub (2017) filled this gap by studying how bundling alters previous conclusions. Indeed, and fully in line with what we do here, Lindenlaub (2017) focuses on Task-Biased Technological Change (TBTC, using her words) within a Quadratic-Gaussian model with bi-linear production function, delivering a closed-form solution to her problem. Whereas we increase the productivity of firms having a "preference" for cognitive tasks, she models TBTC by decreasing the relative efficiency of the manual task w.r.t. the cognitive task. In this environment, she presents sufficient conditions for wage inequality to increase. Her work is important for us because it shows how bundling (for jobs in her approach) has implications that widely differ from those obtained in a fully unbundled world. Our analysis differs from hers because we define SBTC (or TBTC) in a world with firms (rather than jobs). Indeed, we are able to predict changes in firms' sizes and labor shares after SBTC.

We could also contrast SBTC without and with an increase in the supply of college-educated workers (as took place in the US or in France, as examined in Bittarello, Kramarz, and Maitre, forthcoming), increasing as a result the supply of cognitive skills. Hence, in comparison with our previous Figures, the price of cognitive skills will increase

[40]Mobility is likely to be too large without such costs when compared to year-to-year realized moves seen in data. But estimated costs in structural analyses tend to be unrealistically large too.

less. As a consequence, all the effects described will still exist but will be attenuated, potentially resulting in an absence of impact on wage inequality (as suggested in Lemieux, 2006).[41]

D A Dixit-Stiglitz Environment

In the paper, we assume that the price of the final good is exogenous and normalized to one, and quantities (output, labor demand, etc.) are determined by decreasing returns to scale. We now present a Dixit-Stiglitz environment where firms operate under constant returns to scale and quantities are set by monopolistic competition.[42] We then check that this environment is isomorphic to the model presented in the main text.

For simplicity of exposition, we assume in the following that the workers' skills are two-dimensional. Firms indexed by (α, z) produce a differentiated good under constant returns to scale. The production function takes the form $y(\alpha_N, z) = zF(T_C, T_N; \alpha_N)$, where F is homogenous of degree one. A representative consumer has income I and preferences over baskets $\mathbf{y} = (y(\alpha_N, z))$ given by

$$U(\mathbf{y}) = \left(\int y(\alpha_N, z)^{\frac{\sigma-1}{\sigma}} H^f(\mathrm{d}\alpha_N, \mathrm{d}z) \right)^{\frac{\sigma}{\sigma-1}},$$

with $\sigma > 1$. Let $p(\alpha_N, z)$ denote the price of good (α_N, z). The Marshallian demand is given by

$$y(\alpha_N, z) = I \frac{p(\alpha_N, z)^{-\sigma}}{\int [p(\alpha_N, z)]^{1-\sigma} H^f(\mathrm{d}\alpha_N, \mathrm{d}z)}. \tag{D.1}$$

As in the text, we parameterize aggregate skill vectors as $\mathbf{T}^{\mathrm{D}} = (T_C, T_N) = \Lambda^{\mathrm{D}} \boldsymbol{\theta}^{\mathrm{D}}$, where Λ^{D} is the aggregate quality of its employees (our notion of firm's size) defined in (8) and $\boldsymbol{\theta}^{\mathrm{D}}$ is the firm-aggregated skill profile defined by (9). The qualitative properties of the wage schedule $w(s_C, s_N)$, which come from the linear aggregation of skills within firms, are the same as in our baseline environment, namely convexity and one-homogeneity. There is monopolistic competition on the product market and firm (α_N, z) chooses its

[41]It is hard, not to say impossible, to go beyond this limited observation without structurally estimating our model, an endeavor beyond the purpose of the present article.

[42]See for instance Costinot and Vogel (2010) with one-dimensional skills.

aggregate skill vector $\mathbf{T}^{\mathrm{D}} = \int s\, n^{\mathrm{D}}(\mathrm{d}s)$ to maximize its profit

$$\begin{aligned} p(\alpha_N, z)y - w(\mathbf{T}^{\mathrm{D}}) &= y\left[p(\alpha_N, z) - \frac{w(\mathbf{T}^{\mathrm{D}})}{y}\right] \\ &= y\left[p(\alpha_N, z) - \frac{w(\boldsymbol{\theta}^{\mathrm{D}})}{z\; F(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N)}\right]. \end{aligned} \tag{D.2}$$

(To derive the last equality above, we use the one-homogeneity of w and F and eliminate Λ^d.) For any aggregate skill profile $\boldsymbol{\theta}^{\mathrm{D}}$, the firm chooses its aggregate worker quality Λ^{D} (or equivalently its output y) under the demand equation, which yields the standard mark-up condition

$$p(\alpha_N, z) = c(\theta; \alpha_N, z)\; \frac{\sigma}{\sigma - 1}, \tag{D.3}$$

and the output

$$y(\alpha_N, z) = I\; \frac{\sigma - 1}{\sigma}\; \frac{c(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N, z)^{-\sigma}}{\int c(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N, z)^{1-\sigma} H^f(\mathrm{d}\alpha_N, \mathrm{d}z)}, \tag{D.4}$$

where

$$c(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N, z) = \frac{w(\boldsymbol{\theta}^{\mathrm{D}})}{z\; F(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N)}$$

is the firm's constant marginal cost. The firms' profit is decreasing in the unit cost, hence the aggregate skill profile $\boldsymbol{\theta}^{\mathrm{D}}$ is chosen to minimize the cost:

$$\tilde{c}(\alpha_N, z) = \min_{\boldsymbol{\theta}} c(\boldsymbol{\theta}; \alpha_N, z) = \min_{\boldsymbol{\theta}} \frac{w(\boldsymbol{\theta})}{z\; F(\boldsymbol{\theta}; \alpha_N)}. \tag{D.5}$$

From output (D.4), we obtain the resulting labor demand:

$$\Lambda^{\mathrm{D}}(\alpha_N, z) = \frac{y(\alpha_N, z)}{zF(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N)} = I\frac{\sigma - 1}{\sigma}\frac{1}{\int \tilde{c}(\alpha_N, z)^{1-\sigma} H^f(\mathrm{d}\alpha_N, \mathrm{d}z)}\frac{[zF(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N)]^{\sigma-1}}{w(\boldsymbol{\theta}^{\mathrm{D}})^{\sigma}} \tag{D.6}$$

and the wage bill:

$$\begin{aligned} w(T) &= \Lambda^{\mathrm{D}}(\alpha_N, z)\, w(\boldsymbol{\theta}^{\mathrm{D}}) \\ &= I\frac{\sigma - 1}{\sigma}\frac{1}{\int \tilde{c}(\alpha_N, z)^{1-\sigma} H^f(\mathrm{d}\alpha_N, \mathrm{d}z)}\left[\frac{zF(\boldsymbol{\theta}^{\mathrm{D}}; \alpha_N)}{w(\boldsymbol{\theta}^{\mathrm{D}})}\right]^{\sigma-1}. \end{aligned} \tag{D.7}$$

The two environments, the one presented in this Appendix and the competitive one from the main text, have deep similarities. In fact, the Dixit-Stiglitz variant is isomorphic to the model presented in the main text. Using the demand equation (D.1), we rewrite the firm's profit (D.2) as

$$p(\alpha_N, z)y - w(T) = C^{1/\sigma}\left[zF(\mathbf{T}^{\mathrm{D}}; \alpha_N)\right]^{1-1/\sigma} - w(\mathbf{T}^{\mathrm{D}}),$$

where C is a constant. We obtain expression (A.9) for the firms' profit used in our main model by replacing z with $z^{1-1/\sigma}$ and the one-homogenous function F with $C^{1/\sigma}F^{1-1/\sigma}$, which is homogenous of degree $\eta = 1 - 1/\sigma$. Equations (D.6) and (D.7) correspond to Equations (A.10) and (A.11) after the above replacements of z and F. The labor demand elasticity is $\sigma > 1$ here and $1/(1-\eta) > 1$ in the main text. The equilibrium conditions (24) and (25) must be modified according to (D.6) and (D.7). Notice that the underlying primal problem maximizes pF, the gross revenue of firms to be shared with workers, and ignores the surplus of final consumers.

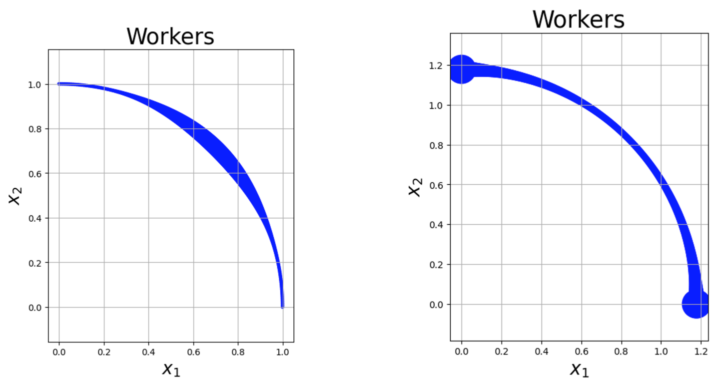
All results obtained under Bundling have their counterpart in the DS world despite differences in the resulting formulas. In particular, the logic of the horizontal matching (link between worker's skill profiles θ and firms' technological intensity α_N) is unchanged. The determination of the aggregate profile θ from (D.5) is the same as in the main text, see for instance Figure 2. The analysis of bunching is also similar as well as the AKM-like decomposition of the log-wage. Furthermore, the analysis of unbundling, presented in our companion paper, can also be carried out within the DS setting.

E Numerical simulations

E.1 The Relative Prevalence of Generalist and Specialist Workers in the Economy

To illustrate how the presence of specialists affects the markdown, we contrast two polar cases in our two-skill environment. In Scenario (A) the skill supply comprises essentially generalists, whereas in Scenario (B) it comprises essentially specialists, i.e., workers are mostly endowed with either cognitive skills or non-cognitive skills. In both scenarios, the production technology is CES of the form (12), with returns to scale parameter $\eta = .5$ and $\sigma = -1$ (complementary skills). The technical intensities α_C and α_N being uniformly distributed on $[0, 1]$. All firms have the same total factor productivity $z = 1$ and all workers have the same quality $\lambda = 1$. In Scenario (A) skill profiles $\theta = \arctan s_N/s_C$ are distributed as a Beta(9,9) random variable, whereas in Scenario (B) they are distributed as a Beta(.8,.8) variable.

In Figure E.2, we present for these two scenarios: (a) the generalist markdown $m(\theta)$; (b) the matching patterns $\theta^{\mathrm{D}}(\alpha_N)$; (c) the implicit prices (marginal product



(a) Mostly Generalists (A) (b) Mostly Specialists (B)

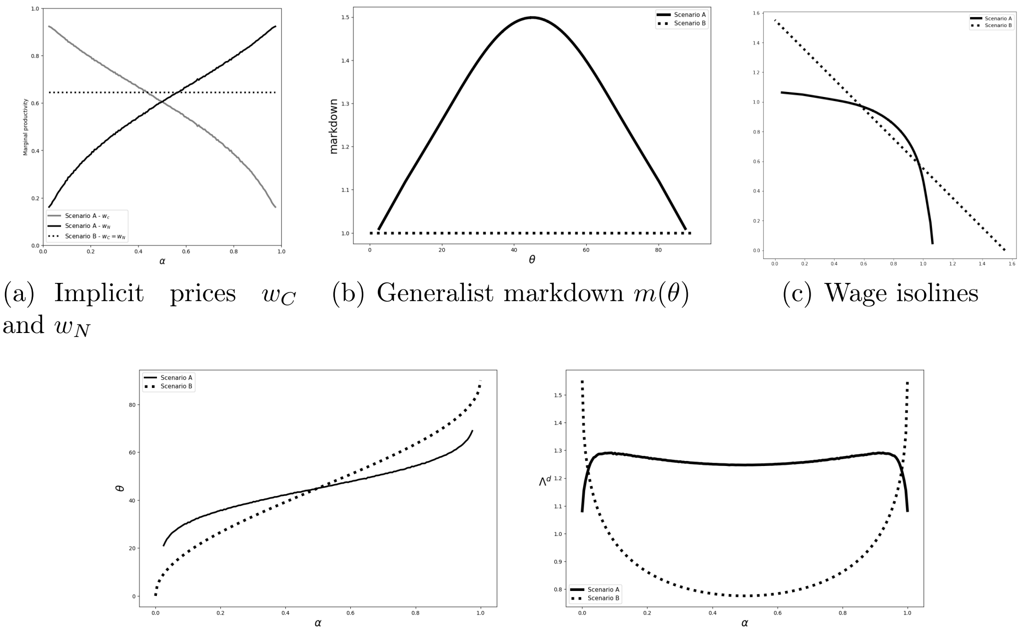
Figure E.1: Mostly Generalists (A) vs Mostly Specialists (B)

of each skill) $w_C(\theta^{\mathrm{D}}(\alpha_N))$ and $w_N(\theta^{\mathrm{D}}(\alpha_N))$ as a function of α; (d) the wage isolines $w(s_C, s_N) = 1$, i.e., the quantity of each skill available for one dollar; (e) the firm sizes $\Lambda^d(\alpha_N; 1)$. In Scenario (A), implicit prices differ across firms, see Figure 2(a): the specialist prices (paid by specialist firms) are $p_C = p_N \approx .92$ while generalist workers are paid $w_C = w_N \approx .6$ for each of their skills. The markdown wage is thus close to 50% as shown on Figure 2(b). In this scenario with many generalists, the wage schedule is nonlinear, there is no bunching, see the nonlinear isoline wage isoline on Figure 2(c). Specialist firms (with either low or large technical parameters αs) are forced to use workers that possess too much of the other specialist skill, see Figure 2(d). Specialist firms cannot attain a large size because their favored employees are expensive, Figure 2(e).

In stark contrast, bunching is pervasive under Scenario (B). The wage schedule is linear, the law of one price applies with the two skills being paid $w_C = w_N \approx .62$ by all firms, see Figure 2(a). There is no markdown (Figure 2(b)) and the isolines are flat (Figure 2(c)). The matching map that connects firms (α) and workers (θ) lies closer to the 45-degree line than under Scenario (A). Because more specialist workers are available in Scenario (B), specialist firms have access to skills that are more aligned with their technologies (Figure 2(d)). The large supply of specialists decreases their wage, thus allowing specialist firms to increase their size, see Figure 2(e).

E.2 Non-Homothetic Production Functions

To numerically solve for optimal workers-to-firms matching, we use, in the spirit of Simonovska (2015) and Jung, Simonovska, and Weinberger (2019), the following non-



(a) Implicit prices w_C and w_N (b) Generalist markdown $m(\theta)$ (c) Wage isolines

(d) Sorting $\theta^{\mathrm{D}}(\alpha_N)$ (e) Firms' sizes $\Lambda^{\mathrm{D}}(\alpha, 1)$

Figure E.2: The Effect of Skills Supply: Mostly Generalists (A) vs Mostly Specialists (B)

homothetic CES function:

$$F(T;\alpha,z) = \frac{z}{\eta}\left(\left[\alpha_C(T_C+A)^\rho + \alpha_N(T_N+B)^\rho\right]^{\eta/\rho} - \left(\alpha_C A^\rho + \alpha_N B^\rho\right)^{\eta/\rho}\right), \tag{E.1}$$

where the technical intensities in the cognitive and non-cognitive skills, α_C and α_N, satisfy $\alpha_C + \alpha_N = 1$.[43] Setting $(\mathbf{T}_C^{\mathrm{D}}, \mathbf{T}_N^{\mathrm{D}}) = \Lambda^{\mathrm{D}}(\boldsymbol{\theta}_C^{\mathrm{D}}, \boldsymbol{\theta}_N^{\mathrm{D}})$ and using the first-order conditions (7) as well as the 1-homogeneity of the wage, we get

$$\frac{F_C(\Lambda^{\mathrm{D}}, \theta^{\mathrm{D}}; \alpha_N)}{F_N(\Lambda^{\mathrm{D}}, \theta^{\mathrm{D}}; \alpha_N)} = \frac{\alpha_C}{\alpha_N}\left[\frac{\Lambda^{\mathrm{D}}\boldsymbol{\theta}_N^{\mathrm{D}} + B}{\Lambda^{\mathrm{D}}\boldsymbol{\theta}_C^{\mathrm{D}} + A}\right]^{1-\rho} = \frac{w_C(\theta^{\mathrm{D}})}{w_N(\theta^{\mathrm{D}})}. \tag{E.2}$$

The marginal product of T_C versus T_N, F_C/F_N, increases with Λ^{D} for skill profiles $\boldsymbol{\theta}^{\mathrm{D}}$ such that $\tan\boldsymbol{\theta}^{\mathrm{D}} = \mathbf{T}_N^{\mathrm{D}}/\mathbf{T}_C^{\mathrm{D}} > B/A$. In this case, the marginal productivity of *Cognitive* skills relative to that of *Non-Cognitive* skills increases with the size of firms, hence big firms use relatively more *Cognitive* skills, implying that θ decreases with z. The "worker-to-firm sorting effect" now becomes linked to the firm's productivity z and becomes closer to a true "firm-effect", see the above discussion. Furthermore, under non-homotheticity, the "worker-to-firm sorting effect" (which captures the intensity of the

[43] Sato (1977)'s approach of non-homotheticity, followed for instance by Lashkari, Bauer, and Boussard (2024), models production isolines and delivers the value of output as an implicit function, which would be hard to integrate within a numerical solver. All simulations presented in the paper use the mirror-descent algorithm of Paty, Choné, and Kramarz (2022).

relative use of the two skills) and the total quality of the firms' workers will be correlated. Indeed, in our model, the strength of the correlation between individual worker quality and the "worker-to-firm sorting effect" will vary: zero under homotheticity whereas, under non-homotheticity, this individual-level correlation will be positive and small when the productive firms employ many average workers but positive and large when the productive firms employ a small number of very high-quality workers.

Our baseline simulation imposes $A = 1.0$ and $B = 0.3$ in the production function (affecting firms' demand). Firms are uniformly distributed with two productivity levels ($z = 1$ and $z = 2$) when the skills' supply follows a Beta(7,4) distribution with relatively few "Cognitive" workers. The distribution of firms and the workers' supply are presented in Figure E.3. In this example, the equilibrium condition can be written as

$$\Lambda^{\mathrm{S}}(\theta)h^{w}(\theta) = \sum_{l=1}^{2} h^{f}(\alpha_N(\theta; z_l))\, \Lambda^{\mathrm{D}}(\alpha_N(\theta; z_l); z_l)\frac{\partial \alpha_N}{\partial \theta}(\theta; z_l), \tag{E.3}$$

where $z_1 = 1$ and $z_2 = 2$, $\Lambda^{\mathrm{S}}(\theta) = 1$, $h^w(\theta)$ is the density of the Beta(7,4) law, and $h^f(\alpha_N) = 1$. The sorting functions are denoted by $\theta(\alpha_N; z_l)$ for $l = 1, 2$ and are determined by (E.2). We denote their inverses by $\alpha_N(\theta; z_l)$ in (E.3).

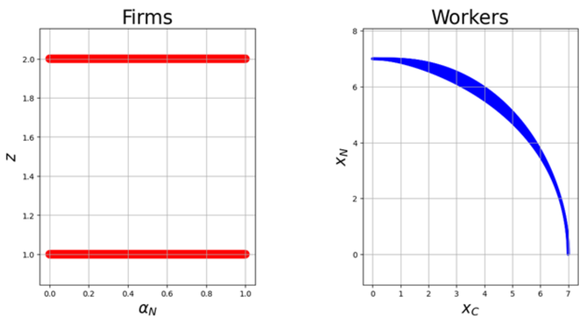


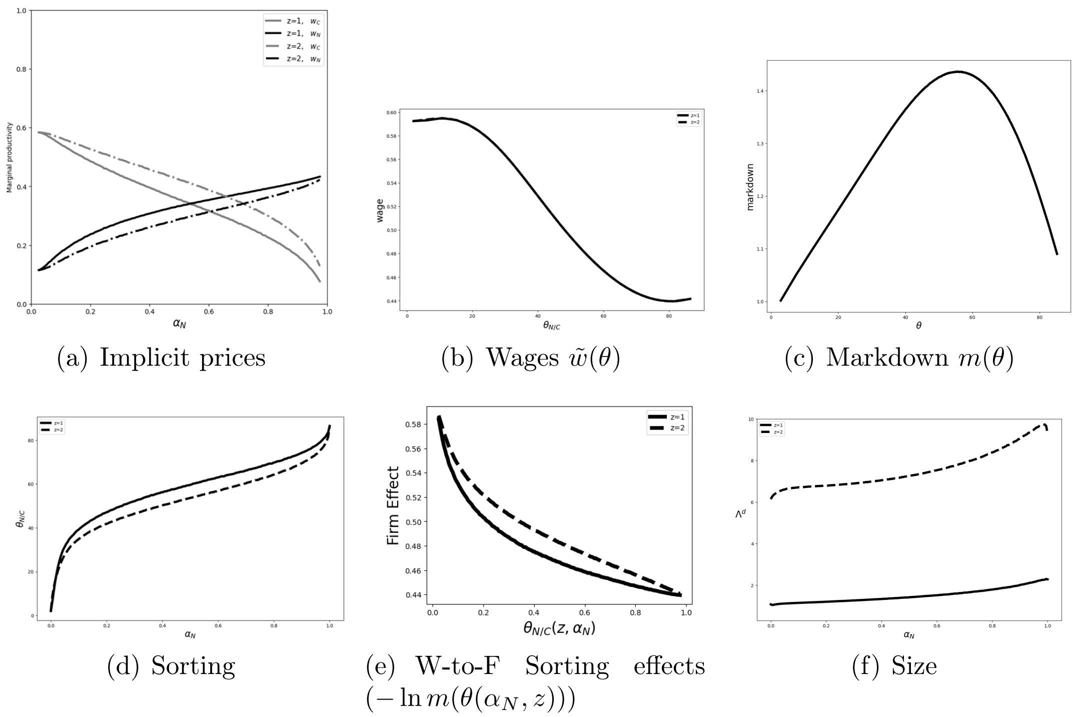
Figure E.3: Firms (uniform with $z = 1, 2$); Skill Supply (Beta(7,4) i.e. "shortage" of skill C)

Figure E.4 presents the resulting Sorting patterns ($\theta_{N/C}$ as a function of α_N), Prices (as a function of α_N), and Wages (as a function of $\theta_{N/C}$) in the top panel, when "worker-to-firm sorting effects" (as a function of (α_N, z), Firm size, (both as functions of α_N ans z) are presented in the bottom panel.

The deficit in supply of cognitive workers drives most of the resulting patterns described below. In Figure 4(d), more productive firms (dashed line) use relatively more cognitive workers than less productive firms (solid line) as evidenced by the lower value of $\theta_{N/C}$ for a given α_N: $\theta_{N/C}(z = 2, \alpha_N) < \theta_{N/C}(z = 1, \alpha_N)$. This increased demand makes the wage decrease in the skill profile $\theta_{N/C}$, and, as a consequence increasing for workers with a comparative advantage in cognitive skills.

As a result, in this configuration, "worker-to-firm sorting effects" (W-to-F sorting effects, hereafter) depend on z, specifically they are larger for more productive firms ($z = 2$), see Figure 4(e). Hence markdowns in more productive firms are lower (closer to – but still above – one) than in less productive firms. The high wages commanded by the C-specialists explain why *Cognitive* firms, *conditional* on z, are smaller than *Non-Cognitive* firms. As predicted by Proposition 4, firms' sizes Λ^{D} increase with their TFP parameter z.

Competitive wages do not depend on firms' productivity z directly (Figure 4(b)). However, workers employed in high-z firms will be better paid than those employed in low-z firms because the former have a more cognitive profile (higher $\theta_{N/C}(z, \alpha_N)$ at given α_N) and, hence, command higher wages and larger Workers-to-Firms sorting effects.



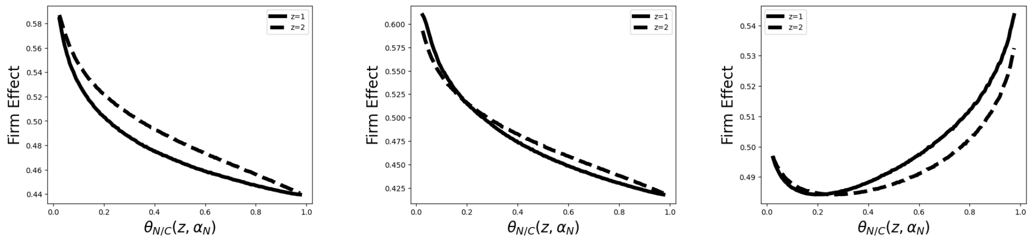
(a) Implicit prices (b) Wages $\tilde{w}(\theta)$ (c) Markdown $m(\theta)$

(d) Sorting (e) W-to-F Sorting effects ($-\ln m(\theta(\alpha_N, z))$) (f) Size

Figure E.4: Wages and sorting with non-homothetic production function (E.1)

In the above configuration, the Workers-to-Firms sorting effects are monotonic in the total factor productivity z, with the variation in z being the same for all technological intensities α_n. This result is not general as Proposition 7 shows and Figure E.5 illustrates. Figures 5(a) reports again the case when both the production function and the skill distribution are strongly asymmetric: $(A, B) = (1, .3)$ in (E.1) and Beta(7,4). In Figure 5(b), we make the production function almost symmetric $(A, B) = (1, .9)$ while keeping the skill distribution asymmetric. In Figure 5(c), we consider a symmetric skill distribution Beta(6,6) with the original asymmetric production function

$(A, B) = (1, .3)$. Comparing these three cases shows that the variations of the Workers-to-Firm effects in the TFP parameter z depend on the details of the primitives and complex general equilibrium effects.



(a) Asym. supply Beta(7,4) with $(A, B) = (1, .3)$ (b) Asym. supply Beta(7,4) with $(A, B) = (1, .9)$ (c) Sym. Supply Beta(6,6) with $(A, B) = (1, .3)$

Figure E.5: Workers-to-Firms Effects ($-\ln m(\theta(\alpha_N, z))$) with non-homothetic production function

F Connection to optimal transport theory

In this section, we explain how our setup is related to optimal transport theory. To fully understand how OT is connected to our contribution, a small detour is required. Our approach requires to account for workers' skill aggregation within firms and for endogenous firm size. These requirements demand a new mathematical framework, developed in Choné, Gozlan, and Kramarz (2023).

Weak optimal transport (WOT) Given two probability measures μ and ν, and a cost function $c(\phi, m)$ that is convex in m, Gozlan, Roberto, Samson, and Tetali (2017) consider the problem of minimizing

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(\phi, p^{\phi}) \mu(\mathrm{d}\phi), \tag{F.1}$$

where $\Pi(\mu, \nu)$ is the set of all couplings π of μ and ν (i.e., the set of probability measures over $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν) and p^{ϕ} is the (μ-almost surely unique) probability kernel such that

$$\pi(\mathrm{d}s, \mathrm{d}\phi) = p^{\phi}(\mathrm{d}s)\, \mu(\mathrm{d}\phi). \tag{F.2}$$

Gozlan, Roberto, Samson, and Tetali (2017) prove existence and duality results for Problem (F.1) under the main requirement that $c(\phi, m)$ is convex in m.

The problem of maximizing total output in the economy, which is given by (4), has the same form as (F.1), with $\mu = H^f$, $\nu = H^w$, and the transport cost defined (for any given $x_0 \in \mathcal{X}$) by

$$c(\phi,\, n^{\mathrm{D}}) = -F\left(\int s\; n^{\mathrm{D}}(\mathrm{d}s); \phi\right) + F(s_0; \phi) + \nabla_s F(s_0; \phi). \left(\int s\; n^{\mathrm{D}}(\mathrm{d}s) - s_0\right).$$

The above cost function is nonnegative by concavity of F in X. Under the equilibrium condition (2), minimizing (F.1) is equivalent to maximizing (4) because $\iint x\, p^{\phi}(\mathrm{d}s)\mu(\mathrm{d}\phi)$ equals $\int x\, \nu(\mathrm{d}s)$, which is a fixed and exogenous quantity.

Unnormalized kernels and endogenous firms' sizes As mentioned in Section 2, the framework developed in the present article has an important difference with the WOT problem described above. Specifically, we do not impose that the workers-to-firms assignments, $n^{\mathrm{D}}(\mathrm{d}s; \phi)$, are *probability* measures, as is required in the kernel disintegration (F.2). Accordingly, Choné, Gozlan, and Kramarz (2023) relax the assumption that the kernel p^{ϕ} in (F.1) is a *probability* measure. Denoting by $\mathcal{M}(\mathcal{Y})$ the set of nonnegative finite measures over $\mathcal{Y}$, they introduce the weak optimal transport problem with unnormalized kernel (WOTUK) as

$$\mathrm{WOTUK}(\mu, \nu) \overset{\mathrm{d}}{=} \sup_{\substack{q \in \mathcal{M}(\mathcal{Y})^{\mathcal{X}} \\ \int q^{\phi}\, \mu(\mathrm{d}\phi) = \nu}} \int_{\mathcal{X}} F(\phi, q^{\phi})\, \mu(\mathrm{d}\phi), \tag{F.3}$$

where $\mathcal{F} : \mathcal{X} \times \mathcal{M}(\mathcal{Y}) \to \mathbb{R}$. The constraint $\int q^{\phi}\, \mu(\mathrm{d}\phi) = \nu$ expresses that the unnormalized kernel (q^{ϕ}) transports μ onto ν. CGK connect the WOTUK problem (F.3) to a WOT problem as follows. Letting

$$\Pi(\ll \mu, \nu) \overset{\mathrm{d}}{=} \{P \in \Pi(\eta, \nu)\,,\, \eta \in \mathcal{P}(\mathcal{X}), \eta \ll \mu\},$$

denote the set of probability measure over $\mathcal{X}$ that are absolutely continuous with respect to μ, CGK show that

$$\mathrm{WOTUK}(\mu, \nu) = \sup_{\eta \in \Pi(\ll \mu, \nu)} \sup_{\pi \in \Pi(\eta, \nu)} \int F\left(\phi, \frac{\mathrm{d}\eta}{\mathrm{d}\mu}(\phi)\pi^{\phi}\right)\, \mu(\mathrm{d}\phi) \tag{F.4}$$

where $\pi^{\phi} \in \mathcal{P}(\mathcal{Y})$ is the unique disintegration of π with respect to η, *i.e.* such that $\pi(\mathrm{d}s, \mathrm{d}\phi) = \eta(\mathrm{d}\phi)\pi^{\phi}(\mathrm{d}s)$. At given η, we get a WOT problem. Instead of constraining the first marginal of π to be μ, the WOTUK problem only imposes that the first marginal is absolutely continuous with respect to μ.

In the economic setting of this paper, the kernel $q^{\phi}(\mathrm{d}s)$ is the assignment of workers to firms $n^{\mathrm{D}}(\mathrm{d}s; \phi)$. The modified firms' distribution is $\eta = \tilde{H}^f = N(\phi)H^f(\mathrm{d}\phi)$, where $N(\phi) = \mathrm{d}\eta/\mathrm{d}\mu = \mathrm{d}\tilde{H}^f/\mathrm{d}H^f$, represents the number of workers employed by firms of type ϕ. The constraint $\int q^\phi \, \mu(\mathrm{d}\phi) = \nu$ is the equilibrium condition (2), with $\nu = H^w$ and $\mu = H^f$. Allowing q^ϕ or n^{D} to be an unnormalized measure instead of a probability measure avoids having to assume that all firms have the same size.

Introducing $\bar{n}^{\mathrm{D}}(\mathrm{d}s;\phi) = n^{\mathrm{D}}(\mathrm{d}s;\phi)/N(\phi)$ the probability measure that describes the distribution of workers' types in firms ϕ, we see that the matching between workers' and firms' types

$$\pi(\mathrm{d}s, \mathrm{d}\phi) = n^{\mathrm{D}}(\mathrm{d}s;\phi) H^f(\mathrm{d}\phi) = \bar{n}^{\mathrm{D}}(\mathrm{d}s;\phi)\tilde{H}^f(\mathrm{d}\phi) \tag{F.5}$$

is a transport plan between the original skill distribution $H^w(\mathrm{d}s)$ and the modified firm distribution $\tilde{H}^f(\mathrm{d}\phi)$.

Conical WOTUK problems The specification studied in the present paper corresponds to a special class of WOTUK problems, which CGK call conical WOTUK problems. It corresponds to the case where

$$\mathcal{F}(x, p) = F\left(x, \int_{\mathcal{Y}} y \, p(\mathrm{d}y)\right)$$

for some $F : \mathcal{X} \times \mathrm{cone}(\mathcal{Y}) \to \mathbb{R}$, where the conical hull of $\mathcal{Y}$ is given by

$$\mathrm{cone}(\mathcal{Y}) \overset{\mathrm{d}}{=} \left\{ \sum_{i=1}^{n} \lambda_i y_i \, , \, \lambda_C, \dots, \lambda_n \in \mathbb{R}_+, y_C, \dots, y_n \in \mathcal{Y}, n \geq 1 \right\}.$$

CGK establish the existence of solutions for the dual problem, which guarantee the existence of a competitive equilibria in our setting where a firm's output depends on the *conical* combination of its employees' types, $\int y \, dq_x(y)$. The combination is said to be "conical" because the mass of q_x is not necessarily equal to one. In other words, the aggregate skill of the workers hired by a firm is not their average skills as in the WOT setting, but their average skills *scaled by the positive factor* $q_x(\mathcal{Y})$ that represents the number of employees.

G Existence of an optimal market-clearing assignment

To prove existence, we extend Theorem 6.4 of Choné, Gozlan, and Kramarz (2023), hereafter CGK. Specifically, we relax their assumption that $\inf_{\phi \in \mathcal{X}^f, s \in \mathcal{X}^s} F(\mu s; \phi)$ tends to $+\infty$ as $\mu \to +\infty$. This assumption indeed does not hold if some workers have zero skill in some dimensions (i.e., the support of H^w intersects the boundary of $\mathbb{R}^d_+$) and the skills are complements as is the case in our baseline CES example (11) with $\rho < 0$.

In Appendix A.1, we use new tools in optimal transport theory to prove the above results. These tools allow us to overcome two challenges. First, contrary to the classic OT framework, the primal problem (4) is a nonlinear function of the assignment n^{D}. For this reason it belongs to the class of weak optimal transport problems introduced by Gozlan, Roberto, Samson, and Tetali (2017). Second, firms' sizes are endogenous in equilibrium, and thus $n^{\mathrm{D}}(\mathrm{d}s;\phi)$ is not necessarily a probability measure for all firms' types ϕ. In the terminology of Choné, Gozlan, and Kramarz (2023), assignments of workers to firms are *unnormalized* kernels. The primal problem can be rewritten as a transport problem over the set of probability distributions whose first marginal is absolutely continuous with respect to the distribution of firms' types, H^f, recall the discussion after (3). (Appendix F presents in greater detail the connection between competitive matching equilibria and optimal transport theory.)

Lemma G.1. *Under the regularity assumption made at the beginning of Section 2, the production function F satisfies Condition (A') of CGK, i.e., there exists sequences $(u_n(\phi))_{n\geq 0}$ and $(v_n(\phi))_{n\geq 0}$ of continuous functions on $\mathcal{X}^f$ and taking values respectively in $\mathbb{R}$ and $\mathbb{R}^d$ such that*

$$F(s;\phi) = \inf_n u_n(\phi) + v_n(\phi)\, s. \tag{G.1}$$

Furthermore, the value of the primal problem is finite: $\mathcal{J}^b(H^f, H^w) < \infty$.

Proof. Take $(s_n)_{n\geq 0}$ a dense subset of $\mathcal{X}^s$. For each n and all $s \in \mathcal{X}^s$, the concavity of F yields

$$F(s;\phi) \leq F(s_n;\phi) + \nabla_s F(s_n;\phi)(s - s_n).$$

By assumption, the functions u_n and v_n defined as

$$u_n(\phi) = F(s_n;\phi) - \nabla_s F(s_n;\phi)s_n \;\text{ and }\; v_n(\phi) = \nabla_s F(s_n;\phi)$$

are continuous in ϕ. For any ϕ, the function $\inf_n u_n(\phi) + v_n(\phi)\, s$ is concave in s, is greater than or equal to $F(s;\phi)$, and coincides with F at all points s_n. This yields the result by continuity. Furthermore, assuming that the total skills in the economy are finite and that $\mathcal{X}^f$ is compact, we have for all market-clearing assignment n^{D}:

$$\int F\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)\right) H^f(\mathrm{d}\phi) \leq \bar{U} + \bar{V}\int sH^w(\mathrm{d}s) < \infty,$$

with $\bar{U} = \sup_{\phi\in\mathcal{X}^f} u_0(\phi)$ and $\bar{V} = \sup_{\phi\in\mathcal{X}^f} v_0(\phi)$.

The following a priori estimate holds:

$$\begin{aligned}\int F\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)\right) H^f(\mathrm{d}\phi) &\leq \int \left\{u_0(\phi) + v_0(\phi)\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi)\right\} H^f(\mathrm{d}\phi)\\ &\leq \bar{U} + \bar{V}\iint s\, n^{\mathrm{D}}(\mathrm{d}s;\phi) H^f(\mathrm{d}\phi)\\ &= \bar{U} + \bar{V}\int sH^w(\mathrm{d}s),\end{aligned}$$

implying that $\mathcal{J}^b(H^f, H^w) < \infty$. □

As explained below Equation (3), one can associate to any market-clearing assignment n^{D} a probability distribution π on $\mathcal{X}^f \times \mathcal{X}^s$ whose first marginal $\tilde{H}^f$ is absolutely continuous with respect to the firms' distribution H^f. Disintegrating such a distribution as $\pi(\mathrm{d}s, \mathrm{d}\phi) = \pi_1(\phi)\, q(\mathrm{d}s;\phi)$, where the first marginal π_1 is absolutely continuous with respect to H^f, one can define the functional

$$\mathcal{I}(\pi; H^f) = \int F\left(\int \frac{\mathrm{d}\pi_1}{\mathrm{d}H^f}\int sq(\mathrm{d}s;\phi)\right) H^f(\mathrm{d}\phi).$$

CGK introduce the set of distributions whose first marginal $\tilde{H}^f$ is absolutely continuous with respect to H^f and whose second marginal is the distribution of skills H^w:

$$\Pi(\ll H^f, H^w) \stackrel{\mathrm{d}}{=} \{\pi \in \Pi(\tilde{H}^f, H^w)\,,\, \tilde{H}^f \in \mathcal{P}(\mathcal{X}^f), \tilde{H}^f \ll H^f\},$$

with $\Pi(\mu,\nu)$ denoting the set of all couplings between μ and ν. An assignment n^{D} is optimal if and only if the associated coupling π maximizes $\mathcal{I}(\pi; H^f)$ on $\Pi(\ll H^f, H^w)$. The latter set, however, is generally not closed, and for this reason the existence of an optimal assignment, which CGK call a "strong solution", is not guaranteed. Accordingly, they introduce a notion of weak solution. They say that a finite measure π on $\mathcal{X}_f \times \mathbb{R}^d_+$ is a weak solution of the optimal transport problem if there exists a sequence of market-clearing assignments n^{D}_n such that the total output $\int F\left(\int s\, n^{\mathrm{D}}_n(\mathrm{d}s;\phi);\phi\right) H^f(\mathrm{d}\phi)$ tends to $\mathcal{J}^b(H^f, H^w)$. They prove that weak solutions always exist (Proposition 2.4).

To investigate the existence of strong solutions, CGK show that the closure of $\Pi(\ll H^f, H^w)$ is the set $\Pi(\mathcal{X}^f; H^w)$ of all probability measures such that $\pi_1(\mathcal{X}^f) = 1$ and $\pi_2 = H^w$. Assuming that $F'_\infty(s;\phi) = \lim_{\mu\to\infty} F(\mu s;\phi)/\mu$ is a continuous function on $\mathcal{X}^f \times \mathbb{R}^k_+$, they introduce the functional

$$\bar{\mathcal{I}}(\pi; H^f) = \int F\left(\int \frac{\mathrm{d}\pi_1^{\mathrm{ac}}}{\mathrm{d}H^f}\int sq(\mathrm{d}s;\phi)\right) H^f(\mathrm{d}\phi) + \iint F'_\infty(s;\phi) q(\mathrm{d}s;\phi)\pi_1^{\mathrm{sing}}(\mathrm{d}\phi),$$

where $\pi(\mathrm{d}\phi, \mathrm{d}s) = \pi_1(\mathrm{d}\phi)\, q(\mathrm{d}s;\phi)$ and $\pi_1 = \pi_1^{\mathrm{ac}} + \pi_1^{\mathrm{sing}}$ is the decomposition of the first marginal π_1 into absolutely continuous and singular parts with respect to H^f. They show that $\bar{\mathcal{I}}$ is a lower semicontinuous extension of $\mathcal{I}$ on $\Pi(\mathcal{X}^f; H^w)$. Under Assumption (A') mentioned in Lemma G.1, Theorem 3.7 of CGK states that a measure π is a weak solution if and only if it maximizes $\bar{\mathcal{I}}(\pi; H^f)$ over $\pi \in \Pi(\mathcal{X}^f; H^w)$. Moreover, the value of the maximum is nothing else than $\mathcal{J}^b(H^f, H^w)$. Assumption 1 states that $F'_\infty = 0$, implying that the functional simplifies into

$$\bar{\mathcal{I}}(\pi; H^f) = \int F\left(\int \frac{\mathrm{d}\pi_1^{\mathrm{ac}}}{\mathrm{d}H^f} \int sq(\mathrm{d}s;\phi)\right) H^f(\mathrm{d}\phi).$$

We now extend the results of CGK to prove the existence of an optimal market-clearing assignment under weaker assumptions. We start from any weak solution $\bar{\pi} = \bar{\pi}(\mathrm{d}s;\phi)\bar{\pi}_1(\mathrm{d}\phi)$. This means that $\bar{\pi}$ maximizes $\bar{\mathcal{I}}(\bar{\pi}; H^f)$ on $\Pi(\mathcal{X}^f; H^w)$ and $\bar{\mathcal{I}}(\bar{\pi}; H^f) = \mathcal{J}^b(H^f, H^w)$.

We set $\gamma(\mathrm{d}s) = \int \bar{\pi}^\phi(\mathrm{d}s)\bar{\pi}^{\mathrm{sing}}(\mathrm{d}\phi)$ and consider the assignment n^{D} defined by

$$n^{\mathrm{D}}(\mathrm{d}s;\phi) = \frac{\mathrm{d}\bar{\pi}^{\mathrm{ac}}}{\mathrm{d}H^f}(\phi)\bar{\pi}^\phi(\mathrm{d}s) + \gamma(\mathrm{d}s).$$

The assignment n^{D} clears the market because

$$\begin{aligned}\int n^{\mathrm{D}}(\mathrm{d}s;\phi)H^f(\mathrm{d}\phi) &= \gamma(\mathrm{d}s) + \int \bar{\pi}^{\mathrm{ac}}(\mathrm{d}\phi)\bar{\pi}^\phi(\mathrm{d}s)\\ &= \int \left[\bar{\pi}^{\mathrm{ac}}(\mathrm{d}\phi) + \bar{\pi}^{\mathrm{sing}}(\mathrm{d}\phi)\right]\bar{\pi}^\phi(\mathrm{d}s) = \int \bar{\pi}(\mathrm{d}\phi)\bar{\pi}^\phi(\mathrm{d}s) = H^w(\mathrm{d}s).\end{aligned}$$

Since $F(.;\phi)$ is nondecreasing on $\mathbb{R}_+^k$ for all ϕ,[44] it holds

$$F\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi);\phi\right) = F\left(\frac{\mathrm{d}\bar{\pi}^{ac}}{\mathrm{d}H^f}(\phi)\int s\bar{\pi}^\phi(\mathrm{d}s) + \int s\gamma(\mathrm{d}s);\phi\right) \geq F\left(\frac{\mathrm{d}\bar{\pi}^{ac}}{\mathrm{d}H^f}(\phi)\int s\bar{\pi}^\phi(\mathrm{d}s);\phi\right).$$

Integrating with respect to $H^f(\mathrm{d}\phi)$ and using the fact that $\bar{\pi}$ is a weak solution, we find

$$\int F\left(\int s\, n^{\mathrm{D}}(\mathrm{d}s;\phi);\phi\right) H^f(\mathrm{d}\phi) \geq F\left(\frac{\mathrm{d}\bar{\pi}^{ac}}{\mathrm{d}H^f}(\phi)\int s\bar{\pi}^\phi(\mathrm{d}s);\phi\right) H^f(\mathrm{d}\phi) = \mathcal{J}^b(H^f, H^w)$$

and hence that n^{D} an optimal market-clearing assignment.

[44] Any concave function $F : \mathbb{R}_+^k \to \mathbb{R}_+$ is nondecreasing. To see this, consider $T_1 \leq T_2$, use concavity to write: $F(T_2) = F(T_1 + T_2 - T_1) \geq tF(T_1/t) + (1-t)F((T_2 - T_1)/(1-t)) \geq tF(T_1)$, and let t go to 1.

H Data Appendix

H.1 Data Description

The empirical exercises presented in this paper make use of Swedish administrative data. They connect information from the military draft that tested virtually all Swedish males at the age of 18. We use two principal scores—capturing Cognitive and Non-Cognitive skills—from these tests. These have been used extensively in previous research on wages and labor market sorting (see e.g. Lindqvist and Vestman, 2011; Fredriksson, Hensvik, and Skans, 2018; Skans, Choné, and Kramarz, 2023). A detailed description of the data and the testing procedure is presented in Mood, Jonsson, and Bihagen (2012). We follow Skans, Choné, and Kramarz (2023) and use the scores on their nine -step Stanine scale throughout the paper. Our skills data cover tests administered during 1970 to 1996, with the exception of 1978 where two thirds of the data are missing, anecdotally because of a fire. The data cover birth years 1951 to 1978.

The cognitive skills data are collected from written exams. The tests analyze the subjects' skills in terms of spatial ability, verbal comprehension, inductive and technical skills. The Non-Cognitive skills are assessed during a structured 20-minute interview with a trained psychologist. These assessments cover sub-aspects such as social maturity, psychological energy, intensity and tolerance to stress. We measure both Congitive and Non-Cognitive skills, we use the combined score reported at a nine-step Stanine scale (see below) as this is a consistent metric available for a large set of cohorts. Using a two-dimensional skill-vector also simplifies exposition.

The skill scores are reported at the 1-9 integer Stanine scale, built to approximate a normal distribution by nine discrete bins. Each step corresponds to approximately 0.5 standard deviations and the mean skill is 5. In contrast to percentile ranks that approximate a uniform distribution, Stanine bins approximate a normal distribution and are therefore of unequal size, each capturing 4, 7, 12, 17, 20, 17, 12, 7, 4 percent of the sample as we move from 1 to 9. Our measure of skill endowments is thus defined w.r.t. a worker's economy-wide peers. Hence, the difference between Cognitive and Non-Cognitive scores captures whether the worker is relatively stronger in one dimension than in the other, allowing to use the score difference as a direct measure of each worker's comparative advantage.

We link the skills data to outcome data for those employed in the private sector for the period 2002 to 2013. The outcome data cover information on wages, occupations, firm identities as well as information from the firms' financial accounts. We restrict the data to stop at 2013 to ensure that we have consistent information about occupations (the codes change in 2013). Because of the intersection between restrictions on outcome years and birth cohorts, workers in the sample are aged 24 to 62. The wage and occupation data are from the structure of earnings statistics firm-level sample. The data over-samples large firms and the data cover around 50 percent of all private sector workers. We only keep firm-years with at least 10 workers with complete skills data and drop all firm-years where we have skill information for less than 1/4 of all workers. Wage data are Winzorized at the $1st$ and $99th$ percentile within years. The used data have $550,000$ unique individuals, 3.3 million worker-years and $30,000$ firm-years.

H.2 Details regarding subsection 2.1 ("suggestive evidence")

The subsection presents evidence using the difference between Cognitive and Non-Cognitive scores as the outcome. As the scores are measured at the 1-9 scale, the differences can in principle fall anywhere in the $[1-9]$ range. However, end-point results are rare by definition and scores are mildly positively correlated ($\rho = 0.37$ in our used sample), so very few end up in the extreme end-points of the difference distribution. Figure 1(a) shows the resulting comparative advantage distribution, using one observation per individual in our used worker-level sample. As the figure shows, about half of all workers have a comparative advantage in one of the skill dimensions of at least 1 standard deviation (2 Stanine steps). More precisely, 20 percent have the exact same score and 17 percent (on each side) have a gap of exactly 1, and the rest have a larger gap as illustrated by the graph.

Figure 1(b) relates workers' comparative advantage to their sorting patterns across firms using one observation one observation per firm in our used data. The figure shows the distribution of *firm-averaged* comparative advantages. The data are thus computed by simply taking the firm-level mean of the worker-level data used in panel a). The distribution illustrates the firm-level heterogeneity of specific skill use. At the worker-level, the dispersion implies that half of all workers are employed by a firm where the average coworker has a comparative advantage in one dimension of at least 0.5 on the Stanine scale (thus 1/4 standard deviations).

We perform two exercises to show that these patterns reflect heterogeneous firm-level labor demand and not random noise. First, we show that firm-averaged comparative advantages correlate across occupations within the same firm. To perform the analysis, we first subtract the economy-wide occupational mean from the individual-level comparative advantage. Next, we aggregate the ensuing residual to the occupation-times-firm level and compute the leave-one(-occupation)-out association across occupations within the same firm. The results presented in Figure 1(c) show that firms with an unusually high cognitive advantage in one occupation, also have an unusually high cognitive advantage in its other occupations. Hence, firms are systematically heterogeneous in their use of cognitive vs non-cognitive skills in way that goes beyond occupational boundaries.

Second, we show that the firm-averaged comparative advantage is predictable from firm-level economic fundamentals. Here, we use data from the financial accounts under the presumption that heterogeneity in accounting data across firms is an indication of differences in production technologies. This is a natural extension of the fact that accounting data tend to be used for firm-level production-function estimations. We use a fully data-driven approach using 25 different variables (all we have in this version of the data) from the accounts as prediction features in a Random Forest model we train to predict the average firm-level comparative advantage. The features capture costs, revenues, assets and liabilities, but contain no information related to worker composition. Features are Winzorized at the $1st$ and $99th$ percentile and standardized. Because all measures are a function of size, we de-correlate them by dividing all measures by the overall wage cost, and include the wage cost itself in log form.

We estimate the Random Forest prediction models with the firm averaged comparative advantage as the outcome.[45] We use scikit-learn accessed via Stata with default parameter settings. We perform 5-fold cross-validations with firm-clustered folds to ensure that predictions are always across firms even though we use data from several years. We illustrate the out-of-fold prediction properties in Figure 1(d). The model explains around 20 percent of the variation in comparative advantages out-of-sample. The clear positive slope implies that our fairly crude accounting data can be used to learn about the demand for workers' comparative advantages across firms. The most relevant predictors are displayed in Figure H.1 below.

H.3 Estimation of Size and Markdown

In subsection 3.2 we estimate the size, as defined in equation 8 relying on estimates of the price of skills in pure specialist skill segments. Furthermore, in subsections 3.5, 4.4 and 5.2 we provide estimates of the wage markdown and these also require estimates of the returns to skills in specialized skill segments. The approach we use is outlined in 3.5 in the paper. Here, we provide some additional details.

Firms are grouped based on the average comparative advantage among its employees. We compute the average across all years of observation. We then define a Generalist firm as having an average skill difference in the $(-0.5, 0.5)$ range. This corresponds to a gap of 1/4 of an individual level standard deviation. Around half of all workers

[45] The approach can be related to the unsupervised K-means clustering of firms used in Bonhomme, Lamadon, and Manresa (2019). The K-means puts equal weight on all explanatory features and create K groups with similar features. The Random Forest instead creates a prediction by weighting outcomes of firms with similar features as the focal firm. The Random Forest only pays attention to feature combinations that help predict the outcome.

Figure H.1: Feature importance in the Random Forest used for figure 1(d)
Note: Mean Impurity Decrease as a measure of feature importance, using Python's skcit built-in feature importance algorithm. Distribution across the 25 included features with labels on the most important ones. See the text for details on the Random Forest estimation.

are working in these generalist firms. The remaining workers are either working in N-specialized or C-specialized firms.

In subsection 3.2, size of firm j in year t is computed by multiplying the number of workers (Emp_{jt}) overall, i.e. including women and untested males, with the price-weighted skill composition among those males that are tested:

$$Size_{jt} = Emp_{jt}(\hat{\phi_C^C}\bar{C}_j + \hat{\phi_N^N}\bar{N}_j)$$

where $\hat{\phi_N^N}$ and $\hat{\phi_C^C}$ refers to the skill estimates in specialized segments as explained in Section 3.5 and $\bar{C}_j, \bar{N}_j$ refers to the average skill-levels of workers employed by the firm. The underlying assumption is that the skill composition of tested workers are reflective of skills of untested workers.

Estimates displayed in Table 2 are based on splits of generalist firms across markets and firms. We first split the generalist sample across markets, into a one sample where bunching is more prevalent, i.e. markets where the generalist firms are more prone to relying on a mix of specialists, and one sample drawn from markets where generalist firms use more generalist workers. To achieve this split, we define a market as a combination of a 3-digit ISCO occupation and a 3-digit NUTS region. We define generalist workers as those having a most a 1-Stanine step between their cognitive and non-cognitive score (about half of the sample, as shown above). For each market, we compute the share of generalist workers within generalist firms and then rank the markets based on this share. We define markets in the lowest quartile as "bunching" and the highest quartile as "no bunching". We ignore the generalist firms in the middle range of markets to get a clearer spread across markets. The last two columns instead split the generalist firms by labor productivity using a simple split at the worker-weighted median.

H.4 Non-homothetic Production Functions

In subsection 5.2 we provide estimates of the relationship between firm productivity and a) relative wage returns to the two types of skills and b) sorting patterns. Firm-years are ranked based on labor productivity (value added per worker) using the financial accounts and then put into decile bins. For each decile, we regress log wages on Cognitive and Non-cognitive skills and plot the difference in estimates across these two skills. Next, we show firm-averaged cognitive advantage across the log labor productivity distribution.