Review of Economic Dynamics 38 (2020) 59–93

Contents lists available at ScienceDirect

Review of Economic Dynamics

www.elsevier.com/locate/red

Size-dependent policies, talent misallocation, and the return to skill ☆

José Joaquín López [a,*], Jesica Torres [b]

[a] University of Memphis, United States of America
[b] The World Bank, United States of America

ARTICLE INFO

Article history:
Received 16 June 2018
Received in revised form 16 March 2020
Available online 23 March 2020

Keywords:
Organization of production
Misallocation
Size-dependent distortions
Returns to skill
Entrepreneurship
Size-distribution of plants

ABSTRACT

We study the allocation of talent in knowledge-based hierarchies subject to a payroll tax that increases with establishment size. The tax distorts the allocation of talent across occupations, as well as the sorting of all infra-marginal agents, thus attenuating the strength of the positive sorting throughout the entire economy. This talent misallocation results in lower output, smaller plants, higher self-employment, less wage employment, and lower returns to skill. To quantify these effects, we first estimate a tax policy to match the establishment-level evidence on the size-dependent compliance of social security contributions in Mexico, and conduct two sets of numerical exercises. Introducing this size-dependent policy into an undistorted economy calibrated to the U.S. generates a reduction in average plant size and output losses of 10 percent. The returns to skill for wage workers decrease by 75 percent. When we isolate the margins of misallocation in our model, we find that slightly over half of the output losses are due to talent mismatch, whereas the rest is accounted for by the best wage workers turning into self-employment. Similarly, eliminating the labor distortion in Mexico increases average plant size by 12 percent, and output by 9 percent, while the average return to skill for workers increases by 14 percent. Perfect enforcement of the average effective tax accounts for one fifth of the output gains from removing the tax. Size-dependent policies in our model generate a reduction in average plant size that is only a fifth of that obtained using a standard span-of-control model, yet output losses are three times as large. The main losers from this type of policies are high-skill wage workers.

☆ We are especially grateful to Santiago Levy and Erwan Quintin for many useful suggestions and their constant encouragement. We also thank the editor, Diego Restuccia, as well as two anonymous referees whose comments greatly improved the quality of the paper. We have benefited from discussions with Enghin Atalay, Chang-Tai Hsieh, Erik Hurst, Rafael Lopes de Melo, Bernabé López-Martín, Casey Mulligan, Andrés Neumeyer, Paulina Restrepo, Ctirad Slavík, and John Van Reenen, as well as participants at various conferences and seminars. This work was supported in part by Competitive Summer Faculty Grant from the Fogelman College of Business & Economics at the University of Memphis. This research support does not imply endorsement of the research results by either the Fogelman College or the University of Memphis. All errors are our own.

* Corresponding author. Department of Economics, Fogelman College of Business & Economics, University of Memphis, 3675 Central Ave., Memphis, TN 38152, United States of America.
E-mail address: jjlopez@memphis.edu (J.J. López).

https://doi.org/10.1016/j.red.2020.03.005
1094-2025/

1. Introduction

Workers in large establishments are, on average, more skilled than workers in smaller plants. Further, larger establishments tend to be run by more educated managers, relative to those of smaller size. If more talented individuals form larger establishments, which, in turn, tend to have a larger share of highly skilled workers, then policies that affect plants of different sizes differently will distort the allocation of talent, as well as the return to skill. In this paper we argue, in particular, that regulations which favor small businesses misallocate talent throughout the entire economy, and lower the return to skill.

The relationship between skill composition and establishment size holds across countries, and for different measures of skill and size. Headd (2000) reports that in the U.S. small firms are more likely to employ workers with a high school diploma or less, whereas workers with at least some colleges are more likely to work in larger firms. Also for the U.S., Cardiff-Hicks et al. (2014) find that higher quality workers are sorted into large firms and large establishments in retailing. Fox (2009) documents evidence for Sweden consistent with hierarchical matching, while Busso et al. (2012) document a positive relation between cognitive skills and firm size for both employers and employees in OECD and Latin American countries.

In this paper, we study the allocation of talent in a knowledge economy where plants are subject to size-dependent regulations, which we model as a tax on labor that increases with the number of employees in the establishment. To this end, we develop a model where production is organized in knowledge hierarchies, building on the work of Garicano and Rossi-Hansberg (2004, 2006). We make two extensions to their original framework that allow us to analyze how individuals with heterogeneous abilities are sorted into occupations and into plants of varying sizes, while delivering a more realistic distribution of sizes relative to the original model. First, we allow communication costs between production workers and managers to decrease with the skill of wage workers. Second, we embed this environment into a production economy subject to decreasing returns to scale, as in Lucas (1978). The latter modification provides a way to organize production under certain distortions that break down positive sorting, as well as comparability with the rest of the misallocation literature. Under a specific parameterization detailed below, our environment generates closed-form solutions for all equilibrium objects, including a Pareto distribution of plant sizes.

In the model, the skill of an agent completely determines his occupation, the quality of his match, as well as the size of his plant, and the return to skill derives from matching with more talented workers in larger plants. The size-dependent tax encourages managers to constrain the size of their productive unit, which they achieve by matching with workers of lower skill. The tax distorts not only the allocation of the marginal individual in each occupation, but also the sorting of the infra-marginal workers, thus attenuating the strength of the positive sorting throughout the entire economy. Both talent misallocation and a lower return to skill originate from the same source: distortions on occupational choices, which reorganizes production by re-sorting everyone within occupations, generating a talent mismatch.

To understand the magnitude of these effects, we first estimate a tax policy to match the establishment-level evidence on the size-dependent compliance of social security contributions in Mexico, as documented in Busso et al. (2018). Our tax policy features a finite asymptotic rate characterized by a single parameter that can be set to match any statutory rate, whereas the degree of progressivity (enforcement) is also summarized by one parameter. Thus, our proposed tax policy improves on some of the issues, while maintaining the advantages, of the popular tax policy featured in Benabou (2002), which is used by Guner et al. (2018) to study the effects of hypothetical size-dependent policies.

We then calibrate two benchmark economies: an undistorted economy with parameters set to match some features of the distribution of establishment sizes and the allocation of workers across occupations in the U.S., and a distorted version with our fitted tax policy, where the parameters are set to match the observed allocation of talent, as well as several moments of the distribution of establishment sizes in Mexico. In a recent paper, Poschke (2018) argues that differences in entrepreneurial technology and—to a lesser degree—the severity of (general, hypothetical) size-dependent distortions, account for most of the variation in average firm sizes across countries. Our findings are largely consistent with this view. Our calibrated model economy for the U.S. features a superior communication technology, higher returns to scale, more abundant skill, and more complicated production problems than our calibrated distorted benchmark for Mexico, which suggests that technology and skills are of first-order importance to understand the observed differences in establishment size distributions and occupational shares across countries.

We then conduct two sets of counterfactual experiments. We start by introducing two different policies in our undistorted U.S. benchmark. First, we introduce the size-dependent tax calibrated from the micro-evidence on labor regulations in Mexico. Under this tax, self-employment increases more than two-fold, to a level close to that observed in Mexico, and the returns to skill for wage workers (managers) decrease by 75 (3) percent. The resulting misallocation of talent has the best wage workers turn into self-employment, thus reallocating the best managers to smaller teams comprised of employees of lower skill. When we isolate the margins of misallocation in our model, we find that slightly over half of the output losses are due to talent mismatch, whereas the rest is accounted for by the best wage workers turning into self-employment.

Even though the average establishment size decreases by only 10 percent as a result of the size-dependent tax, output and average plant productivity decrease by 10 and nine percent, respectively.[1] Thus, the positive sorting mechanism generates a nearly one-to-one positive relationship between percentage changes in plant size and output, which is considerably stronger than what has been previously documented (e.g., see Bento and Restuccia (2017)). Size-dependent policies in our model generate a reduction in average plant size that is only a fifth of that obtained using a standard span-of-control model calibrated to match the same moments of the establishment size distribution, yet the implied output losses are three times as large.

In the second experiment using our U.S. benchmark, we introduce a perfectly-enforced proportional tax of the same level as the average tax obtained under size-dependent enforcement. The effects of a flat tax are large: they account for more than half the output and productivity losses from the size-dependent tax. However, the effects on the returns to skill are much different. Workers at the bottom of the skill distribution are the main losers from the flat tax, which implies that the returns to skill for wage workers increase modestly.

Next, we investigate how our model can address a set of documented facts about the Mexican economy: persistently-high levels of business ownership (especially self-employment), a proliferation of small businesses, a growing mismatch between the supply and demand for skill, and high statutory labor costs that are more stringently enforced on larger establishments. To this end, we take our distorted benchmark calibrated to Mexico, and then consider two possible counterfactual policy scenarios: complete elimination of the payroll tax, and perfect enforcement of the average effective tax rate under size-dependent enforcement. A complete elimination of the payroll tax has large positive effects: self-employment drops by more than 40 percent, and wage employment increases by 13 percent. Output and average plant productivity increase by nine and eight percent, respectively. Average returns to skill for workers increase by 14 percent, and, while all workers see an increase in their wages, high-skill workers experience the largest gains. Perfect enforcement of the average tax accounts for slightly over one fifth of the output and productivity gains from eliminating the tax altogether. Wage workers reap the highest benefits from this policy, as their average returns to skill increase by 35 percent, which represents almost two thirds of the gains from dispensing with the payroll tax. However, these gains are not evenly distributed across skill levels: high skill wage workers benefit the most, while low skill workers see a reduction in wages, brought about by the higher taxes paid by smaller plants.

We find that the skill composition of productive units, which arises endogenously from positive sorting, is crucial to fully understand the effects of plant-specific distortions. In doing so, we fill an important gap in the vast literature on misallocation spurred by Restuccia and Rogerson (2008) and Hsieh and Klenow (2009), where the skill composition of plants has remained absent. Further, by fully considering the role of the skill distribution, our work complements the literature on size-dependent policies. For instance, the works of Braguinsky et al. (2011), Garicano et al. (2016), and Guner et al. (2008), consider heterogeneity in managerial skill, but this talent becomes useless if agents choose to work for a wage. The evidence on positive sorting in quality and quantity discussed above suggests potentially larger aggregate effects through talent misallocation. Indeed, output losses from size-dependent policies in our model are at least three times larger than those in a standard span-of-control model calibrated to the same moments of the establishment distribution in the U.S. In a closely related paper, Alder (2016), shows that deviations from positive sorting between CEO and projects (firms) can have sizable aggregate effects, depending on the degree of complementarity between projects and managers, as well as the correlation between mismatch and project quality. Our paper differs from Alder's in that we focus on the effects of specific, observable size-dependent distortions on talent misallocation across the entire skill distribution—including wage workers, the self-employed, and managers—as opposed to the allocation between heterogeneous managers and projects of varying quality.

Another important contribution of our paper is showing that size-dependent regulations could significantly lower the average return to skill in the economy. Our model features superstar effects, in the sense that the earnings schedule is convex in ability, as in Rosen (1981) and, more recently, Scheuer and Werning (2017). Therefore, because the best managers are matched with the best wage workers, distortions that increase with establishment size act as an increasing marginal tax schedule that attenuates the strength of the positive sorting, lowering the return to skill for both wage workers and managers. The effects of size-dependent regulations on the returns to skill has remained absent from the literature in part because most misallocation models treat labor as a homogeneous input. In our model, just as in Garicano et al. (2016), the main losers from size-dependent regulations are wage workers. However, we are able to provide more nuance to that effect: it is the most talented wage workers that suffer from size-dependent enforcement.

Our model predicts misallocation of talent and a lower return to skill as consequences of size-dependent policies for a given distribution of skills. That means that even if individuals became more skilled for other reasons, they would still be misallocated as long as the size-dependent policies persisted. This prediction is in line with the evidence for Mexico presented in Levy and López-Calva (2016), who document a growing mismatch between the supply and demand for skilled labor: while the amount of available skilled workers has grown, the earnings of more educated workers have decreased. While Levy and López-Calva (2016) suggest a link between size-dependent policies and the return to skill, our paper is the first to formalize such link.

[1] While a 10 percent reduction in establishment size may seem modest, notice that we are studying changes in a very specific, directly observable, size-dependent tax. Changes in a single source of misallocation are unlikely to explain a large share of the observed differences between countries. A notable exception is Adamopoulos and Restuccia (2014).

Our work also contributes to the theory of Pareto distributions of plant sizes by providing a specific example of how a Pareto team-size distribution can arise from positive sorting between heterogeneous workers and managers. Our result is closely related—yet independently and contemporaneously developed—to the more general theory of Pareto distributions by Geerolf (2017). The Pareto shape is the result of positive sorting in a purely static environment, and stands in contrast with previous work where Pareto size distributions arise from either assuming that some primitive distribution is itself Pareto—e.g. managerial talent, or firm-level productivity—as in Lucas (1978), or from a long sequence of large and persistent positive shocks, as in Luttmer (2007), investment in managerial skills as in Guner et al. (2018) and Bhattacharya et al. (2013), or skill-biased change in entrepreneurial technology, as in Poschke (2018).

Finally, we argue that the positive-sorting framework offers new insights on the sources and consequences of misallocation. In economies that exhibit positive sorting, the best managers match with the best wage workers to form the largest plants. Thus, distortions that affect the production decisions of the largest productive units will generate a trickle-down effect, re-sorting workers within and across occupations, which is ultimately reflected in reductions in average size, and output. We find that the level of the tax is of first-order importance for the magnitude of the output losses, while the degree of enforcement affects the returns to skill. In other words, flat taxes can have large effects on output, but do not impact the returns to skill in a negative way, whereas size-dependent enforcement disproportionately lowers the wages of the best employees.

The presence of two-sided heterogeneity—where heterogeneous labor is allocated to heterogeneous producers—implies that larger productive units exhibit a higher marginal product of labor relative to smaller ones. This positive relationship between size and marginal products is strongest when complementarities in production are not hindered by the presence of size-dependent distortions. That is, an undistorted economy with knowledge hierarchies exhibits a larger dispersion of marginal products, and a steeper relationship between size and marginal product, than an economy under size-dependent regulations. These results imply that some of the dispersion in marginal products observed in the data may be "good" dispersion—in the sense that plants differ in the skill composition of their workforce—and, in fact, in economies affected by size-dependent policies, there may be too little of it.

The rest of the paper is organized as follows. In the next section, we present an overview of our modified version of the hierarchical model of Garicano and Rossi-Hansberg (2004, 2006), and completely characterize a parametric example that serves as the basis of our analysis. We also describe the qualitative effects of different tax policies. Section 3 contains a description of the data used in our analysis—including a set of motivating facts about the Mexican economy that our model can address—as well as a description of our calibration strategy. The results from our counterfactual experiments for the U.S. and Mexico are reported in Section 4. Section 5 details the main takeaways from our analysis and Section 6 concludes.

2. Theory

The basic undistorted environment is the general equilibrium, continuous assignment model of Garicano and Rossi-Hansberg (2004, 2006), where not only managerial rents but also wages vary across workers, and which we modify in two ways. First, communication costs are not constant but rather decrease with the skill of production workers. Second, we subject the economy to decreasing returns to scale as in Lucas (1978). These extensions allow for a more realistic support for the distribution of plant sizes, which in turn facilitates the implementation of useful quantitative exercises. Further, the resulting environment serves as a useful—and familiar—setup to study the effects of distortions on the optimal sorting of workers and the return to skill.

We start by describing the general environment, as well as the algorithm to find the equilibrium. Next, we fully characterize a specific parametric example, which we will rely upon in our calibration and counterfactual policy experiments. Finally, we show the effects of a general payroll tax especially enforced on larger teams, and discuss the effects of size-dependent policies that potentially break positive sorting in a segment of the skill distribution.

2.1. The allocation of talent: knowledge hierarchies and decreasing returns

Individuals differ in a single trait—call it talent—and own one unit of time. They choose the use of their time and talent that maximizes their earnings. They can either produce alone or in a team (plant) with other workers. They work together to specialize either in managerial activities—running the establishment—or in production activities—working as one of the plant's employees—and thus exploit complementarities in production. Individuals have then three options: to produce alone, to manage others, or to work for a wage. The formation of teams in the model is also endogenous—the managers optimally select the size of their team, as well as the quality of their employees, whereas the wage workers optimally select which of the continuum of teams to join. Thus, our economy features establishments with two and one layers, and the latter represent the self-employed in the economy—entrepreneurs without employees.

To produce in this economy, workers solve problems which vary in their difficulty, z, according to some density $g(z)$. Skill is cumulative—a worker of skill z can solve all problems of difficulty less than or equal to z. Workers draw and attempt to solve one problem in their unit of time, and produce only if they know the answer. The (expected) earnings of worker z are therefore the percentage of problems he is able to solve: $G(z)$. The skill endowment z varies continuously in the population according to the (given) skill distribution $F(z)$, with support $[L, H]$—where L and H denote the lowest and highest levels of skill in the population—and density $f(z)$.

Agents can also form teams, each consisting of identical production workers and one manager. In these teams, the manager attempts to solve the problem whenever his production workers do not know the answer, and production workers do not interact with each other. More precisely, a team with n employees draws n problems, and the (expected) output of the team is the percentage of tasks that a manager with skill z_m is able to solve in his n units of time,

$$y = G(z_m)\, n^{\alpha}$$

where $\alpha \in (0,1)$ denotes the degree of decreasing returns to scale in the use of time, as in Lucas (1978).

Communication in teams is costly: communicating with a worker of skill z_p costs the manager a fraction $\tilde{h}(z_p)$ of his time, which occurs with probability $1 - G(z_p)$. Therefore, helping a worker of skill z_p requires the manager to spend $[1 - G(z_p)]\tilde{h}(z_p)$ of his time. We assume that $\tilde{h}'(z_p) < 0$. Thus, expected communication costs are decreasing in worker skill via two channels: communicating with a more skilled production worker takes less time, and more knowledgeable workers ask the manager for help less often.[2]

Communication costs then limit the entrepreneur's span of control, or team size, n. More precisely, define $h(z_p) \equiv [1 - G(z_p)]\tilde{h}(z_p)$. If wage workers are of quality z_p, a manager can coordinate at most $n(z_p) \equiv 1/h(z_p)$ with his unit of time. Notice that the potential size of the team $n(z_p)$ is then increasing in the quality of its members.

The problem of a manager with ability z_m is to choose the quality of his employees, z_p, and the size of his team n, so as to solve

$$\begin{aligned} R(z_m) = \max_{z_p, n}\ & G(z_m)\, n^{\alpha} - w(z_p)\, n, \\ s.t.\ \ & n \leq n(z_p), \end{aligned} \tag{1}$$

where $w(z_p)$ denotes the equilibrium wage rate for workers of quality z_p.

Without any distortion, the inequality constraint will always bind: from the manager's point of view, team size l is feasible hiring wage workers of any type above $n^{-1}(l)$, but costs are minimal when employing workers of skill $n^{-1}(l)$, who earn the lowest wage, and that is true for all team sizes. Sizes without distortions are therefore simply a function of the knowledge in the bottom layer of the team. Thus, we can rewrite output as

$$y(z_p, z_m) = G(z_m)\left[n(z_p)\right]^{\alpha}, \tag{2}$$

and managerial rents are thus

$$R(z_m) = \max_{z_p} G(z_m)\left[n(z_p)\right]^{\alpha} - w(z_p)\, n(z_p). \tag{3}$$

Individuals take wages and managerial rents as given, and choose the occupation that yields the highest earnings given their skill:

$$\max\{w(z), G(z), R(z)\}. \tag{4}$$

The equilibrium exhibits positive sorting in both quality and quantity, as in Garicano and Rossi-Hansberg (2004, 2006) and Garicano and Hubbard (2012). Specifically, the best managers match with the best wage workers to form the largest teams, the second-best managers match with the second-best wage workers to form the second-largest teams, and so on. The smallest teams in the market are thus the match of the least-skilled entrepreneurs to the least-skilled wage workers.[3] The equilibrium also exhibits perfect stratification of individuals into occupations based on their skill: the less skilled agents become production workers, those in the middle produce alone, while the most skilled ones work managing others.[4] More formally, a competitive equilibrium is defined as follows:

[2] In the original framework only the latter channel is present. More precisely, in the setup in Garicano and Rossi-Hansberg (2004), $\alpha = 1$ and $h(z_p) = b(1 - G(z_p))$. In that case,

$$\lim_{z_p \to H} n(z_p) = \infty,$$

and because output exhibits constant returns to scale, managers have incentives to set teams as large as possible by matching with the worker right next to them, which in turn can only occur by adding additional layers, and the model has no equilibrium. High communication costs—the value of b—is what prevents them from adding more layers, but high values of b significantly limit the support of the equilibrium distribution of team sizes. In our alternative setup the incentives to set large teams are diminished by $\alpha < 1$ and, in addition to bounding $n(z_p)$, there is a broader range of values of b which allow for an equilibrium to exist.

[3] The positive sorting of workers is guaranteed as long as the second-order sufficient condition of the entrepreneur's problem is satisfied.

[4] Unlike the equilibrium in the original framework, the equilibrium in our model could exhibit segregation of workers, as in Kremer and Maskin (1996), in which the top agents match together whereas workers at the bottom produce alone. Whether the equilibrium entails segregation or not depends on 1) the sensitivity of earnings to skill in the outside option (self-employment), 2) the skill distribution, and 3) the strength of complementarities in production (α). In our analysis we focus on the equilibrium without segregation.

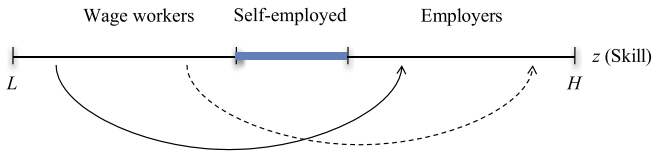

Fig. 1. Allocation of workers across occupations and teams.

Definition. Given a distribution of talent $F(z)$ over $[L, H]$, a distribution of problems $G(\tilde{z})$, a communication technology $\tilde{h}(z)$, and a production technology $y = G(z_m)n(z_p)$, a competitive equilibrium consists of (i) an assignment function sorting individuals into occupations and into teams; (ii) a wage function $w(z)$; (iii) managerial rents $R(z)$; and (iv) a pair of cutoffs $\{z_1 < z_2\}$, where $[L, z_1]$ is the set of wage workers, $[z_1, z_2]$ is the set of owners producing alone, and $[z_2, H]$ is the set of managers, such that: (E1) no agent desires to switch to another occupation or team; (E2) the supply of workers equals the demand for workers; (E3) the supply of managers equals the demand for managers.

Since the assignment function matches workers to managers and team sizes, conditions (E2) and (E3) are equivalent. Fig. 1 illustrates the equilibrium allocation of workers into occupations and into teams according to their skill level.

To find the equilibrium, we follow the algorithm in Sattinger (1993), as described by Garicano and Rossi-Hansberg (2004). We conjecture the existence of the two thresholds, z_1 and z_2, such that agents with skill $z \leq z_1$ optimally become wage workers, agents with skill $z_1 < z < z_2$ optimally choose to produce alone, and agents with skill $z \geq z_2$ optimally choose to manage others. The first step is to find the equilibrium assignment function $z_m = m\left(z_p\right)$, which denotes the manager z_m that corresponds to workers of skill z_p. In equilibrium the market for managers must clear: for all $s \in [L, z_1]$ it has to be the case that

$$\int_L^s \frac{f(x)}{n(x)} dx = F\left(z_m(s)\right) - F\left(z_2\right), \tag{5}$$

which implies

$$\frac{f(s)}{n(s)} = f\left(m(s)\right) \frac{dm}{ds}. \tag{6}$$

We find the assignment function $m\left(z_p\right)$ by solving the differential equation in (6), with boundary condition $m(L) = z_2$—which states that the worst wage workers are matched with the worst employers.

The second step is to find the equilibrium wage function $w\left(z_p\right)$. The manager's problem is to select z_p to maximize

$$R(z_m) = G\left(z_m\right)\left[n\left(z_p\right)\right]^{\alpha} - w\left(z_p\right) n\left(z_p\right),$$

with corresponding FOC

$$G\left(z_m\right)\alpha\left[n\left(z_p\right)\right]^{\alpha-1} \frac{dn}{dz_p} = \frac{dw}{dz_p} n\left(z_p\right) + w\left(z_p\right) \frac{dn}{dz_p}. \tag{7}$$

We find the wage profile by substituting $z_m = m\left(z_p\right)$ from step one above into equation (7) and solving the resulting differential equation for $w\left(z_p\right)$ with boundary condition $w\left(z_1\right) = G\left(z_1\right)$—which states that the best wage workers are indifferent between working for a wage or producing alone.

The third and final step is to pin down the equilibrium cutoffs z_1 and z_2 to completely characterize the equilibrium assignment function $z_m = m\left(z_p\right)$, and the equilibrium earnings functions $w(z)$ and $R(z)$. We do so by solving the two remaining equilibrium conditions, $m\left(z_1\right) = H$ and $R\left(z_2\right) = G\left(z_2\right)$, which state that the best managers match with the best workers, and that the worst managers are indifferent between running a team and producing alone. If, as conjectured in the algorithm, it is the case that $w(z) \geq G(z)$, and $w(z) \geq R(z)$, for all $z \in [L, z_1]$, $G(z) \geq w(z)$ and $G(z) \geq R(z)$ for all $z \in [z_1, z_2]$, and $R(z) \geq G(z)$ and $R(z) \geq G(z)$ for all $z \in [z_2, H]$, the conjecture is correct and the allocation is an equilibrium.

The two cutoffs $\{z_1, z_2\}$ determine the equilibrium allocation of talent, that is, the sorting of workers into occupations, as well as the assignment of wage workers to managers. In turn, the allocation of talent determines team sizes, earnings, and aggregate output in this economy. Therefore, policies that distort the allocation of talent have a rippling effect, and may result in significant aggregate losses even if the effect on z_1 or z_2 is seemingly negligible.

The skill distribution, the span of control parameter α, and the costs of communication in teams determine the earnings distribution and the establishment size distribution in equilibrium. For example, when low-skill labor is relatively more abundant, wages are lower and managerial rents are higher (because talent is more scarce), which results in a larger share of managers. Further, teams are smaller on average, because managers spend more time communicating with their employees (low-skill workers ask questions more often). The span of control parameter α governs the strength of complementarities

in production. When α increases, the returns from matching in larger teams increase, and therefore more workers choose to form teams and fewer workers choose to produce alone. Teams are larger in equilibrium, and the earnings of both wage workers and managers increase. Finally, when communication costs are higher, team sizes are smaller, but the effect on earnings will now depend on whether the sensitivity of costs with respect to skill is higher or lower, which we carefully explain using our parametric example.

2.2. Parameterization of the undistorted environment

We now fully characterize a specific parametric example, which we will rely upon in the calibration and counterfactual policy experiments that follow. Our choice of functional forms and assumptions deliver an environment that is both analytically tractable and suitable for the implementation of useful quantitative exercises.

The population skill distribution, $F(z)$, is assumed to be a double-truncated exponential distribution over the interval $[L, H]$, with parameter λ, that is,

$$F(z) = \frac{\exp[-\lambda L] - \exp[-\lambda z]}{\exp[-\lambda L] - \exp[-\lambda H]}.$$

Larger values of λ result in a relative abundance of low skill labor. We assume that the distribution of problems, $G(z)$, is uniform in the interval $[L, \tilde{H}]$, with $\tilde{H} \geq H$. That is, the most difficult production problem is potentially unsolvable. Communication costs take the form

$$\tilde{h}(z) = \frac{a \exp[b[1 - G(z)]]}{1 - G(z)},$$

where $a, b > 0$ govern both the level and the sensitivity of costs to changes in the skill of wage workers. Expected communication costs are thus

$$\begin{aligned} h(z) &= a \exp[b[1 - G(z)]], \\ &= a \exp\left[b\left(\frac{\tilde{H} - z}{\tilde{H} - L}\right)\right]. \end{aligned}$$

Potential team sizes follow then

$$n(z) = \frac{1}{a} \exp\left[-b\left(\frac{\tilde{H} - z}{\tilde{H} - L}\right)\right].$$

Notice that $n'(z) > 0$ and $n''(z) > 0$. That is, matching with more talented wage workers allows the manager to increase more than proportionally the size of his team. However, decreasing returns in team sizes ($\alpha < 1$) rein in the incentives to form arbitrarily large teams.

Fig. 2 displays the equilibrium assignment, occupational choices, and earnings in a hypothetical version of this economy where parameter values are such that the assignment function is linear. The wage profile is an increasing and concave function of skill—which is a result of decreasing returns to the span of control of managers—whereas managerial rents are increasing and convex—because higher-skill managers can leverage their knowledge by forming larger teams comprised of better workers.

The comparative statics of $\tilde{H}$, b, and a in our model economy are in line with a simple framework of supply and demand for varying levels of skill. An increase in the maximum difficulty of tasks ($\tilde{H}$) translates into a higher level of communication costs, but a lower sensitivity with respect to skill, as seen in Panel A of Fig. 3. Workers in this economy now possess skills that are not as useful to solve the set of available production problems. This shock results in smaller potential team sizes, and lower earnings for wage workers, the self-employed, and managers. Conversely, an increase in b translates into both a higher level of communication costs and a higher sensitivity with respect to skill, as shown in Panel B of Fig. 3. Team sizes are smaller, but the effect on wages now depends on whether the worker is low or high skill. Skilled wage workers become comparatively more attractive because they ask the manager for help less often. In other words, the demand for high-skill wage workers is higher relative to their low-skill counterparts. Wages at the bottom decrease whereas wages at the top increase, that is, the skill premium for wage workers is higher. Last, Panel C of Fig. 3 shows that an increase in a yields the same qualitative effects as an increase in b: the level of communication costs and the sensitivity with respect to skill are higher. In this case, however, the magnitude of the effects is significantly lower, because a enters the communication cost function linearly, whereas b does so exponentially.

2.3. The allocation of talent under size-dependent regulations

We examine the effects of a general payroll tax whose enforcement increases with team size. More precisely, we assume that the effective tax rate can be summarized by a function $\tau(n)$, with $\tau'(n) > 0$. Under this policy, the manager's optimization problem is now:

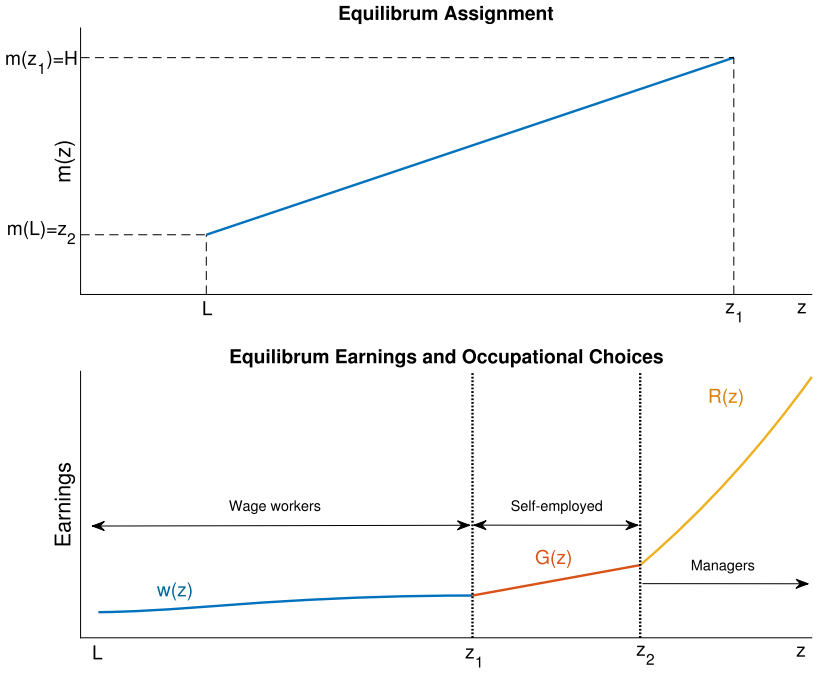

Fig. 2. Equilibrium assignment and earnings under parametric assumptions.

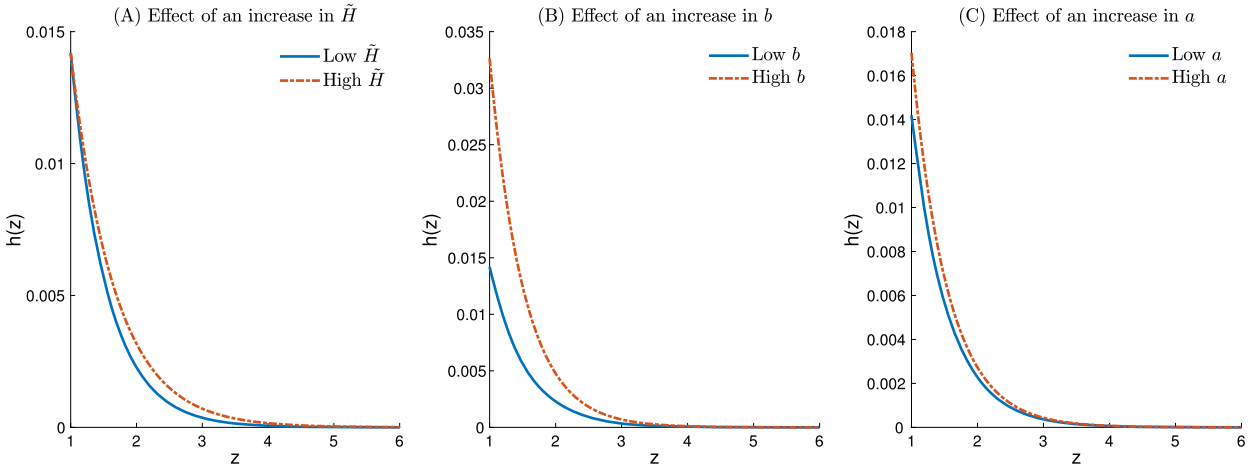

Fig. 3. Effect of changes in $\tilde{H}$, b, and a on the relationship between skill and communication costs.

$$\max_{z_p,n} G\left(z_m\right) n^{\alpha} - \left[1+\tau\left(n\right)\right] w\left(z_p\right) n, \text{ s.t. } n \leq n\left(z_p\right),$$

where $n(z_p) \equiv 1/h(z_p)$, as before. The tax policy we consider takes the following form

$$\tau\left(n\right) = \tau_s\left[1-\exp\left(-\kappa n\right)\right], \tag{8}$$

where $\tau_s \in [0,1)$ is the statutory tax rate, and $\kappa \geq 0$ is an enforcement parameter. The parameter τ_s is a level shifter, whereas the enforcement parameter κ affects the steepness of the tax policy, with $\tau(n)$ approaching τ_s as κ becomes arbitrarily large. Further, the size-dependent policy assumes perfect enforcement of the statutory tax on large establishments. That is, this policy amounts to a proportional payroll tax of τ_s for establishments of sizes above some threshold, and this size threshold is lower the larger the value for κ.[5] Fig. 4 plots this tax function for different parameter values.

Because $\tau''(n) < 0$ and $n''(z_p) > 0$, the size constraint in the manager's problem always binds. Then, we can write the first order condition to the manager's problem as

[5] This type of negative exponential tax functions can be used to model progressive taxation, as in López (2020), and tax evasion, as in López (2017).

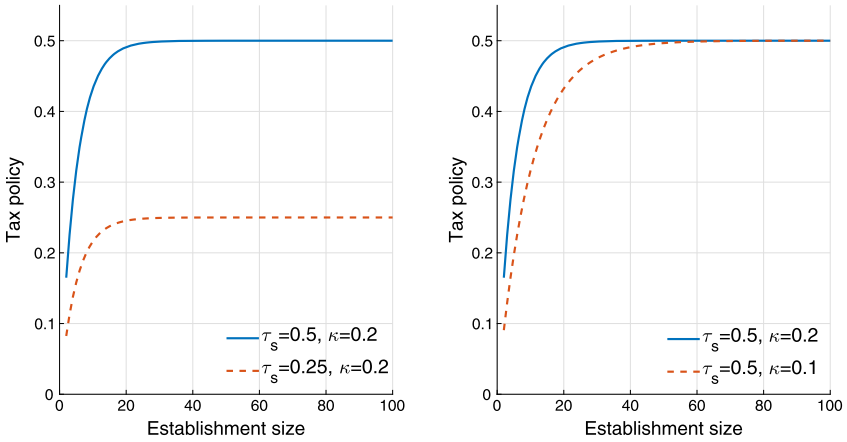

Fig. 4. Size-dependent tax policy.

$$G(z_m)\alpha[n(z_p)]^{\alpha-1}n'(z_p) = \left[w'\left(z_p\right)n\left(z_p\right) + w\left(z_p\right)n'\left(z_p\right)\right]\left[1+\tau\left(n\left(z_p\right)\right)\right] + w\left(z_p\right)n\left(z_p\right)\tau'\left(n\left(z_p\right)\right)n'\left(z_p\right). \tag{9}$$

The algorithm to find the equilibrium under this type of size-dependent policy is similar to the one described above for the undistorted case. The labor market clearing condition in equation (6) used to derive the assignment function $z_m = m(z_p)$ remains unchanged, including the boundary condition $m(L) = z_2$. Once we solve for $m(z_p)$, we plug it into our first order condition in equation (9), and solve the differential equation for $w(z_p)$, using the indifference condition $w(z_1) = G(z_1)$ as the boundary condition. The final two equilibrium conditions used to pin down the cutoffs $\{z_1, z_2\}$ are $m(z_1) = H$—which states that the best manager matches with the best workers—and $R(z_2) = G(z_2)$—which states that the worst manager is indifferent between running a team and being self-employed without employees.[6]

2.3.1. Occupational choices

The size-dependent tax creates an incentive to form smaller teams, which managers can achieve by matching with workers of lower quality—who will ask questions more often—compared to the undistorted equilibrium, thus lowering the demand for high-skill wage workers. As a result, the best wage workers in the undistorted equilibrium become self-employed after the tax, which in turn decreases the average team size. A smaller mass of wage workers requires a smaller mass of managers, and in the bid for talent only the most skilled managers will find a match. In other words, clearing in the market for managers occurs from top to bottom, and so the worst managers in the undistorted equilibrium turn to self-employment as a result of the tax.[7]

In short, the size-dependent tax only reallocates individuals in the middle of the skill distribution into another occupation, but does not distort the occupational choices of the least and the most talented individuals in the economy. However, by changing the marginal individuals in each occupation—z_1 decreases, whereas z_2 increases—the size-dependent tax resorts everyone within occupations, weakening the positive sorting in this economy, as we illustrate in Fig. 5.

2.3.2. Team assignments

The size-dependent tax lowers the average quality of wage workers and increases the average quality of managers, wasting managerial talent on wage workers of lower skill. Fig. 6 shows the effects in a hypothetical economy of two size-dependent taxes with the same level of enforcement, but different statutory tax rates. As the tax increases, the quality of the best worker, and therefore the size of the largest team, decreases, and as a result wage workers are matched with more talented managers relative to the undistorted environment (alternatively, managers are matched with less talented wage workers).

Notice that the assignment of wage workers to managers is a many-to-one matching problem, so that a small decrease in the share of managers is consistent with a large decrease in the share of wage workers, especially because this decrease occurs in the right tail of the skill distribution of wage workers (who form the largest teams). For this reason, in Fig. 6, a decrease of 19% in z_1 is consistent with an increase of 0.59% in z_2 (which is visually unnoticeable).

[6] The fact that the algorithm remains the same means that, starting from a distorted benchmark, we can derive closed form solutions for all equilibrium objects using a distorted version of our parametric example (details are provided in Appendix A), which is particularly useful in the quantitative exercises in the next section.

[7] Note that the wage workers and managers that switch into self-employment were not working together before the tax: the wage workers used to work for the best managers, whereas the managers used to work with the worst wage workers.

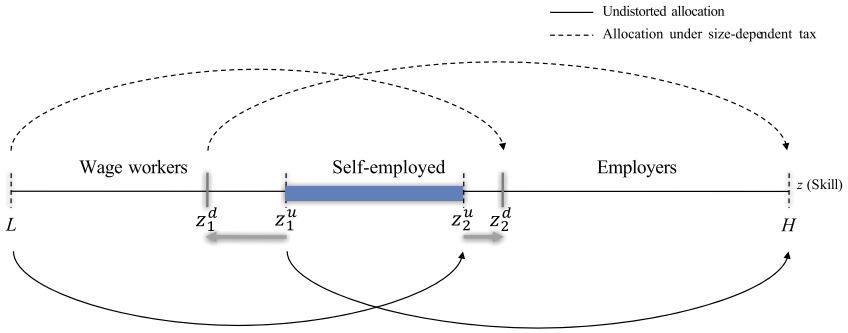

Fig. 5. Talent reallocation as a result of the size-dependent tax.

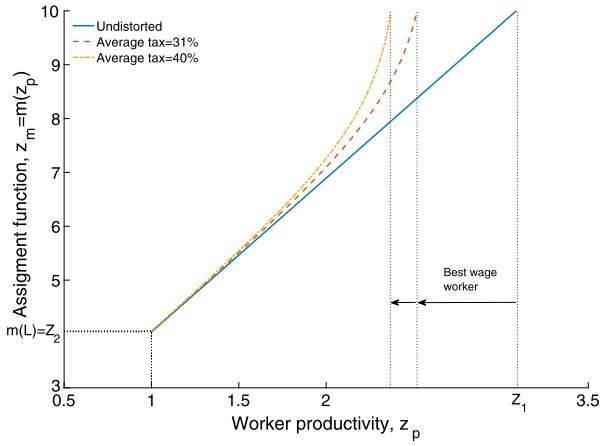

Fig. 6. Effect of size-dependent policies on equilibrium assignments.

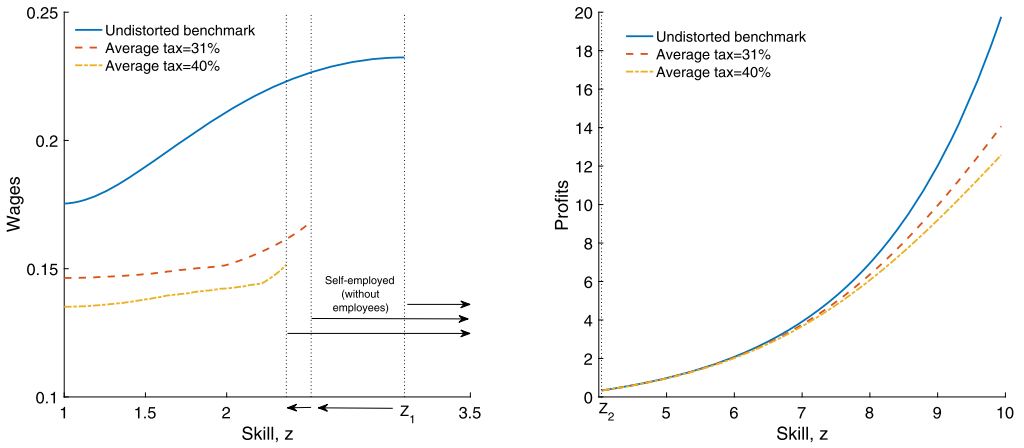

Fig. 7. Effect of size-dependent policies on equilibrium earnings.

2.3.3. *Earnings*

Fig. 7 shows the effects on earnings of the two tax policies used in Fig. 6. The tax reduces the demand for skill, which depresses wages—thus pushing the most talented wage workers into self-employment. The wage profile becomes flatter on average because the earnings of the best wage workers (now of lower skill) are necessarily lower. In addition, the earnings of the least skilled wage workers are lower as well—under positive sorting, market forces act from top to bottom, and therefore the trickle-down effect necessarily results in lower wages for workers at the bottom when the wages of those at the top decrease.

The managerial rents profile is flatter as well because managers are matched to lower quality wage workers in smaller teams. Managerial rents at the top of the skill distribution are lower because the effective cost of labor (wages plus payroll

taxes) is markedly higher relative to the undistorted economy. Managerial rents at the bottom are only marginally lower: the least talented managers in the undistorted environment turned into self-employment but, because the assignment is many-to-one, the effect on the marginal employer is relatively small.

In short, in small teams the earnings of wage workers are significantly lower but the earnings of managers remain more or less the same, whereas in large teams both wage workers and managers share the burden of the tax.

2.3.4. Knowledge hierarchies and the "bunching" of teams

The size-dependent distortion in equation (8) preserves the qualitative properties of the undistorted equilibrium and generates a continuous establishment size distribution. When the size-dependent policy results in a "bunching" of plants around a particular threshold N in the equilibrium size distribution, then positive sorting in a segment of the skill distribution is not guaranteed, which aggravates both the losses from talent misallocation and the effect of the policy on the return to skill.

As a simple illustration, consider the step tax policy examined by Garicano et al. (2016), where all establishments above a given size threshold $N > 0$ have to pay the same tax, but pay zero if they stay below the threshold.[8] Now the size constraint $n \leq n\left(z_p\right)$ will not bind for managers in a segment above the threshold N. That is, the tax will encourage employers within a range of the size threshold N to constrain the size of their team to avoid the tax, discouraging their wage workers from fully exploiting their talent. If the size constraint is binding, prices and assignments in equilibrium are solved for as in the undistorted case; if the size constraint is not binding, then optimal sizes will depend only on managerial skill, just as in a Lucas (1978) span-of-control model. We show how to solve for the equilibrium assignment and the wage profile under this policy in Appendix A.

The solution now requires wages to be independent of skill if the size constraint is not binding—if managers constrain the size of their team to avoid the tax, then the wages paid to their workers must be independent of their skill (that is, wages in this segment equal a constant). Moreover, positive sorting in the constrained teams who constrain their size is not guaranteed: positive sorting is one of infinitely many possible outcomes among constrained teams and workers. This segment of the assignment function is therefore indeterminate. Managers could potentially mix employees from different types in a single team, all of them earning the same wage.

The step tax policy unambiguously lowers the return to skill for workers in the constrained teams. Managers in these teams reduce the size of their plant to avoid the tax, but by doing so wage workers lose the ability to differentiate themselves, and therefore their returns to skill turn flat, whereas managers lose the reward for pairing with more skilled employees, which also lowers the return on their skill (their earnings are now linear in skill). Policies that produce a bunching of plants in the size distribution then only aggravate the misallocation of talent and its effects on both productivity and the return to skill.

2.3.5. Improvements in the enforcement technology

When κ is high enough, the size-dependent tax amounts to a proportional tax on the right tail of the distribution of establishments. Managers and wage workers at the top are therefore indifferent to more stringent enforcement—increases in κ. Indeed, moving from size-dependent to perfect enforcement (when κ is initially high enough and keeping the statutory rate constant) will have no effect on the production decisions in the largest teams. Under positive sorting, market forces act from top to bottom. Consequently, the trickle-down effect that causes everyone to re-sort and reallocate within and across occupations is absent if the managers of the largest teams do not re-optimize their decision, and thus increasing κ or moving from size-dependent to perfect enforcement—maintaining the statutory rate constant—has no effect on plant sizes, output, or top wages.

However, when we move from size-dependent enforcement to a perfectly-enforced tax, the situation for smaller plants changes, as they now face a higher cost of labor. Because no other margins adjust to the policy change, the burden of the higher effective tax is borne entirely by low-skill wage workers, who see their earnings decrease. That is, the earnings of wage workers at the bottom are lower when the payroll tax is perfectly enforced, but the earnings of wage workers at the top are identical with or without size-dependent enforcement (as long as κ is initially high enough). As a result, the returns to skill for wage workers under perfect enforcement are higher relative to the environment with the size-dependent tax (since the top wages are identical under both policies), and may not significantly differ from the returns to skill in the undistorted environment.

This result suggests that governments that care about earnings inequality, but lack the resources for perfect tax enforcement and redistribution, may prefer size-dependent policies over flat taxes (either statutorily or through enforcement): under size-dependent labor taxes, workers of different skills earn more equal wages compared to a situation where the labor tax is flat and perfectly enforced.

[8] Because some of the size-dependent regulations studied by Garicano et al. (2016) are subject to different definitions of size, the step function is not a precise characterization of the tax policy. One way to model this complication is to consider a Sigmoid tax with midpoint N, with a steepness that can be adjusted to match the observed level of enforcement or compliance, or certain characteristics of the observed "bunching" in the establishment or firm size distribution. That is the size-dependent policy we consider in Appendix A. Note that when the steepness parameter becomes arbitrarily large, the Sigmoid policy converges to the step tax policy examined in Garicano et al. (2016).

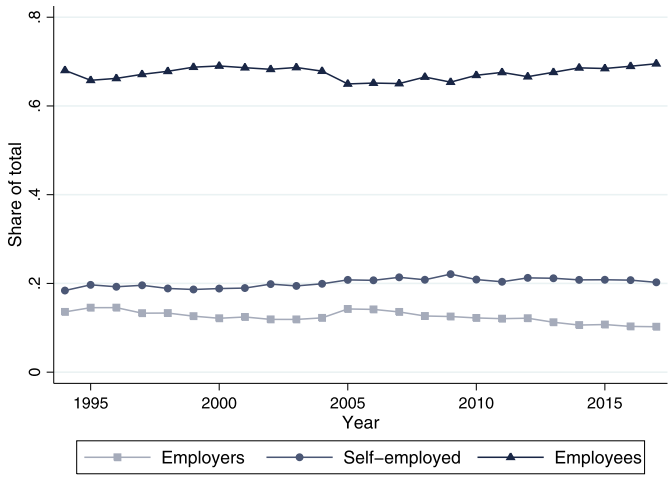

Fig. 8. Distribution of workers across occupations in Mexico. 1994-2017.

In short, when κ is initially high enough, a more stringent enforcement of the statutory tax results in higher returns to skill without significant effects on the establishment size distribution and aggregate output. In economies with positive sorting in both quality and quantity, increases in κ combined with reductions in the statutory tax rate τ_s is a more effective tool to boost both aggregate output and the returns to skill.

3. Data and calibration

In this section, we describe the data used to discipline our analysis, and provide the details of our calibration. We start by highlighting the stark differences between the establishment size distributions and occupational shares in the U.S. and Mexico. We also provide a set of motivating facts about the Mexican economy that we address in our analysis: persistently-high levels of business ownership (especially self-employed), a proliferation of small businesses, a growing mismatch between the supply and demand for skill, and high statutory labor costs that are more stringently-enforced on larger establishments. Last, we discuss the calibration results and their implications for the observed differences between the U.S. and Mexico.

3.1. Data

Our quantitative analysis uses establishment-level as well as individual-level data from the U.S. and Mexico. Appendix B describes our data in detail. We measure size using the number of employees (without including the self-employed in our calculations). For the distribution of plant sizes in the U.S., we rely on tables from the Longitudinal Business Database for the year 2016, published by the U.S. Census Bureau. For the distribution of establishments in Mexico, we use micro-data for manufacturing, retail and wholesale, and services sectors, which we obtain from the Economic Census conducted by Mexico's National Institute of Statistics (INEGI, by its Spanish acronym).[9]

We use individual-level data to estimate occupational shares in both countries. For the U.S., we rely on calculations by the Bureau of Labor Statistics (BLS), as reported in Hipple and Hammond (2016), to determine the shares of wage workers (89.9 percent), non-employers (7.6 percent), and employers (2.5 percent).[10] We estimate occupation shares in Mexico using the National Survey of Occupations and Employment and the National Survey of Urban Employment (ENOE and ENEU, by their Spanish acronyms). The resulting shares for Mexico are 70 percent for wage workers, 20 percent for non-employers, and 10 percent for employers. Thus, around 30 percent of workers in Mexico run their own business—three times the entrepreneurship rate observed in the U.S. (U.S. Bureau of Labor Statistics). The fraction of non-employers, in particular, is one of the highest among OECD countries (OECD, 2017) and, as Fig. 8 shows, has remained more or less stable since 1994.

Table 1 contains descriptive statistics on the distribution of establishment sizes, as well as occupation shares in the U.S. and Mexico. Micro-businesses make up for most of Mexican plants, a feature that has also remained stable for at least the past two decades (Busso et al., 2018). The average establishment in Mexico employs 5.65 workers, and more than 9 out of

[9] The Economic Census conducted by INEGI surveys all establishments with a fixed physical location in municipalities with more than 2,500 inhabitants. A thorough description of the establishment-level data from INEGI used in our analysis can be found in Busso et al. (2018). Self-employment in Mexico then is not entirely captured by the Economic Census, since a large portion of the self-employed do not have a business address (they work from home, or are street vendors).

[10] The BLS estimate for the share of wage workers is identical to our own calculation using the March supplement of the CPS.

Table 1
Establishment size and occupation shares in the U.S. and Mexico.

Average establishment size	U.S.	Mexico
	17.72	5.65
Share of establishments of size $1-4$	49%	85.6%
Share of establishments of size $5-9$	21.2%	7.7%
Share of establishments of size $10-19$	14.3%	3.5%
Share of establishments of size $20-49$	9.7%	2%
Share of establishments of size ≥ 50	5.8%	1.3%
Share of wage workers	90%	70%
Share of non-employers	7.6%	20%
Share of employers	2.4%	10%

Note: Size is measured using the number of employees. The statistics on the size distribution of plants do not include non-employers. Data sources: US – Longitudinal Business Database for 2016; Mexico – Establishment Census for 2013. See Appendix B for details.

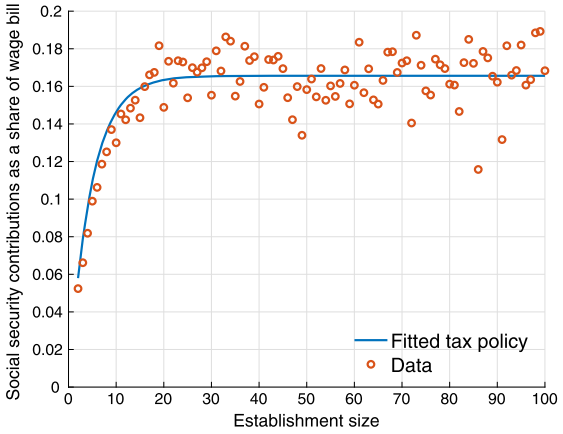

Fig. 9. Average social security contributions as a fraction of wages in fixed establishments in Mexico.

10 plants have less than 10 employees. In contrast, the average U.S. establishment employs 17.72 workers, and 70 percent of U.S. establishments have less than 10 employees.

The last 20 years in Mexico have also seen a growing mismatch between the supply and demand for skilled labor: while the amount of available skilled workers has grown, years of schooling average 9.1, barely above the 9 years of mandatory schooling.[11] Moreover, the earnings of more educated workers have decreased (Levy and López-Calva, 2016), which suggests the presence of distortions that reduce the demand for, and hence, the returns to skill. This stands in sharp contrast with the evidence for advanced economies, such as the U.S., where the skill premium is often described as the result of supply not being able to keep up with the demand for skill.

In economies where workers sort positively in quality and quantity, size-dependent distortions disrupt the allocation of talent increasing self-employment, which results in smaller teams and lower returns to skill. Labor regulations in Mexico are not only stringent, but their enforcement displays a very clear size-dependent quality. According to Alaimo et al. (2017), labor regulations in Mexico amount to 44% of the average wage of a formal employee: 18 percentage points from employer contributions to social security (which covers pension benefits and the provision of medical services), 5 percentage points from contributions to a housing fund (which offers mortgages at relatively low rates), and the rest from required days of paid leave, some dismissal restrictions, and a mandatory yearly bonus. Additionally, Busso et al. (2018) document evidence of two size-dependent distortions in Mexico that are directly tied to labor costs: smaller establishments employ a higher fraction of non-salaried workers (which are exempt from complying with labor regulations), and pay on average a lower fraction of social security contributions for their salaried employees. The latter distortion is displayed in Fig. 9, which shows the relationship between establishment size (number of employees) and the effective social security tax as reported in the Mexican establishment census, along with estimates of the tax policy described in equation (8). The size gradient is more pronounced along medium-sized establishments and flattens out as plant sizes increase.

We do not observe the relationship between compliance and plant size for the remaining labor regulations, but in what follows we assume that they share the same enforcement parameter κ to examine the effects of size-dependent enforcement of labor regulations using our hierarchical model with decreasing returns.

[11] Source: Mexico's National Institute of Statistics (INEGI, by its Spanish acronym).

3.2. Calibration strategy

We conduct separate calibrations for the U.S. and Mexico, allowing key parameters in our model to differ across countries. This is important considering the vast differences in the occupational shares as well as the distribution of establishment sizes observed in these two economies. These differences are due to a number of potential reasons, size-dependent regulations being only one of them. For instance, communication technology between production workers and managers might be better in the US than in Mexico, and skills might be distributed differently in the two countries.[12]

Our model has seven parameters: $\Phi \equiv \left\{L, \lambda, H, \tilde{H}, a, b, \alpha\right\}$. In what follows we impose two restrictions that allow the problem to be solved and calibrated in a relatively straightforward way. While we conduct a joint calibration of the key parameters of the model, we also provide an intuitive discussion on how the moments we choose are related to different parameters.

Assumption 1. $\tilde{H} = H$. That is, the hardest production problem is solvable by the most skilled workers in the economy.

Assumption 2. Define

$$\gamma \equiv \frac{a\lambda\left[H - L\right]}{\lambda\left(H - L\right) + b}.$$

Assume

$$\gamma \exp\left[\lambda\left(z_2 - L\right) + b\right] = 1. \tag{10}$$

Assumption 1 is without loss of generality, and is consistent with the work of Geerolf (2017), who also studies the equilibrium distribution of teams under hierarchical matching. Under Assumption 2, the equilibrium assignment is linear, and closed form solutions exist for all differential equations. Further, we can analytically solve for the equilibrium cutoffs $\{z_1, z_2\}$, and the equilibrium distribution of plant sizes is a double-truncated Pareto distribution with power $1 + {}^{\lambda(H-L)}/_{b}$. All derivations are provided in Appendix A.

The Pareto team-size distribution is the result of positive sorting in a purely static environment, and stands in contrast with previous work where Pareto size distributions arise from either assuming that some primitive distribution is itself Pareto, or from a long sequence of large and persistent positive shocks, investment in managerial skills, or skill-biased change in entrepreneurial technology (Lucas, 1978; Luttmer, 2007; Bhattacharya et al., 2013; Guner et al., 2018; Poschke, 2018). Our result for the power of the distribution is consistent with the findings in Rossi-Hansberg and Wright (2007), who show that the right tail of the distribution of firm sizes in the U.S. is much closer to the unit Pareto benchmark than the establishment size distribution—which is in fact steeper (> 1) in absolute terms.

We emphasize that successfully matching the disproportionately large employment shares at the top observed in the U.S. would require either missing the average establishment size using our current model, or a model with more than two layers. The former is a result of the Pareto shape obtained for the *entire* distribution of team sizes: a better match of the right tail would force us to miss the average establishment size. A model with more than two layers would take us away from the traditional span-of-control setup, and would preclude us from analyzing the role of self-employment—two features that we find particularly important to frame and understand the effects of size-dependent policies. Moreover, we are not able to observe higher layers of knowledge in the data. The model can potentially be more successful at matching employment shares in Mexico precisely because the establishment size distribution—which is not markedly different from the firm size distribution—exhibits a thinner tail compared to the U.S.[13]

To exploit the analytical and computational convenience of a closed-form solution, one parameter has to be left free to adjust so as to satisfy Assumption 2. Because a and b have similar roles in the communication technology, but changes in b lead to much larger changes in costs—a feature we would like to exploit in our calibration—we choose a as the parameter left to adjust in equation (10). Last, we fix the location parameter L, which denotes both the lowest level of skill in the population and the difficulty of the easiest production problem.

In the case of the U.S., we calibrate the benchmark undistorted economy. In the case of Mexico, we exploit the fact that closed-form solutions exist in a distorted version of the benchmark economy described above to conduct our calibration (see Appendix A for these closed-form solutions in the distorted economy). Our size-dependent distortion follows equation (8). We use the estimates from Alaimo et al. (2017) on the cost of labor in Mexico, and set the tax on labor τ_s at 44 percent. We use $\kappa = 0.2154$, which matches the evidence on the size-dependent compliance of social security contributions shown in Fig. 9 above.

The four parameters in our joint calibration are then $\Phi^c \equiv \{\lambda, H, b, \alpha\}$. We calibrate the model by minimizing the sum of absolute deviations between four empirical moments and their model-generated counterparts, letting a adjust so that

[12] Moreover, there are forces outside our model that are likely to play important roles in shaping the observed differences between the two countries.

[13] In Mexico, only around 3% of firms are multi-establishment (INEGI).

Table 2
Calibration targets for U.S. and Mexico.

U.S.	Mexico
Average establishment size	Average establishment size
Share of non-employers	Share of non-employers
Share of establishments of size < 20	Share of establishments of size ≤ 5
Share of establishments of size ≥ 50	Share of establishments of size > 10

Table 3
Calibrated parameters for the U.S. and Mexico.

Parameter	Definition	US	Mexico
L	Lower bound for $G(\cdot)$ and $F(\cdot)$	1	1
H	Upper bound for $G(\cdot)$ and $F(\cdot)$	10.03	7.78
λ	Exponential parameter (skill dist.)	0.989	0.993
a	Communication cost (linear)	8.96×10^{-09}	5×10^{-10}
b	Communication cost (exponential)	16.57	20.71
α	Returns to scale	0.789	0.735

Assumption 2 is satisfied. More precisely, let $M(\Phi^c)$ denote the vector of model-generated moments and M^E the vector of empirical moments. Our calibration solves the problem

$$\min_{\Phi^c} \sum_{i=1}^{4} \frac{|M_i^E - M_i(\Phi^c)|}{|M_i^E|},$$

Table 2 contains the targets chosen for the U.S. and Mexico. In both countries, we try to match the average establishment size (measured using the number of employees and without including the self-employed), as well as the share of non-employers. Rather than focusing on fixed size bins for both countries, we choose the remaining two moments to capture similar sections of the establishment size distribution: for the U.S., we choose the share of establishments with less than 20 employees (84%), as well as the share of establishments with more than 50 employees (5.8%), whereas for Mexico, we choose the share of plants with 5 employees or less (88%), as well as the share of plants with more than 10 employees (6.7%).

In Appendix A we report the elasticity of each target moment to exogenous changes in the calibrated parameters. Moments in general are more responsive to changes in α and b, while the shape of the skill distribution λ and the difficulty of tasks H mainly affect only the right tail of the size distribution and the employment share of these plants.

The share of the self-employed is highly responsive to both α and b, whereas the share of wage workers does not significantly respond to changes in these parameters. The span of control parameter α governs the incentives to match in teams, and affects more strongly the right tail of the distribution and the employment share of these plants (relative to the left tail). Larger values for b (which increase both the level and the sensitivity of communication costs to skill) increase the share of employers and reduce the share of the self-employed, and moments from the left tail are much more responsive to changes in this parameter (relative to the elasticity of moments from the right tail).

3.3. Calibration results

Table 3 shows the parameter values obtained from our calibration exercise. The lower bound for skills and problems is set to one, which implies that the easiest production problem and the lowest skill level are the same for both countries. The upper bound for skills and problems is higher for the U.S., which is consistent with better workers and more difficult production problems in the U.S. than in Mexico. The skill parameter λ is nearly identical for both countries. Thus, the resulting skill distribution in the U.S. benchmark has a longer tail than that in the benchmark for Mexico, but lower skill levels are distributed more or less similarly across countries. The calibrated values for a and b imply that communication costs are higher in Mexico than in the U.S., as shown in Fig. 10. Last, the returns to team sizes, α, is higher in the U.S. than in Mexico. Higher communication costs and lower returns to team size imply that the span of control of managers in Mexico is more limited compared to their U.S. counterparts.

Tables 4 and 5 report the results from the independent calibrations for the U.S. and Mexico. The model matches perfectly the average establishment size (which does not include non-employers), and the share of non-employers in both countries. The model slightly underestimates both tails of the distribution, which implies that it generates more establishments in the middle than those observed in the data for both the U.S. and Mexico. In terms of non-targeted moments, the model replicates well the share of wage workers in both countries, as well as other non-targeted segments of the establishment size distribution. In the case of the U.S., the model does not capture well the distribution of employment across different size categories. In the case of Mexico, the model is more successful, particularly at capturing the employment share of the smallest establishments in the economy. Given the analytical simplicity of the model and the limited number of parameters available for the calibration, we find this performance satisfactory.

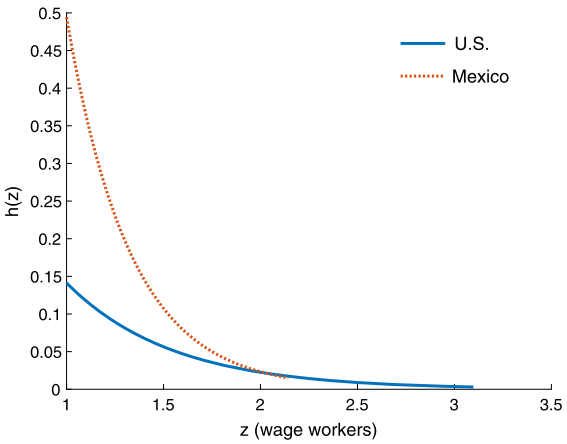

Fig. 10. Communication costs in the U.S. and Mexico calibrations.

Table 4
Calibration results for the undistorted benchmark: U.S.

Targeted	Data	Model
Average establishment size	17.72	17.72
Share of non-employers	0.076	0.076
Share of establishments of size < 20	0.85	0.80
Share of establishments of size ≥ 50	0.058	0.047
Non-targeted	Data	Model
Share of wage workers	0.90	0.87
Share of employers	0.024	0.054
Share of establishments of size > 100	0.025	0.014
Share of establishments of size > 250	0.007	0.002
Emp. share of est. < 20	0.25	0.49
Emp. share of est. ≥ 50	0.58	0.25

Table 5
Calibration results for the distorted benchmark: Mexico.

Targeted	Data	Model
Average establishment size	5.65	5.65
Share of non-employers	0.20	0.20
Share of establishments of size ≤ 5	0.88	0.77
Share of establishments of size > 10	0.067	0.097
Non-targeted	Data	Model
Share of wage workers	0.70	0.68
Share of employers	0.10	0.12
Share of establishments of size > 20	0.033	0.039
Share of establishments of size > 30	0.022	0.018
Emp. share of est. ≤ 5	0.36	0.38
Emp. share of est. > 10	0.58	0.40

The model captures well other non-targeted moments and relationships observed in the data. In the U.S. benchmark, the largest establishment size is 333 workers. In the data, the average size of establishments with more than 100 employees (top 2.5%) is 309 workers. Similarly, in the Mexico benchmark, the largest plant size is 62 employees. In the data for Mexico, the average size of establishments with more than 10 employees (top 6.7%) is 61 workers. More importantly, the model satisfactorily replicates the relationship between wages (relative to average) and plant sizes observed in Mexico, as seen in Fig. 11.[14] We find this last result to be a particularly important test of the model, considering there is a well-documented positive relationship between establishment size and wages, which has remained absent in the literature on misallocation that treats production workers as a homogeneous input.

[14] Our microdata from Mexico allows us to estimate mean wages for every employment size category, without resorting to bins. Unfortunately, we do not have access to the same data for the U.S. and, while there are some data available based on surveys like the CPS, that information is more likely to capture firm rather than establishment size.

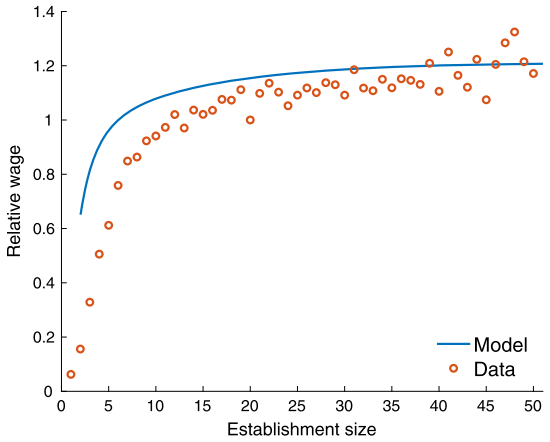

Fig. 11. Establishment size and relative wages in Mexico.

Table 6
The losses from size-dependent enforcement of labor regulations.*

	(1) Tax policy from Mexico ($\tau_s = 0.44$, $\kappa = 0.0658$)	(2) Perfect enforcement of average tax ($\tau_s = 0.31$, $\kappa = \infty$)
Share of wage workers	0.890	0.923
Share of non-employers	2.275	1.886
Share employers	0.989	0.994
Avg. est. size	0.900	0.929
Output	0.902	0.932
Avg. plant productivity	0.912	0.938
Top wage	0.726	0.794
Bottom wage	0.862	0.765
Ret to skill (employees)	0.247	1.114
Ret to skill (employers)	0.966	0.977

* Relative to undistorted benchmark calibrated to the U.S.

4. Counterfactual experiments

In this section we conduct two sets of counterfactual experiments. First, we introduce two different policies in our undistorted benchmark: the size-dependent tax calibrated from the micro-evidence on payroll taxes in Mexico, and a perfectly-enforced proportional tax of the same level as the average tax obtained under size-dependent enforcement. Next, we consider two counterfactual policy scenarios in our distorted benchmark: complete elimination of the payroll tax, and perfect enforcement of the average effective tax rate under size-dependent enforcement.

We find that the level of the tax is the main driver of reallocation and output losses, while the nature of enforcement determines the effects on the returns to skill. We show that the output losses generated by the size-dependent tax are the result of talent mismatch between managers and wage workers, and the reallocation of the best wage workers into self-employment, whereas the effect of managerial reallocation is null.

4.1. The effects of size-dependent enforcement in the U.S. benchmark

We introduce a tax on labor with size-dependent enforcement like the one observed in Mexico into our undistorted benchmark economy calibrated to the U.S. In our distorted benchmark for Mexico the average plant (size 5.65) pays an effective payroll tax of 30%, and we calibrate the enforcement parameter, κ, so that the average establishment in the U.S. (size 17.72) pays the same effective tax rate as the average establishment in Mexico. This way the size gradient of the enforcement policy exhibits comparable variation across the team size distribution in both countries. Results from this exercise are reported in column (1) of Table 6. Then, in column (2), we report the effects of a perfectly enforced proportional tax that is equal to the average tax under size-dependent enforcement.

Introducing a size-dependent tax into our U.S. benchmark economy reduces wage employment by 11 percent, while self-employment more than doubles—the share increases from 7.6 to 17.3 percent, almost as high as in Mexico. Because the wage workers moving into self-employment represent a small share of plants (as they were forming the largest teams in the economy), the share of employers decreases by only one percent. The average establishment size decreases 10 percent.

Table 7
Output loss decomposition from size-dependent enforcement.

	(1) Reallocation of managers into self-employment	(2) Talent mismatch	(3) Reallocation of wage workers into self-employment
Contribution to output loss	-0.01%	50.78%	49.23%

Aggregate output—which we measure using production in teams—decreases 10 percent. Distortions in this economy reduce aggregate output because they reallocate the marginal worker in each occupation, which in turn changes the sorting of the infra-marginal workers. As a result, employers coordinate less talented employees and run smaller teams relative to the equilibrium without distortions. We can decompose the effects of size-dependent policies into losses from the reallocation of workers into self-employment, plus losses from talent mismatch. First, note that to measure aggregate output, we can either add up the contribution of each wage worker to production (using the left tail of the skill distribution) or add up the output of each team (using the right tail of the skill distribution):

$$Y = \int_{L}^{z_1} G\left(m(i)\right)\left[n\left(i\right)\right]^{\alpha} \frac{f\left(i\right)}{n(i)} di = \int_{z_2}^{H} G\left(i\right)\left[n\left(p\left(i\right)\right)\right]^{\alpha} f\left(i\right) di, \tag{11}$$

where $m(i)$ is the equilibrium assignment function from equation (6) that matches wage workers of quality i to managers of quality $m(i)$, and $p(i)$ is the equilibrium assignment function that matches managers of quality i to wage workers of quality $p(i)$—the inverse of $m(i)$. We measure aggregate output using production in teams only—self-employment is the outside option and does not feature any production decision.

Then, we combine the two alternative measures of aggregate output in equation (11) to isolate each margin of misallocation—the reallocation of managers into self-employment, the reallocation of wage workers into self-employment, and the talent mismatch:

$$\begin{aligned} Y^d - Y^u = & \left[-\int_{z_2^u}^{z_2^d} G\left(i\right)\left[n\left(p^u\left(i\right)\right)\right]^{\alpha} f\left(i\right) di + \int_{L}^{p^u\left(z_2^d\right)} G\left(m^d\left(i\right)\right)\left[n\left(i\right)\right]^{\alpha-1} f\left(i\right) di \right] \\ & \left[-\int_{z_2^d}^{m^u\left(z_1^d\right)} G\left(i\right)\left[n\left(p^u\left(i\right)\right)\right]^{\alpha} f\left(i\right) di + \int_{p^u\left(z_2^d\right)}^{p^d\left(m^u\left(z_1^d\right)\right)} G\left(m^d\left(i\right)\right)\left[n\left(i\right)\right]^{\alpha-1} f\left(i\right) di \right] \\ & \left[-\int_{z_1^d}^{z_1^u} G\left(m^u\left(i\right)\right)\left[n\left(i\right)\right]^{\alpha-1} f\left(i\right) di + \int_{m^u\left(z_1^d\right)}^{H} G\left(i\right)\left[n\left(p^d\left(i\right)\right)\right]^{\alpha} f\left(i\right) di \right], \end{aligned} \tag{12}$$

where the superscripts u and d denote the undistorted and distorted equilibria. We provide the details of this decomposition in Appendix A.

Managers in $\left[z_2^u, z_2^d\right]$ become self-employed with the policy, but their former wage workers in $\left[L, p^u\left(z_2^d\right)\right]$ are matched with better managers relative to the undistorted equilibrium. The first term captures the net loss from this reallocation. Similarly, wage workers in $\left[z_1^d, z_1^u\right]$ become self-employed with the policy, which destroys their teams, but their former managers $\left[m^u\left(z_1^d\right), H\right]$ are assigned into smaller teams. The third term captures the net cost of this re-assignment. The second term is a pure talent mismatch: wage workers and managers in these segments of the skill distribution do not switch occupations, but managers are matched with employees of lower skill in smaller teams.

Table 7 separates this output loss into the cost from talent mismatch and the net cost from the reallocation of workers across occupations following equation (12). The reallocation of the worst managers into self-employment does not affect aggregate output. Matching managers with less talented employees into smaller teams explains 51% of the loss in output, while the reallocation of the best wage workers into self-employment explains the remaining 49%.

To provide additional context for the magnitude of the effects of size-dependent policies on output, we introduce the same distortion of column (1) in Table 6 into a standard span-of-control model in the spirit of Lucas (1978), calibrated to match the same moments of the distribution of establishment sizes in the U.S. as in our benchmark undistorted calibration. We detail this exercise in Appendix A. Table 8 summarizes our results. The average establishment size in that economy is cut nearly in half with the size-dependent statutory tax of 44%, but output losses are only three percent. When we introduce

Table 8
Losses from size-dependent enforcement in a Lucas (1978) economy.*

	Tax policy from Mexico ($\tau_s = 0.44$, $\kappa = 0.0658$)	Lower tax rate ($\tau_s = 0.05$, $\kappa = 0.0658$)
Avg. est. size	0.517	0.904
Output	0.973	0.999

* Relative to undistorted scenario calibrated to the U.S.

Table 9
The gains from removing size-dependent labor regulations.*

	(1) No payroll tax	(2) 34.5% flat tax (perfectly enforced)
Share of wage workers	1.126	1.026
Share of non-employers	0.569	0.909
Share of employers	1.007	1.002
Avg. est. size	1.118	1.024
Output	1.091	1.020
Avg. plant productivity	1.084	1.018
Top wage	1.272	1.051
Bottom wage	1.152	0.858
Ret to skill (employees)	1.14	1.367
Ret to skill (employers)	1.019	1.005

* Relative to distorted benchmark under size-dependent tax with $\tau_s = 0.44$, and $\kappa = 0.2145$ calibrated to Mexico.

instead a tax rate low enough generate the same reduction in average establishment size as in our benchmark economy (10 percent), output losses are nearly zero. Thus, in our economy with positive sorting, moderate reductions in average plant size can generate large output losses. The amplification mechanism is the talent mismatch absent in the standard model with homogeneous wage workers.

Our model does not feature a measure of total factor productivity (we cannot separate size from the productivity of both managers and wage workers), but in Table 6 we also report changes in average plant productivity (output per plant)—which captures the re-assignment of managers to worse wage workers in response to the tax. Average plant productivity drops by nearly nine percent as a result of the size-dependent tax.

The main losers from the size-dependent tax are wage workers. Top wages—wages for the most skilled wage workers who are indifferent between self-employment and wage working—decrease by as much as 27 percent, whereas bottom wages—wages for wage workers of skill level L—drop 14 percent. The resulting wage schedule is flatter, and thus the returns to skill for wage workers, which we compute regressing log wages on log skill, decrease by 75 percent. In contrast, returns to skill for managers decrease by less than four percent.[15]

Finally, in column (2) of Table 6 we introduce a perfectly-enforced proportional labor tax of the same level as the average effective tax under size-dependent enforcement (31%). It turns out the flat tax has relatively large effects: it accounts for more than half of the reductions in average establishment size and output generated by the size-dependent tax. While this result may seem surprising, recall that in our model equilibrium forces act from top to bottom, so any policy that modifies the production decisions of the largest teams—including a proportional payroll tax—will create a trickle-down effect on the allocation of workers across and within occupations, which is ultimately reflected on output. However, the effects of the flat tax on the returns to skill are very different compared to those of the size-dependent tax: a flat tax means a higher cost of labor for all teams in the economy, and while the managers at the bottom have self-employment as an outside option, that is not the case for low-skill workers, who see their earnings reduced by a larger proportion than high-skill workers. Therefore, in this exercise, the wage profile is steeper relative to the undistorted benchmark, and the returns to skill for wage workers actually increase by 11 percent as a result of the flat tax.

4.2. Introducing perfect enforcement in Mexico

Table 9 displays the results from changing the enforcement intensity, as well as the level of the payroll tax in our distorted benchmark economy calibrated to Mexico. In column (1) we show the effects of completely eliminating the tax. Wage employment increases by nearly 13 percent, while self-employment decreases by more than 40 percent. Because all the additional wage workers require only a small fraction of managers (since they form the largest teams), there is a rather small increase in the share of employers—less than one percent.

[15] To compute the return to skill, we obtain 300 draws from the skill distribution in each occupation, obtain the distribution of wages and the distribution of managerial rents, and then regress log earnings on log skill in each occupation.

Table 10
Output gain decomposition from removing size-dependent policies.

	(1) Reallocation of self-employed into managerial activities	(2) Talent mismatch	(3) Reallocation of self-employed into wage working
Contribution to output gain	0.00%	55.67%	44.33%

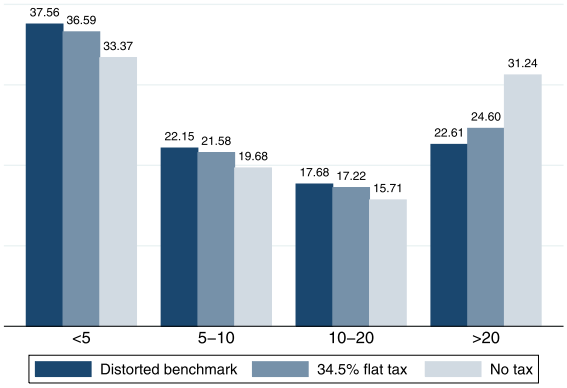

Fig. 12. Effects of changes in the tax policy on employment shares across size categories.

The average establishment size increases by 12 percent when the payroll tax is removed (from 5.65 to 6.33 employees), whereas output and average productivity increase by nine and eight percent, respectively. That is, in economies with positive sorting, small increments in the average size (less than one employee) may produce large output and productivity gains. Moreover, policy alone does not seem to explain the size distribution of plants in Mexico—removing distortions does not result in significantly larger plants. Technology—communication costs and the degree of decreasing returns—and the skill distribution are decisive factors as well.

Table 10 separates the change in aggregate output into the effect of talent mismatch and the effect from the reallocation of talent away from self-employment. Matching managers with more talented employees in larger teams explains 56% of the output gain, while the reallocation of workers away from self-employment and into wage working—adding some new very large teams—generates the remaining 44%. The reallocation of some self-employed into managerial activities does not affect aggregate output.

The main winners from the elimination of the tax are wage workers, especially those of higher skill. Whereas wages increase for all production workers, top wage workers (who are indifferent between self-employment and working for a wage) benefit more than those at the bottom (27% vs. 15%). The larger wage increase for high-skill employees compared to their low-skill counterparts is reflected in a 14 percent increase in the returns to skill for wage workers.

Last, in column (2) we examine the effect of a perfectly-enforced payroll tax of the same level as the average effective tax rate under size-dependent enforcement (34.5 percent). Under this policy, wage employment increases by 2.6 percent, while self-employment decreases by nine percent. Output and average plant productivity increase by two percent. Returns to skill for wage workers increase by nearly 37 percent, but the winners from this policy change are high-skill wage workers, whose earnings increase by as much as five percent. Low skill workers experience a decrease in earnings as large as 14 percent, and consequently are worse off when the policy changes from a size-dependent tax to a flat tax.

In Fig. 12 we show how the distribution of employment across plant sizes changes under our two counterfactuals. When we reduce the level of distortions, some of the formerly self-employed become employees in the largest teams. Because this optimal reassignment occurs from top to bottom, the employment shares of the largest establishments increases, while those in the smaller establishments decrease. The share of employment in the largest establishment increases by nearly 40 percent when we remove the payroll tax, but this increment is considerably more modest when we move to a flat tax equal to the average effective tax under size-dependent enforcement.

4.3. Talent misallocation in Mexico

Our results shed light on some of the driving forces behind the observed allocation of talent and the returns to skill in Mexico. First, size-dependent policies distort the allocation of the marginal individual in each occupation, and in particular, induce the best wage workers to enter into self-employment. Recently, Torres (2018) finds evidence of a misallocation of the most skilled workers into micro-business ownership. In particular, she finds that medium and high skill self-employed individuals would earn significantly more if they instead worked for a wage.

Second, size-dependent policies also re-sort the infra-marginal workers and attenuate the strength of the positive sorting throughout the entire economy. Fig. 13 shows the distribution of workers in the U.S. and Mexico across employer sizes,

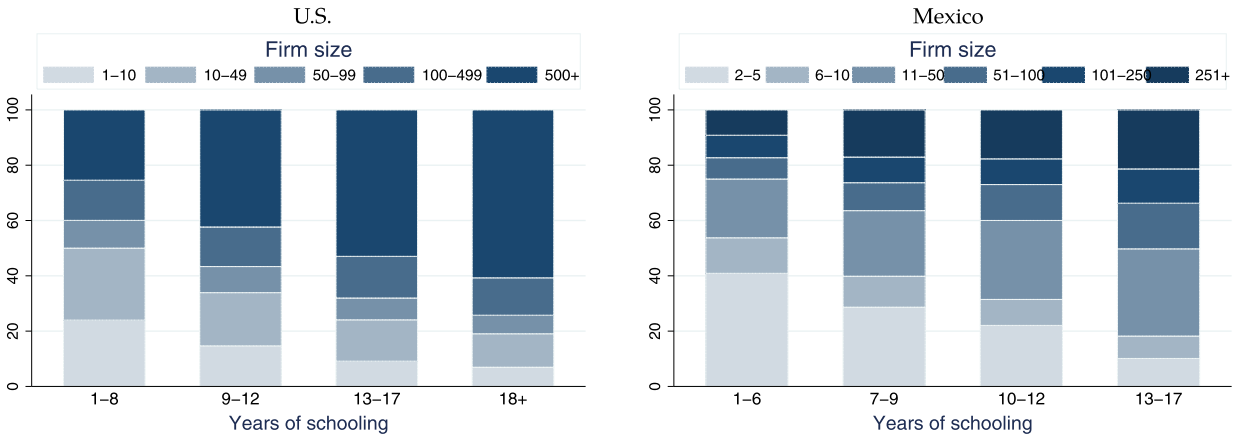

Fig. 13. Sorting of wage workers across years of schooling and firm size in the U.S. and Mexico.

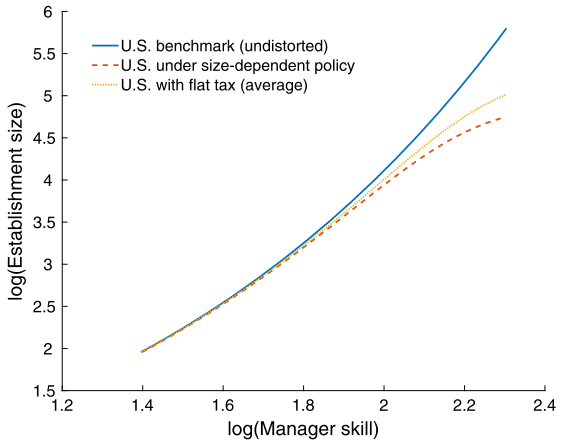

Fig. 14. Relationship between plant productivity and size with and without distortions.

conditional on years of schooling. In both countries, workers with more years of schooling are more likely to work in larger employers, but in Mexico the relationship is not as strong as in the U.S. for those at the top of the skill distribution.

5. Taking stock: resource misallocation in economies with positive sorting

Our analysis provides new insights into the role of plant-level distortions in an environment where heterogeneity matters for both wage workers and managers. First, talent misallocation amplifies the losses from size-dependent policies. Size-dependent policies in our model generate a reduction in average plant size that is only a fifth of that obtained using a standard span-of-control model, yet the implied output losses are three times as large. Matching managers with wage workers of lower skill accounts for over half of the losses from the size-dependent tax in our framework. The rest is accounted for by the reallocation of the best wage workers into self-employment—which is also absent in the standard span of control model because all wage workers are homogeneous—whereas the effect of the worst managers reallocating into self-employment is null.

Second, common wedges in an environment with positive sorting in quantity and quality cause larger distortionary effects than in the traditional span-of-control model. In the traditional framework, fixed amounts of homogeneous resources are allocated across heterogeneous productive units. Maximizing aggregate output requires allocating more labor to the more productive plants—and away from less productive units—to the point where the marginal product of labor is the same across all producers. In equilibrium, more productive plants are allocated proportionately more labor. In our framework, resources are heterogeneous and optimality requires positive sorting: the best, most productive wage workers match with the best managers to form the largest and most productive teams, and in equilibrium, more productive plants are allocated disproportionately more labor. In other words, optimality in our framework requires not only the optimal amount of resources flowing to every productive unit, but also allocating the optimal match. Fig. 14 shows the relationship in logs between managerial skill (the typical stand-in for team-level productivity) and establishment size in our U.S. benchmark, without and with distortions. The relationship between the log of the manager's productivity and the log of the size of

his team in our framework is not only positive but convex, whereas this relationship is only linear in the standard span of control model.

In the traditional framework, a flat, proportional payroll tax distorts occupational choices, but its effect on the allocation of workers into teams is subdued because, in equilibrium, the marginal product of labor remains constant across producers. In our framework there is an optimal assignment of managers and wage workers. A proportional tax distorts the production choices of the largest establishments, which reallocates the best wage workers into self employment, which in turn misallocates the assignment of managers into teams. As a result, higher quality managers produce with lower quality wage workers in smaller teams, and the most talented managers in the most productive plants are not able to leverage their skills with better workers. This misallocation of talent is potentially costly: in our undistorted benchmark, the proportional payroll tax of 31% results in output losses of 7%.

In the end, in terms of output and size effects, the level of the tax in our framework is of first-order importance. However, the degree of enforcement has important implications for the returns to skill, especially for wage workers: when enforcement increases with plant size, employees are not able to reap the benefits of higher skill levels, and their wage profile turns flatter relative to an undistorted equilibrium. On the contrary, when taxes are perfectly enforced, wages at the bottom take the largest hit, which creates a steeper wage profile, and thus can even increase the average returns to skill for wage workers.

Finally, the environment with knowledge hierarchies provides new insights on measures of misallocation in the data. In our framework, plants of identical sizes—and therefore identical managerial and wage working skill—do not exhibit dispersion in marginal products. Some of the dispersion in marginal products in the data may thus simply indicate differences in the skill composition of establishments that have not been controlled for. However, distortions in our environment attenuate the strength of the positive sorting and dampen the relationship between managerial skill and plant size, which may result in less (unconditional) heterogeneity in marginal products. That is, the undistorted counterfactual may actually exhibit more—not less—unconditional dispersion in marginal products.

Ultimately, our understanding and measurement of marginal products comes from models designed to understand how production is organized. The model with positive sorting in quality and quantity generates a marginal product of labor that is directly tied to different levels of skill for both wage workers and managers. Thus, to try to measure this version of marginal product in the data, we would ideally need information on skill proxies for workers and managers in each productive unit. The relatively recent emergence of high-quality datasets that match employees and employers offers a promising source of such valuable information.

We believe that the way we think about the relationship between plant size, marginal products, and misallocation should be refined to take into account the skill composition of establishments, as well as the evidence on positive sorting. Moreover, the models used to address the aggregate effects of idiosyncratic distortions among productive units should acknowledge these forces. We think our paper makes an important contribution towards this goal.

6. Concluding remarks

We show that size-dependent regulations disproportionately lower the returns to skill for the most able workers and managers. Both the allocation of talent and the returns to skill are prominent concepts in the fields of growth and development—some might say, in the entire discipline of economics—yet remain largely unexplored in the vast literature on firm-specific distortions and misallocation. Our framework allows us to bridge this gap, and its analytical convenience opens many exciting lines of inquiry. Some of these include studying the effects of size-dependent policies on the acquisition of skills, as well as the endogenous evolution of occupational choices and the skill distribution throughout the process of development.

A drawback of the knowledge hierarchy model with two layers lies in its inability to reproduce employment shares of the largest establishments, especially in economies such as the U.S., where the largest productive units account for a disproportionately large share of workers. Future work on misallocation using the positive sorting framework should address this issue. The addition of more layers, as in the recent work by Tamkoç (2019), is the most straightforward way of generating large employment shares at the top, but a thorough quantitative exploration of that environment requires rich datasets on the organization of production at the establishment level.

Appendix A

A.1. Knowledge hierarchies and the "bunching" of teams

We describe the qualitative effects of a policy that can potentially break down positive sorting, and generates bunching in the distribution of sizes. Specifically, we consider the effects of the following Sigmoid tax policy:

$$\tau(n) = \frac{\tau_s}{1 + \exp[-\kappa(n - N)]},$$

where τ_s is the general, statutory tax rate, N is the Sigmoid's midpoint, which represents the threshold around which marginal taxes significantly increases with size. The parameter κ reflects the steepness of the Sigmoid function around this threshold. Fig. 15 shows an example with $\tau = 0.3$, and $N = 50$, for three different levels of κ.

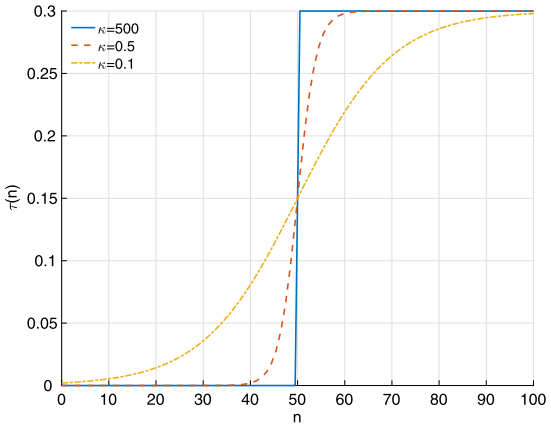

Fig. 15. Sigmoid tax function ($\tau = 0.3$, $N = 50$).

When κ becomes arbitrarily large, this continuous tax approaches a step tax policy of the following form: managers who hire more than N workers pay a tax $\tau \in (0, 1)$ on their wage bill, and managers who employ fewer than N workers pay no tax. Garicano et al. (2016) estimate the effects of this policy in a standard Lucas (1978) setup, in which wage workers are homogeneous in their abilities. A step-tax policy generates a mass point at the enforcement size threshold N. Because in their data on French firms they observe bunching across a range of sizes, and not a mass point at a single size value, Garicano et al. (2016) work around this complication by assuming measurement error in their data. The Sigmoid tax function we consider is the theoretical counterpart to their empirical methodology to fit the observed distribution of sizes under size-dependent policies. Moreover, the Sigmoid tax policy is continuous, and we can then easily obtain the first order conditions in the manager's problem to characterize the solution.

Consider the FOC of the entrepreneur's optimization problem:

$$\begin{aligned} G(z_m)\,\alpha n^{\alpha-1} - w(z_p)\left[1+\tau(n)+\tau'(n)\,n\right] - \mu &\geq 0, \\ -w'(z_p)\left[1+\tau(n)\right]n + \mu n'(z_p) &\geq 0, \\ \mu\left[n(z_p) - n\right] &= 0, \\ \mu &\geq 0. \end{aligned}$$

Where μ is the Lagrange multiplier associated with the optimal size constraint. Note that if the size constraint is binding, prices and assignments in equilibrium are solved for as in the undistorted case. In this case, managers constrain the size of their team by matching with workers of lower skill, but still according to the communication technology $n(z_p)$. On the other hand, if the size constraint is not binding, then $\mu = 0$ and the optimal size solves:

$$\frac{G(z_m)\,\alpha}{w(z_p)} = n^{1-\alpha}\left[1+\tau(n)+\tau'(n)\,n\right], \tag{13}$$

and from the second FOC it must be true that:

$$\begin{aligned} -w'(z_p)\left[1+\tau(n)\right]n &= 0, \\ \iff w'(z_p) &= 0. \end{aligned}$$

That is, the solution requires wages to be independent of skill if the size constraint is not binding: if managers constrain the size of their team to avoid the tax, then the wages paid to their workers must be independent of their skill. Notice, however, that in this case sizes will depend on managerial skill, as dictated by equation (13)—just as in a Lucas (1978) span-of-control model.

Fig. 16 shows the assignment function for our undistorted benchmark economy, as well as for an economy subject to a Sigmoid tax policy. The Sigmoid parameters for this example are $\tau_s = 0.3$, $\kappa = 0.25$, and $N = 20$. The equilibrium assignment under this size-dependent tax can be characterized by six thresholds, $\{z_i\}_{i=1}^{6}$, such that: (i) $[L, z_1]$ is the set of low-skilled workers in small teams who pay low taxes and setup their size according to the communication technology $n(\cdot) \equiv 1/h(\cdot)$; (ii) $[z_1, z_2]$ is the set of medium-skilled workers in constrained teams that choose their size based on the skill of the manager, according to equation (13); (iii) $[z_2, z_3]$ is the set of high-skilled workers in constrained teams who choose their size according to the communication technology $n(\cdot) \equiv 1/h(\cdot)$; (iv) $[z_3, z_4]$ is the set of self-employed (without employees); (v) $[z_4, z_5]$ is the set of managers of small teams matched with workers in $[L, z_1]$; (vi) $[z_5, z_6]$ is the set of managers of medium-sized teams who set their size according to equation (13) and hire workers in $[z_1, z_2]$; (vii) $[z_6, H]$ is the set of managers of large teams matched with workers in $[z_2, z_3]$.

Fig. 16. Equilibrium assignment before and after the tax.

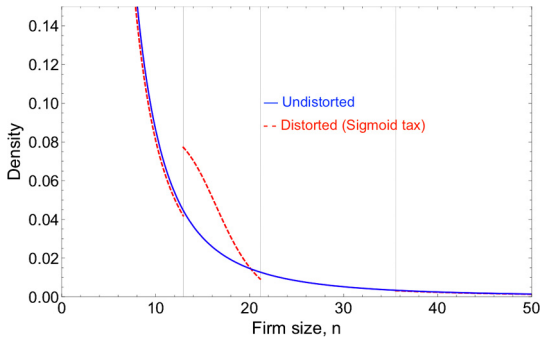

Fig. 17. Equilibrium distribution of sizes before and after the Sigmoid tax.

The tax then encourages employers within a range of the size threshold N to constrain the size of their team to avoid the tax, and discourages their wage workers from fully exploiting their talent. Positive sorting in the constrained teams who choose their size according to equation (13) is not guaranteed. More specifically, let $l^*(z_m)$ denote the size that solves the FOC in equation (13). Then, the minimum worker skill required to support l^* is given by $z_p^* = n^{-1}(l^*)$, where $n(\cdot) \equiv 1/h(\cdot)$. Notice, however, that in this segment of the skill distribution, a manager can setup a team of size $l^*(z_m)$ using any worker with ability $z \in [z_p^*, z_2]$, because they all earn the same wage. Therefore, positive sorting is one of infinitely many possible outcomes among constrained teams and workers. This segment of the assignment function is therefore indeterminate. Managers could potentially mix employees from different types in a single team, all of them earning the same wage.

Fig. 17 displays the equilibrium size distribution before and after the tax. In equilibrium, the size constraint for managers at the bottom and at the top of the managerial talent distribution always binds, and they maximize at an interior solution. To the contrary, managers in a segment around N constrain the size of their team to sizes $[l^*(z_4), l^*(z_5)]$ following the FOC in equation (13), and no manager selects sizes $(l^*(z_5), n(z_2))$. In other words, the size distribution exhibits a bunching of teams in $[l^*(z_4), l^*(z_5)]$, with no mass in the segment $(l^*(z_5), n(z_2))$. When κ becomes arbitrarily large, this bunching of teams occurs at exactly N, as in the model of Garicano et al. (2016).

A.2. Calibration strategy: closed form solutions for equilibrium objects

Differential equations for assignments and wages. The assignment function in equilibrium follows the differential equation

$$\begin{aligned}\frac{f(x)}{n(x)} &= f(m(x))\, m'(x)\,, \\ &= \frac{d}{dx} F(m(x))\,.\end{aligned}$$

Integrating both sides assuming a double truncated exponential distribution of skill yields

$$\int \frac{f(x)}{n(x)} dx = F(m(x))\,,$$

$$\frac{-a\lambda\left[H-L\right]}{\left(\exp\left[-\lambda L\right]-\exp\left[-\lambda H\right]\right)\left(\lambda\left(H-L\right)+b\right)}\left(\frac{\exp\left[-\lambda x\right]}{an\left(x\right)}\right)+c=\frac{\exp\left[-\lambda L\right]-\exp\left[-\lambda m\left(x\right)\right]}{\exp\left[-\lambda L\right]-\exp\left[-\lambda H\right]}.$$

Where c denotes the constant of integration. Using the boundary condition $m\left(L\right)=z_2$ to solve for c yields

$$c=\frac{\exp\left[-\lambda L\right]-\exp\left[-\lambda z_2\right]}{\exp\left[-\lambda L\right]-\exp\left[-\lambda H\right]}-\frac{-a\lambda\left[H-L\right]}{\left(\exp\left[-\lambda L\right]-\exp\left[-\lambda H\right]\right)\left(\lambda\left(H-L\right)+b\right)}\left(\frac{\exp\left[-\lambda L\right]}{an\left(L\right)}\right),$$

and therefore

$$m\left(x\right)=-\frac{1}{\lambda}\ln\left[1+\frac{a\lambda\left[H-L\right]}{\lambda\left(H-L\right)+b}\left(\frac{\exp\left[\lambda\left(z_2-x\right)\right]}{an\left(x\right)}\right)-\frac{a\lambda\left[H-L\right]}{\lambda\left(H-L\right)+b}\left(\frac{\exp\left[\lambda\left(z_2-L\right)\right]}{an\left(L\right)}\right)\right]+z_2.$$

Now define

$$\frac{a\lambda\left[H-L\right]}{\lambda\left(H-L\right)+b}\equiv\gamma,$$

and rewrite the assignment function using this new constant:

$$m\left(x\right)=\left(-\frac{1}{\lambda}\right)\ln\left[1+\gamma\left(\frac{\exp\left[\lambda\left(z_2-x\right)\right]}{an\left(x\right)}\right)-\gamma\left(\frac{\exp\left[\lambda\left(z_2-L\right)\right]}{an\left(L\right)}\right)\right]+z_2.$$

Assume

$$\gamma\exp\left[\lambda\left(z_2-L\right)+b\right]=1.$$

If this assumption holds (Assumption 2), then we can simplify the assignment function even further to obtain

$$m\left(x\right)=x\left[1+\left(\frac{1}{\lambda}\right)b\left(\frac{1}{H-L}\right)\right]+\left(-\frac{1}{\lambda}\right)\ln\gamma+\left(-\frac{1}{\lambda}\right)b\left(\frac{H}{H-L}\right),$$

which is a linear function of skill. To obtain the wage function we must solve the following differential equation (see FOC in manager's problem):

$$G\left[m\left(x\right)\right]\left[\frac{d}{dx}\left[n\left(x\right)^{\alpha}\right]\right]=\frac{d}{dx}\left[w\left(x\right)n\left(x\right)\right].$$

The left-hand side can be rewritten as

$$G\left(m\left(x\right)\right)\left[\frac{d}{dx}\left[n\left(x\right)^{\alpha}\right]\right]=\left[k_1x+k_2\right]\exp\left[k_3+k_4x\right],$$

where

$$k_1=\left[1+\left(\frac{1}{\lambda}\right)\left(\frac{b}{H-L}\right)\right]\alpha\left[\frac{b}{\left[H-L\right]^2}\right]\left(\frac{1}{a^{\alpha}}\right),$$

$$k_2=\left[\left(-\frac{1}{\lambda}\right)\ln\gamma+\left(-\frac{1}{\lambda}\right)\left(\frac{b}{H-L}\right)H-L\right]\alpha\left[\frac{b}{\left[H-L\right]^2}\right]\left(\frac{1}{a^{\alpha}}\right),$$

$$k_3=-b\alpha\left(\frac{H}{H-L}\right),$$

$$k_4=b\alpha\left(\frac{1}{H-L}\right).$$

Then,

$$\int G\left[m\left(x\right)\right]\left[\frac{d}{dx}\left[n\left(x\right)^{\alpha}\right]\right]dx=\int\left[k_1x+k_2\right]\exp\left[k_3+k_4x\right]dx$$

$$=\exp\left[k_3+k_4x\right]\left[\frac{k_2}{k_4}-\frac{k_1}{k_4^2}+\frac{k_1x}{k_4}\right],$$

and therefore (integrating also the right-hand side)

$$\exp\left[k_3+k_4x\right]\left[\frac{k_2}{k_4}-\frac{k_1}{k_4^2}+\frac{k_1}{k_4}x\right]+c=w\left(x\right)n\left(x\right).$$

Where c denotes the constant of integration. Using the boundary condition $w(z_1) = G(z_1)$ to solve for c yields

$$c = G(z_1)\, n(z_1) - \exp[k_3 + k_4 z_1]\left[\frac{k_2}{k_4} - \frac{k_1}{k_4^2} + \frac{k_1}{k_4} z_1\right].$$

To solve for the equilibrium we have two final conditions:

$$G(z_2)\, n(L)^{\alpha} - w(L)\, n(L) = G(z_2),$$
$$m(z_1) = H.$$

Take the second equation to solve for z_1:

$$\begin{aligned} m(z_1) = H &= a^{\alpha}[H - L]\left[\frac{z_1 k_1 + k_2}{k_4}\right] + L, \\ H - L &= a^{\alpha}[H - L]\left[\frac{z_1 k_1 + k_2}{k_4}\right], \\ z_1 &= \left[\frac{\left(\frac{1}{a^{\alpha}}\right) k_4 - k_2}{k_1}\right], \\ &= H + \frac{(H - L)\ln\gamma}{b + (H - L)\lambda}. \end{aligned}$$

Use the first equation to solve for z_2

$$\begin{aligned} G(z_2)\, n(L)^{\alpha} - w(L)\, n(L) &= G(z_2), \\ G(z_2)\left[n(L)^{\alpha} - 1\right] &= w(L)\, n(L), \\ G(z_2) = \frac{z_2 - L}{H - L} &= \frac{w(L)\, n(L)}{n(L)^{\alpha} - 1}, \\ z_2 - L &= \frac{w(L)\, n(L)}{n(L)^{\alpha} - 1}[H - L], \\ z_2 &= L + \frac{w(L)\, n(L)}{n(L)^{\alpha} - 1}[H - L]. \end{aligned}$$

Now recall that from Assumption 2 we need:

$$z_2 = L - \left(\frac{1}{\lambda}\right) b - \left(\frac{1}{\lambda}\right)\ln\gamma.$$

Then, the parametric restriction becomes:

$$\begin{aligned} L - \left(\frac{1}{\lambda}\right) b - \left(\frac{1}{\lambda}\right)\ln\gamma &= L + \frac{w(L)\, n(L)}{n(L)^{\alpha} - 1}[H - L], \\ -\left(\frac{1}{\lambda}\right) b - \left(\frac{1}{\lambda}\right)\ln\gamma &= \frac{w(L)\, n(L)}{n(L)^{\alpha} - 1}[H - L], \\ \left(-\frac{1}{\lambda}\right)[b + \ln\gamma] &= \frac{w(L)\, n(L)}{n(L)^{\alpha} - 1}[H - L]. \end{aligned}$$

Establishment size distribution. To obtain the size distribution of teams, first notice that using the communication technology, we can write the skill of worker z_p as a function of his size,

$$z_p(n) = \frac{H - L}{b}\ln(n) + \frac{H - L}{b}\ln a + H.$$

Then, using the equilibrium assignment function, we can write $m(n) = m(z_p(n))$, which represents the manager that corresponds to size n.

$$m(n) = \left(\frac{H - L}{b}\ln(n) + \frac{H - L}{b}\ln a + H\right)\left[1 + \left(\frac{1}{\lambda}\right) b\left(\frac{1}{H - L}\right)\right] - \frac{1}{\lambda}\ln\gamma - \frac{b}{\lambda}\left(\frac{H}{H - L}\right).$$

Then

$$\frac{dm}{dn} = \left(\frac{H-L}{b} + \frac{1}{\lambda}\right)\left[\frac{1}{n}\right],$$
$$= \lambda^{-1} p n^{-1}.$$

Where $p \equiv \frac{\lambda(H-L)}{b} + 1$. The managerial skill distribution follows $\frac{F(m)-F(z_2)}{1-F(z_2)}$. Using the change of variable technique yields

$$\Pr\left(n \leq \overline{n}\right) = \Pr\left(n\left(m\right) \leq \overline{n}\right),$$
$$= \Pr\left(m \leq n^{-1}\left(\overline{n}\right)\right),$$
$$= \frac{F\left(n^{-1}\left(\overline{n}\right)\right) - F\left(z_2\right)}{1 - F\left(z_2\right)}.$$

Then

$$\frac{f\left(m\left(n\right)\right)}{1-F\left(z_2\right)}\left(\frac{dm}{dn}\right) = \frac{\frac{\lambda \exp[-\lambda m(n)]}{\exp[-\lambda L]-\exp[-\lambda H]}}{\frac{\exp[-\lambda z_2]-\exp[-\lambda H]}{\exp[-\lambda L]-\exp[-\lambda H]}}\left(\frac{dm}{dn}\right)$$
$$= \frac{\exp\left[-\lambda m\left(n\right)\right] p n^{-1}}{\exp\left[-\lambda z_2\right] - \exp\left[-\lambda H\right]}.$$

We can then write the size density as

$$\tilde{f}\left(n\right) = \frac{B p n^{-p-1}}{\exp\left[-\lambda z_2\right] - \exp\left[-\lambda H\right]}.$$

We want to show that

$$C \equiv \frac{B}{\exp\left[-\lambda z_2\right] - \exp\left[-\lambda H\right]} = \frac{[n(L)]^p}{1 - \left(\frac{n(L)}{n(z_1)}\right)^p}$$

Notice that

$$f\left(m(n)\right) = \exp\left[-\lambda\left(\frac{H-L}{b}\ln(n) + \frac{H-L}{b}\ln a + H\right)\left[1 + \left(\frac{1}{\lambda}\right) b\left(\frac{1}{H-L}\right)\right] + \ln\gamma + \frac{bH}{H-L}\right].$$

Then we can write the constant B as

$$B = \exp\left[-\lambda\left(\frac{H-L}{b}\ln a + H\right)\left[1 + \left(\frac{1}{\lambda}\right) b\left(\frac{1}{H-L}\right)\right] + \ln\gamma + \frac{bH}{H-L}\right].$$

Thus,

$$C = \frac{\exp\left[-\lambda\left(\frac{H-L}{b}\ln a + H\right)\left[1 + \left(\frac{1}{\lambda}\right) b\left(\frac{1}{H-L}\right)\right] + \ln\gamma + \frac{bH}{H-L}\right]}{\exp\left[-\lambda z_2\right] - \exp\left[-\lambda H\right]},$$
$$= \frac{\exp\left[-p\ln a + (-p+1)\frac{bH}{H-L} + \ln\gamma\right]}{\gamma\exp\left[-\lambda L + b\right] - \exp\left[-\lambda H\right]},$$
$$= \frac{a^{-p}\exp\left[(-p+1)\frac{bH}{H-L} + \lambda L - b\right]}{1 - \frac{1}{\gamma}\exp\left[-\lambda(H-L) - b\right]},$$
$$= \frac{a^{-p}\exp(-pb)}{1 - \frac{exp(-pb)}{\gamma}},$$
$$= \frac{[n(L)]^p}{1 - \left(\frac{n(L)}{n(z_1)}\right)^p}.$$

In the derivation above we have used Assumption 2 and the fact that $[n(L)]^p = a^{-p}\exp(-pb)$ and $[n(z_1)]^p = a^{-p}\gamma$. Then, the size density writes

$$\tilde{f}(n) = Cpn^{-p-1},$$
$$= \frac{[n(L)]^p pn^{-p-1}}{1-\left(\frac{n(L)}{n(z_1)}\right)^p},$$

which is a double truncated Pareto density.

Equilibrium assignment and wages under size-dependent policies. To obtain closed form solutions for equilibrium objects in the environment with distortions, first notice that if the manager's problem has an internal solution (which is the case with the negative exponential tax function considered), then managers use the communication technology to setup the size of their team. That implies that the market-clearing condition used to derive the assignment function remains the same under the size-dependent policy.

To obtain the wage function we must solve the differential equation from the FOC in the manager's problem:

$$G[m(x)]\left[\frac{d}{dx}\left[n(x)^\alpha\right]\right] = \frac{d}{dx}\left[(1+\tau(n(x)))w(x)n(x)\right].$$

The left-hand side can be rewritten as before,

$$G(m(x))\left[\frac{d}{dx}\left[n(x)^\alpha\right]\right] = [k_1x+k_2]\exp[k_3+k_4x],$$

where the constants k_i, $i \in \{1,2,3,4\}$ are the same as before. Then

$$\int G[m(x)]\left[\frac{d}{dx}\left[n(x)^\alpha\right]\right]dx = \int [k_1x+k_2]\exp[k_3+k_4x]dx$$
$$= \exp[k_3+k_4x]\left[\frac{k_2}{k_4}-\frac{k_1}{k_4^2}+\frac{k_1x}{k_4}\right],$$

and therefore (integrating also the right-hand side)

$$\exp[k_3+k_4x]\left[\frac{k_2}{k_4}-\frac{k_1}{k_4^2}+\frac{k_1}{k_4}x\right]+c = (1+\tau(n(x)))w(x)n(x).$$

Where c denotes the constant of integration. Using the boundary condition $w(z_1) = G(z_1)$ to solve for c yields

$$c = (1+\tau(n(z_1)))G(z_1)n(z_1) - \exp[k_3+k_4z_1]\left[\frac{k_2}{k_4}-\frac{k_1}{k_4^2}+\frac{k_1}{k_4}z_1\right].$$

To solve for the equilibrium we have two final conditions. The first one is $m(z_1) = H$, which is exactly the same as in the undistorted case. The second one is given by

$$G(z_2)n(L)^\alpha - (1+\tau(n(L)))w(L)n(L) = G(z_2),$$
$$G(z_2)\left[n(L)^\alpha - 1\right] = (1+\tau(n(L)))w(L)n(L),$$
$$G(z_2) = \frac{z_2-L}{H-L} = \frac{(1+\tau(n(L)))w(L)n(L)}{n(L)^\alpha-1},$$
$$z_2 - L = \frac{(1+\tau(n(L)))w(L)n(L)}{n(L)^\alpha-1}[H-L],$$
$$z_2 = L + \frac{(1+\tau(n(L)))w(L)n(L)}{n(L)^\alpha-1}[H-L].$$

Then, the parametric restriction from Assumption 2 then becomes

$$L - \left(\frac{1}{\lambda}\right)b - \left(\frac{1}{\lambda}\right)\ln\gamma = L + \frac{(1+\tau(n(L)))w(L)n(L)}{n(L)^\alpha-1}[H-L],$$
$$-\left(\frac{1}{\lambda}\right)b - \left(\frac{1}{\lambda}\right)\ln\gamma = \frac{(1+\tau(n(L)))w(L)n(L)}{n(L)^\alpha-1}[H-L],$$
$$\left(-\frac{1}{\lambda}\right)[b+\ln\gamma] = \frac{(1+\tau(n(L)))w(L)n(L)}{n(L)^\alpha-1}[H-L].$$

Table 11
Moment elasticity with respect to calibrated parameters.

Moment	α	H	λ	b
Share of wage workers	1.122	-0.193	0.174	0.366
Share of non-employers	-12.99	2.691	-2.423	-14.8
Share of employers	0.0599	-0.719	0.648	15.95
Average size	1.062	0.526	-0.474	-15.59
Share of plants with $n \leq 10$	-0.0599	-0.296	0.266	39.66
Share of plants with $n \leq 20$	-0.0599	-0.159	0.143	6.405
Share of plants with $n \geq 50$	1.221	1.26	-1.134	-25.24
Share of plants with $n \geq 100$	4.129	1.956	-1.76	-26.53
Share of wage workers in plants with $n \leq 10$	-1.122	-0.816	0.735	45.98
Share of wage workers in plants with $n \leq 20$	-1.122	-0.634	0.571	10.91
Share of wage workers in plants with $n \geq 50$	3.279	1.261	-1.135	-11.24
Share of wage workers in plants with $n \geq 100$	7.476	2.032	-1.829	-13.31

A.3. Relationship between calibrated parameters and moments in the undistorted benchmark

The four parameters in our joint calibration are $\{H, \lambda, b, \alpha\}$. H and λ affect, respectively, the support and shape of the talent distribution; H in addition affects the difficulty of tasks, which in turn affects communication costs and potential team sizes; b determines both the level and the slope of the function governing communication costs; finally, α determines the returns to scale in the use of time. We examine the numerical relationship in our undistorted environment between these four parameters and some specific moments with the following sensitivity exercise.

We consider the undistorted benchmark economy (the calibration for the US) and separately perturb each individual parameter while maintaining L and the rest of the parameters at their benchmark values. We consider shocks of -1%, -0.5%, $+0.5\%$, and $+1\%$ to H, λ, b, and α. We then compute the average elasticity of each moment to an exogenous shock in each parameter averaging equilibrium moments across shocks of different magnitudes. For example, to compute average elasticities with respect to α, we regress the log of each moment to the log of α pooling together the baseline economy and the economies with exogenous variation in this parameter.

We compute elasticities for the share of agents in each occupation; the average plant size; the fraction of plants with 10 or fewer wage workers, 20 or fewer wage workers, 50 or more wage workers; and 100 or more wage workers; and the employment share of plants in each of these size bins (≤ 10, ≤ 20, ≥ 50, and ≥ 100). Table 11 summarizes our results.

The shape of the skill distribution λ and the difficulty of tasks H mainly affect the right tail of the size distribution and the employment share of these plants, but moments from the size distribution are more responsive in general to changes in α and b. The span of control parameter α affects more strongly the right tail (relative to the left tail)—the share of plants with more than 100 employees increases by 4% when α increases by 1%, and their share of employment increases by 7.5%. In contrast, changes in b affect more strongly moments from the left tail (relative to the elasticity of moments from the right tail). For example, increasing b by 1% increases the share of small plants (≤ 10) by 40%, and increases their share of employment by 46%.

The share of the self-employed is highly sensitive to changes in the calibrated parameters, especially to α and b. The share of the self-employed in economies close to the undistorted benchmark decreases by 13% when α increases by 1%, and by 15% when b increases by 1%. In contrast, the share of wage workers does not significantly respond to changes in the calibrated parameters, whereas the share of employers responds only to changes in b.

A.4. The cost of talent mismatch: decomposing changes in aggregate output

If we use the left tail of the skill distribution, then undistorted aggregate output is:

$$Y_u = \int_L^{z_1^u} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\, di$$

$$= \int_L^{z_1^d} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\, di + \int_{z_1^d}^{z_1^u} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\, di$$

$$= \int_{L}^{p^u\left(z_2^d\right)} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di + \int_{p^u\left(z_2^d\right)}^{z_1^d} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

$$+ \int_{z_1^d}^{z_1^u} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

Alternatively, we can use the right tail to obtain:

$$Y_u = \int_{z_2^u}^{H} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di$$

$$= \int_{z_2^u}^{z_2^d} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di + \int_{z_2^d}^{H} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di$$

$$= \int_{z_2^u}^{z_2^d} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di + \int_{z_2^d}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di$$

$$+ \int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di$$

Note that

$$\int_{L}^{p^u\left(z_2^d\right)} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di = \int_{z_2^u}^{z_2^d} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di \tag{14}$$

$$\int_{p^u\left(z_2^d\right)}^{z_1^d} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di = \int_{z_2^d}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di \tag{15}$$

and

$$\int_{z_1^d}^{z_1^u} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di = \int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di \tag{16}$$

Similarly, we can rewrite distorted aggregate output:

$$Y_d = \int_{L}^{z_1^d} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

$$= \int_{L}^{p^u\left(z_2^d\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di + \int_{p^u\left(z_2^d\right)}^{z_1^d} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

$$= \int_{L}^{p^u\left(z_2^d\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di + \int_{p^u\left(z_2^d\right)}^{p^d\left(m^u\left(z_1^d\right)\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

$$+ \int_{p^d\left(m^u\left(z_1^d\right)\right)}^{z_1^d} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

and

$$Y_d = \int_{z_2^d}^{H} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di$$

$$= \int_{z_2^d}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di + \int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di$$

$$= \int_{z_2^d}^{m^d\left(p^u\left(z_2^d\right)\right)} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di + \int_{m^d\left(p^u\left(z_2^d\right)\right)}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di$$

$$+ \int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di$$

where

$$\int_{L}^{p^u\left(z_2^d\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di = \int_{z_2^d}^{m^d\left(p^u\left(z_2^d\right)\right)} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di \tag{17}$$

$$\int_{p^u\left(z_2^d\right)}^{p^d\left(m^u\left(z_1^d\right)\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di = \int_{m^d\left(p^u\left(z_2^d\right)\right)}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di \tag{18}$$

and

$$\int_{p^d\left(m^u\left(z_1^d\right)\right)}^{z_1^d} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di = \int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di \tag{19}$$

Then the change in aggregate output is

$$Y_d - Y_u = \int_{L}^{p^u\left(z_2^d\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di + \int_{p^u\left(z_2^d\right)}^{p^d\left(m^u\left(z_1^d\right)\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

$$+ \int_{p^d\left(m^u\left(z_1^d\right)\right)}^{z_1^d} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

$$-\int_{z_2^u}^{z_2^d} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di - \int_{z_2^d}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di$$

$$-\int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di$$

Using equations (16) and (19), we rewrite

$$Y_d - Y_u = \int_{L}^{p^u\left(z_2^d\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di + \int_{p^u\left(z_2^d\right)}^{p^d\left(m^u\left(z_1^d\right)\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

$$+\int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di$$

$$-\int_{z_2^u}^{z_2^d} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di - \int_{z_2^d}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di$$

$$-\int_{z_1^d}^{z_1^u} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di$$

Finally, we regroup terms to obtain

$$Y_d - Y_u = \left[-\int_{z_2^u}^{z_2^d} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di + \int_{L}^{p^u\left(z_2^d\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di\right]$$

$$\left[-\int_{z_2^d}^{m^u\left(z_1^d\right)} G(i)\left[n\left(p^u(i)\right)\right]^{\alpha} f(i)\,di + \int_{p^u\left(z_2^d\right)}^{p^d\left(m^u\left(z_1^d\right)\right)} G\left(m^d(i)\right)[n(i)]^{\alpha-1} f(i)\,di\right]$$

$$\left[-\int_{z_1^d}^{z_1^u} G\left(m^u(i)\right)[n(i)]^{\alpha-1} f(i)\,di + \int_{m^u\left(z_1^d\right)}^{H} G(i)\left[n\left(p^d(i)\right)\right]^{\alpha} f(i)\,di\right]$$

With the policy, managers $\left[z_2^u, z_2^d\right]$ become self employed, but their teams $\left[L, p^u\left(z_2^d\right)\right]$ are matched with better managers relative to the undistorted equilibrium. The first term captures the net loss from this reallocation. Similarly, with the policy, wage workers $\left[z_1^d, z_1^u\right]$ become self-employed, which destroys their teams, but their managers $\left[m^u\left(z_1^d\right), H\right]$ are assigned into smaller teams. The third term captures the net cost of this re-assignment. The second term is a pure talent mismatch: managers coordinate lower skill employees in smaller teams.

A.5. Lucas (1978) economy

Production. We try to stay as close as possible to our original environment. While this setup is somewhat restrictive, it allows for a straightforward comparison to our benchmark economy with heterogeneous workers and knowledge hierarchies.

The skill distribution, denoted by $F(\cdot)$, takes the form of a double-truncated exponential with parameter λ, over the interval $[L, H]$. Problems are drawn from a distribution $G(\cdot)$, which is uniform over the interval $[L, H]$. Rather than matching

Table 12
Parameter values in Lucas (1978) Economy—U.S. (undistorted).

Parameter	Value	Source
L	1	Assigned
α	0.76	Joint calibration
H	12.5	Joint calibration
λ	1.25	Joint calibration

Table 13
Calibration results: U.S. (undistorted).

Targeted	Data	Model
Average est. size	17.72	17.46
Share of est. of size < 20	0.845	0.792
Share of est. of size > 50	0.058	0.064
Non-targeted	Data	Model
Share of est. of size > 100	0.025	0.023
Share of est. of size > 250	0.007	0.004
Emp. share of est. < 20	0.25	0.337
Emp. share of est. > 50	0.583	0.422

in terms of quality (skill), teams of varying sizes are formed by managers of different skills as in a standard Lucas (1978) span-of-control economy.

A manager of skill z_m can solve problems of difficulty $G(z_m)$ (his productivity), and has $n(z_m)$ units of labor (his span of control) to work with. There is no communication between the team layers. The manager's problem is then

$$R(z_m) = \max_n \; G(z_m)\, n^{\alpha} - wn. \tag{20}$$

The span of control of the manager is then given by

$$n(z_m) = \left[\frac{\alpha G(z_m)}{w}\right]^{\frac{1}{1-\alpha}}$$

Occupational choices. Individuals draw their skill level from $F(\cdot)$ and then decide whether to become a wage worker, or to manage a team.

$$V(\tilde{z}) = \max\{w, R(\tilde{z})\}.$$

Equilibrium. A competitive equilibrium in this economy is a wage, and managerial rents, such that, given the wage,

1. managers maximize rents,
2. no agent wishes to change occupations, and
3. the labor market clears.

The equilibrium in this economy is characterized by a skill thresholds z^*, such that those with skill level below z^* choose to become workers and earn the same wage w, while those with skill greater than z^* become managers and earn $R(z)$.

Calibration and quantitative exercises. We set the lower bound for the skill distribution equal to 1, and then calibrate jointly the rest of the parameters, which are the span of control parameter α, the upper bound of the skill distribution H, and its exponential coefficient λ. The targets are the average establishment size, the share of establishments with less than 20 employees, and the share of establishments with more than 50 employees in the U.S. economy. We choose the parameters to minimize the absolute relative deviation between the model and the data, as in our original calibration of the benchmark economy. Parameter values are shown in Table 12, while Table 13 contains the calibration results.

Table 14 reports the losses associated with two versions of the size-dependent tax policy introduced into our undistorted U.S. benchmark. First, in column (1) we introduce a policy with the same level and enforcement parameters used in our U.S. counterfactual. The average establishment size is cut nearly in half, and output drops by less than three percent. In our U.S. benchmark economy with positive sorting the same policy generates a decrease in average establishment size that is five times smaller than in the Lucas (1978) economy, but output losses are more than threefold. Then, in column (2) we report the losses associated with a tax that generates the same average establishment size drop as in our U.S. benchmark, which results in output losses that are nearly zero. Thus, size-dependent taxes in an economy with knowledge hierarchies and positive sorting generate output losses much larger than that in a standard span-of-control model.

Table 14
Losses from different policies in the Lucas (1978) version.*

	Size-dependent tax ($\tau_s = 0.44$, $\kappa = 0.0658$)	Size-dependent tax ($\tau_s = 0.05$, $\kappa = 0.0658$)
Avg. est. size	0.517	0.904
Output	0.973	0.999

* Relative to undistorted scenario calibrated to the U.S.

Appendix B

B.1. US data

Data for the U.S. establishment size distribution come from available tabulates from the Census Bureau's Longitudinal Business Database. See https://www.census.gov/programs-surveys/bds/data/data-tables/legacy-establishment-characteristics-tables-1977-2014.html. We use data for 2014.

The occupation shares come from the Bureau of Labor Statistics, March, 2016 report Self-Employment In The United States—see Hipple and Hammond (2016). We do back-of-the-envelope calculations to arrive at our occupation shares for the U.S. The report shows that in 2015, 15 million people, or 10.1 percent of the U.S. labor force, were "self-employed" (this includes those with and without employees, as well as those incorporated and unincorporated). Out of those 15 million, 9.5 million are unincorporated and the remaining 5.5 million are incorporated. Non-employers represent 86% of unincorporated businesses, and 58% of incorporated businesses. Thus, the share of non-employers (what we call self-employed in the model) is

$$0.101 \times (0.58 \times (5.5/15) + 0.86 \times (9.5/15)) = 0.076.$$

That is, 7.6% of the U.S. labor force runs a business without employees, 2.5% runs a business with employees, and the remaining 89.9% work for a wage.

Last, we calculate the education level shares by firm size categories reported in the left panel of Fig. 13 using the 2016 March Supplement of the Current Population Survey available in the NBER repository https://data.nber.org/data/current-population-survey-data.html. We keep full-year workers ages 25-65, in areas with more than 100,000 habitants, not in school, in the private sector, with only one employer. The share of business owners in this population is also 10.1%. Additional details can be found in the readme.txt file that is part of the replication files.

B.2. Mexico data

Data for the establishment size distribution in Mexico come from Busso et al. (2018). They consider only establishments in Manufacturing, Wholesale and retail, and Services. In addition, they exclude around 2% of observations in these sectors due to measurement error or for consistency purposes. We refer the reader to the Data Appendix in Busso et al. (2018) for additional details. We use moments for 2013, the last census wave available.

The size of the establishment in the census data includes the business owner, and therefore we subtract 1 from this reported size for our moments to be comparable with those from the US. To construct the effective social security contribution rates used to calibrate our policy distortion in Fig. 9, we sum total security contributions and total wage bill across establishments within each size bin from 1 to 99, and then take the ratio of these two numbers to obtain the average.

Similarly, to compute the relative average wages in Fig. 11, we first take the ratio of total wage bill to total number of workers to obtain the economy-wide average wage. We then repeat the same procedure to obtain average wages for each size category. Last, we take the ratio of average wages in each size category to the economy-wide average wage.

The series on fraction of workers across occupations over time in Fig. 8—as well as the number for 2017 used in our calibration and in Fig. 13—come from the nationally representative surveys National Survey of Urban Employment, which was administered until 2004, and National Survey of Occupations and Employment (ENEU and ENOE by their Spanish acronyms). The data is publicly available online at https://www.inegi.org.mx/programas/enoe/15ymas for the ENOE (for 2005-2017) and https://www.inegi.org.mx/programas/eneu/2004/ for the ENEU (1994-2004). We work with data for the third quarter of each year.

We select workers ages 25-64, in the private sector, who only have one job (that is, business ownership is not a secondary occupation), who work full year in their occupation, no longer in school, and in the 28 cities that have been sampled between 1994 and 2017. In addition, we drop workers with no schooling or with a graduate degree. Additional details can be found in the readme.txt file that is part of the replication files.

References

Adamopoulos, T., Restuccia, D., 2014. The size distribution of farms and international productivity differences. The American Economic Review 104 (6), 1667–1697.

Alaimo, V., Bosch, M., Gualavisi, M., Villa, J.M., 2017. Measuring the cost of salaried labor in Latin America and the Caribbean. Technical report. Inter-American Development Bank.
Alder, S.D., 2016. In the wrong hands: complementarities, resource allocation, and TFP. American Economic Journal: Macroeconomics 1 (8), 199–241.
Benabou, R., 2002. Tax and education policy in a heterogeneous-agent economy: what levels of redistribution maximize growth and efficiency? Econometrica 70 (2), 481–517.
Bento, P., Restuccia, D., 2017. Misallocation, establishment size, and productivity. American Economic Journal: Macroeconomics 3 (9), 267–303.
Bhattacharya, D., Guner, N., Ventura, G., 2013. Distortions, endogenous managerial skills and productivity differences. Review of Economic Dynamics 16 (1), 11–25.
Braguinsky, S., Branstetter, L.G., Regateiro, A., 2011. The incredible shrinking Portuguese firm. Technical report. National Bureau of Economic Research.
Busso, M., Levy, S., Neumeyer, A., Spector, M., 2012. Skills, informality and the size distribution of firms. Buenos Aires, Argentina: Universidad Torcuato di Tella. Available at: https://dl.dropboxusercontent.com/u/499791/papers/Informality_December2012_2.pdf.
Busso, M., Levy, S., Torres, J., 2018. Labor regulations and resource misallocation in Mexico. Unpublished manuscript (June).
Cardiff-Hicks, B., Lafontaine, F., Shaw, K., 2014. Do large modern retailers pay premium wages? Technical report. National Bureau of Economic Research.
Fox, J.T., 2009. Firm-size wage gaps, job responsibility, and hierarchical matching. Journal of Labor Economics 27 (1), 83–126.
Garicano, L., Hubbard, T.N., 2012. Learning about the nature of production from equilibrium assignment patterns. Journal of Economic Behavior & Organization 84 (1), 136–153.
Garicano, L., Lelarge, C., Van Reenen, J., 2016. Firm size distortions and the productivity distribution: evidence from France. The American Economic Review 106 (11), 3439–3479.
Garicano, L., Rossi-Hansberg, E., 2004. Inequality and the organization of knowledge. The American Economic Review: Papers and Proceedings 94 (2), 197–202.
Garicano, L., Rossi-Hansberg, E., 2006. Organization and inequality in a knowledge economy. The Quarterly Journal of Economics 121 (4), 1383–1435.
Geerolf, F., 2017. A theory of Pareto distributions. Unpublished manuscript.
Guner, N., Parkhomenko, A., Ventura, G., 2018. Managers and productivity differences. Review of Economic Dynamics 29, 256–282.
Guner, N., Ventura, G., Xu, Y., 2008. Macroeconomic implications of size-dependent policies. Review of Economic Dynamics 11 (4), 721–744.
Headd, B., 2000. The characteristics of small-business employees. Monthly Labor Review 123, 13.
Hipple, S.F., Hammond, L.A., 2016. Self-Employment in the United States. Technical report. U.S. Bureau of Labor Statistics.
Hsieh, C.-T., Klenow, P.J., 2009. Misallocation and manufacturing TFP in China and India. The Quarterly Journal of Economics 124 (4), 1403–1448.
Kremer, M., Maskin, E., 1996. Wage Inequality and Segregation by Skill. National Bureau of Economic Research. Working Paper w5718.
Levy, S., López-Calva, L.F., 2016. Labor Earnings, Misallocation, and the Returns to Education in Mexico. IDB Working Paper Series No IDB-WP-671.
López, J.J., 2017. A quantitative theory of tax evasion. Journal of Macroeconomics 53, 107–126.
López, J.J., 2020. Modeling progressive taxation. Unpublished manuscript.
Lucas, R.E., 1978. On the size distribution of business firms. Bell Journal of Economics 9 (2), 508–523.
Luttmer, E.G.J., 2007. Selection, growth, and the size distribution of firms. The Quarterly Journal of Economics 122 (3), 1103–1144.
OECD, 2017. Entrepreneurship at a Glance 2017. OECD.
Poschke, M., 2018. The firm size distribution across countries and skill-biased change in entrepreneurial technology. American Economic Journal: Macroeconomics 3 (10), 1–41.
Restuccia, D., Rogerson, R., 2008. Policy distortions and aggregate productivity with heterogeneous establishments. Review of Economic Dynamics 11 (4), 707–720.
Rosen, S., 1981. The economics of superstars. The American Economic Review 71 (5), 845–858.
Rossi-Hansberg, E., Wright, M.L.J., 2007. Establishment size dynamics in the aggregate economy. The American Economic Review 97 (5), 1639–1666.
Sattinger, M., 1993. Assignment models of the distribution of earnings. Journal of Economic Literature 31 (2), 831–880.
Scheuer, F., Werning, I., 2017. The taxation of superstars. The Quarterly Journal of Economics 132 (1), 211–270.
Tamkoç, M.N., 2019. Production complexity, talent misallocation and development. Arizona State University. Unpublished manuscript.
Torres, J., 2018. The returns to entrepreneurship: selection, non-pecuniary benefits, and necessity in Mexico. Unpublished manuscript.