Labor Market Power with Worker and Firm Heterogeneity

David Berger
Duke University and NBER

Kyle Herkenhoff
University of Minnesota, Federal Reserve Bank of Minneapolis and NBER

Jaehun Jeong
Duke University

Simon Mongey
Federal Reserve Bank of Minneapolis and NBER

July 2026

ABSTRACT

High-wage workers are less responsive to wage differences across firms, and sort into high-wage firms. Hence, the composition of a firm's workforce is central to its wage setting. To accommodate these facts, we extend the neoclassical model of labor supply to include a discrete choice over which firm to work at, and labor demand to include production complementarities and strategic interaction between granular firms. We estimate the model on wage decompositions into firm, worker and sorting components, reduced-form labor supply elasticities across workers and firms, and earnings responses to unanticipated transfers. Half of the aggregate markdown is policy-variant. Worker heterogeneity accounts for 78 percent of this component, and 89 percent of the markdown gap between large and small firms. Analytically, we derive two implications for tax policy: higher progressivity widens markdowns and worsens sorting between workers and firms. These channels are necessary to obtain large declines in earnings following increases in marginal tax rates, consistent with empirical studies. They also reduce optimal marginal tax rates by one-third, and halve the associated welfare gains.

Email: david.berger@duke.edu, kfh@umn.edu, jaehun.jeong@duke.edu, simonmongey@gmail.com. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

Workers are heterogeneous in their responsiveness to differences in wages across firms: high wage workers are less elastic, low wage workers are more elastic (Seegmiller, 2025; Bils et al., 2025; Volpe, 2025). Meanwhile, high wage workers sort into high wage firms and low wage workers sort into low wage firms (Bonhomme et al., 2019; Kline et al., 2020). Together, these imply that answers to two questions are determinative of firms' wage setting: *Who works for me?* and *How elastic are they?* These concerns are absent from the neoclassical theory of labor in which workers trade-off only hours and consumption, and firms are homogeneous.

To address these concerns, this paper extends the neoclassical theory of labor and quantifies the implications of doing so for wages, markdowns, earnings responses to tax changes and optimal tax progressivity. We extend *labor supply* to incorporate evidence on heterogeneity in labor supply elasticities across workers, by adding a discrete choice across heterogeneous employers. Workers' endogenous marginal values of wealth that shape their intensive margin elasticity of how many hours to work in the neoclassical model, now shape their extensive margin elasticity of which firm to work at. Workers are heterogeneous in their productivity, non-labor income, idiosyncratic preferences for jobs, and marginal tax rates. We extend *labor demand* to incorporate evidence on sorting, concentration in labor markets, and labor supply elasticity differences across firms, by adding production complementarities and strategic interaction between granular firms. Firms are heterogeneous in overall productivity, and relative productivity across worker types (Costinot and Vogel, 2010).

With market power, firms internalize differences in workers' extensive and intensive margin labor supply elasticities when setting wages, hence workers' economic circumstances affect wage-setting, markdowns, and sorting. Closed form expressions characterize *Who works for me?* in terms of sorting, and *How elastic are they?* on the intensive (hours) and extensive (firm choice) margin. The latter is shaped by preference heterogeneity, marginal tax rates, the firm's employment share, and the worker's marginal value of additional wages at the firm. If income effects dominate substitution effects, then this last piece implies that (a) *across workers*, richer workers are less elastic, and (b) *within workers, across firms*, all workers are less elastic when being paid higher wages. We find that this new component of labor supply elasticities is determinative of markdowns in the aggregate and across firms, the response of the allocation of labor to changes in taxes, and the shape of optimal taxes.

We identify and estimate key model parameters governing markdowns and sorting from three facts. First, higher-earning workers are less elastic in their labor supply to the firm, both unconditionally and conditional on firm type (Seegmiller, 2025; Bils et al., 2025; Volpe, 2025). Second, high-productivity workers sort to high-productivity firms (Bonhomme et al., 2019; Kline et al., 2020). Third, conditional on worker type, labor supply to high-productivity

firms is less elastic (Seegmiller, 2025; Bils et al., 2025; Yeh et al., 2022; Lobel, 2021).[1] Through the lens of our framework, the first fact points to wealth effects: the extensive-margin elasticity only declines with income when wealth effects are sufficiently strong. However, once high earners become less elastic to wages, they tend to sort to lower-wage firms, a counterfactual implication. Thus matching the positive sorting in the data requires complementarities in production. This raises the wages of high-skill workers relatively more at productive firms. Conditional on complementarities and preferences, any remaining difference in labor supply elasticities across firms pins down parameters that determine oligopsony. Remaining parameters are pinned down by declines in earnings following lottery wins (Golosov et al., 2024), consumption inequality (Heathcote et al., 2020b), and the small share of firm fixed effects in empirical decompositions of earnings dispersion (Bonhomme et al., 2019; Kline et al., 2020).

The calibration itself shows that integrating the discrete choice model into the neoclassical theory of labor supply yields a simple framework capable of rationalizing an important set of empirical facts. The model rationalizes these data under identical preferences across workers, with standard functional forms and values of preference parameters. The calibration reveals that heterogeneity in labor supply elasticities across individuals and firms is almost entirely a matter of the extensive margin of firm choice. The mean overall elasticity is 3.79, while the mean extensive margin elasticity is 3.86. Consistent with empirical estimates, the (Marshallian) elasticity on the hours margin is tightly clustered around zero. This is key for what follows as we shift the focus of wage setting and optimal policy away from small, inelastic components of labor supply (hours) to a large, elastic component (which firm to work at).

We find that worker heterogeneity is a key determinant of markdowns. The earnings weighted markdown in our economy is 35 percent. Half of this is fixed, reflecting dispersion in workers' idiosyncratic tastes for employers, and would not respond to policy. The other half is variable, and heterogeneity in workers' marginal value of wealth is the dominant source. Differences in elasticities across workers account for 65 percent of the variable component, and lower elasticities of any given worker when paid more at higher wage firms account for a further 13 percent. Oligopsony and tax progressivity account for the remainder. The same forces explain differences across firms: whom firms employ and how elastic those workers are, accounts for 89% of the markdown gap between low and high productivity firms.

These results imply a sharp trade-off between efficiency and redistribution, which we study through the lens of tax progressivity. We begin by studying the implications of competitive wages, which are the hallmark of an efficient economy. Since high productivity workers at high productivity firms face the widest markdowns in equilibrium, their wages increase the most. Removing markdowns alone increases the standard deviation of log consumption by 16 percent. In response to higher wages, high productivity workers increasingly sort toward high

[1] These studies are in the U.S. and Brazil. Volpe (2025) finds the higher wage firms face more elastic labor supply in Norway.

productivity firms. Productivity increases and consumption inequality increases by 27 percent. Higher tax progressivity will do the opposite: compress consumption inequality, reduce sorting and reduce productivity.

Higher progressivity reduces workers' elasticity of labor supply across firms. We have three results. First, equilibrium comparative statics show that raising progressivity widens markdowns and worsens sorting and aggregate productivity. An increase in the marginal tax rate at the 95th percentile of the income distribution from 30 percent to 50 percent widens markdowns by 3 ppt, leading wages to fall by nearly 5 percent. As workers become less elastic with respect to wages, sorting deteriorates, reducing productivity. This accounts for more than half of the decline in output, which falls by 3 percent. In the aggregate, an increase in taxes shows up as a negative shift in labor demand as supply rotates and markdowns widen in response. These effects are large, but are they empirically relevant?

Second, we show that endogenous markdowns and sorting are necessary for the model to match large empirical estimates of the response of earnings to changes in marginal tax rates. We analyze the *Elasticity of Taxable Income* (ETI): how much log average earnings decline following an increase in marginal tax rates. In their survey, Saez et al. (2012) put the ETI between 0.12 and 0.40, while Kleven and Schultz (2014) obtain 0.20 to 0.30 for a large tax reform in Denmark. Our model delivers an ETI of 0.24. The endogenous response of markdowns and sorting account for 80 percent of our ETI, while reductions in hours at existing matches and wages account for only 20 percent. This is consistent with Kleven et al. (2025), who estimate a small ETI on *stayers* and much larger ETI on *switchers*. Hence the markdown and sorting responses to taxes are not just theoretical possibilities. They allow the model to be consistent with empirical studies of the effects of taxes on earnings.

Third, we show that the markdown and sorting channels lead to lower optimal tax progressivity. We provide an analytical expression for optimal tax progressivity which balances the marginal welfare gains from redistribution against three distinct costs associated with hours, sorting and markdowns. First, as in the standard model, high progressivity increases the behavioral distortion associated with hours. As hours are generally inelastic, this yields high optimal tax rates. Second, the sorting channel, by which higher progressivity reduces sorting toward high productivity firms, lowers aggregate labor productivity. As firm choice is generally elastic, this substantially reduces optimal tax rates. Third, the markdown channel, by which higher progressivity widens markdowns, represents a revenue loss to the government if profits are not fully taxed. Since we have little to say about profit taxes, we focus on the sorting channel. Optimal marginal tax rates at the 95th percentile of income fall by around one third. For example, from 73% to 50% under a Frisch elasticity of 0.10, or from 43% to 33% under a Frisch elasticity of 0.90. The welfare gains associated with optimizing tax progressivity are halved once the sorting channel is incorporated.

This paper sits at the intersection of two literatures: (i) the treatment of inequality in macroe-

conomics and public economics, with a particular focus on labor income taxes, and (ii) the treatment of firms' market power in labor markets.

Inequality in macroeconomics and optimal labor taxes. In a sequence of papers, Heathcote et al. (2014) and Heathcote et al. (2017) develop and estimate tractable heterogeneous agent economies to (i) measure how consumption insurance has evolved over time and (ii) to prescribe optimal tax progressivity, respectively. In these papers, labor demand is obtained from competitive, homogeneous firms, and workers are perfect substitutes. Sachs et al. (2020) study optimal non-linear taxes in general equilibrium, where wages reflect (endogenous) marginal products of labor. Berger et al. (2024) add progressive taxes to Berger et al. (2022), but abstract from worker heterogeneity and sorting. We contribute to the literature by incorporating non-competitive wage setting, worker heterogeneity, firm heterogeneity and sorting. Importantly, our focus is on a margin of labor supply that is much more elastic—the extensive margin decision of which firm to work at—rather than the intensive margin decision of hours to supply.

Another approach is to study optimal progressivity in Mirrleesian assignment frameworks. Scheuer and Werning (2017) study competitive one-to-one assignment: workers match to firms under supermodular production, generating a convex earnings schedule. Workers reallocate across firms in response to tax changes, which amplifies earnings elasticities and creates a force for lower marginal taxes. Hummel (2023) studies optimal taxes in an economy with a single monopsonist and homogeneous workers. Our approach is quantitative: we model multi-worker firms with a calibrated firm size distribution, oligopsony, endogenous hours, complementarities in production, and heterogeneous marginal utilities of wealth, and identify parameters from micro data on labor supply elasticities, sorting, and wage decompositions.

Taxation has also been studied in search and matching economies (Arseneau and Chugh, 2012, 2023; Bagger et al., 2021). Arseneau and Chugh (2012) study optimal taxes in a representative-household search model based on Andolfatto (1996), showing that tax-smoothing is in general not optimal when wages are rigid. This model features a well-defined labor supply elasticity and smoothing motives for the representative household, but omits the type of firm heterogeneity, sorting and strategic behavior present in our framework. Bagger et al. (2021) study optimal taxation in a linear-utility random search model in which workers climb job ladders to improve job match qualities. They assume that firms are ex-ante identical and workers are risk-neutral. We contribute to this literature by allowing for ex-ante two-sided heterogeneity, heterogeneous marginal utilities of wealth, endogenous labor supply as in MaCurdy (1981), and multi-worker firms calibrated to match the observed firm size distribution.

Labor. The labor literature on the causes and consequences of monopsony power in the U.S. has grown rapidly (e.g. Berger et al., 2022; Lamadon et al., 2022; Jarosch et al., 2024; Volpe, 2025). The majority of this literature maintains the assumption of additive random utility and constant-hours wage contracts. Risk neutrality neutralizes any role for non-labor income in

sorting, labor supply, and firm wage setting decisions. Constant-hours wage contracts disconnect labor supply decisions from their neoclassical treatment (MaCurdy, 1981). Mousavi (2021) considers monopsony with choice of hours and firms, but where firms are atomistic and homogeneous. Lorenzini (2026) extends Berger et al. (2022) to study how complementarity in firm and worker types can increase labor market power even for small firms in a model with full insurance within the family.

Our paper contributes to the ongoing debate over the nature of firm premia in the U.S. (Bonhomme et al., 2019; Borovičková and Shimer, 2024; Kline, 2024). Bonhomme et al. (2019) argue that firm premia explain a small fraction of wage variance in the data; however, they also find that high-type workers are disproportionately employed at firms with high premia. Borovičková and Shimer (2024) argue that these facts are consistent with log supermodular production and selection effects based on match quality. We show that a neoclassical model of labor supply with labor market power is consistent with low firm wage premia and strong complementarities in production. With competitive wages and complementarities, high skill workers would get paid disproportionately more than other workers, leading to a significant departure of wages from log additivity in firm and worker types. With labor market power and worker heterogeneity, these workers' wages are depressed due to wider markdowns, since they are exactly the least wage sensitive workers.

Our model relates to work by Volpe (2025). Volpe (2025) extends Lamadon et al. (2022) to include exogenous heterogeneity across workers in their marginal utilities of wages, and hence their elasticity of labor supply. This generates sorting: workers with the highest marginal utilities sort into the highest wage firms. For a given worker type, higher paying firms will therefore have *more wage-elastic employment*. This is consistent with Norwegian data. However, in the U.S. context, Seegmiller (2025) shows that with higher paying firms face *less wage-elastic employment*. This holds in general, and conditional on worker type. This is consistent with findings in Brazil (Lobel, 2021; Bils et al., 2025). We show that a neoclassical model of labor supply with worker and firm heterogeneity, complementarity in production and labor market power can fit these facts in the U.S., while simultaneously being consistent with decompositions of wage inequality into firm, worker and sorting components as well as recent evidence on declines in earnings following shocks to unearned income (Golosov et al., 2024).

Overview. Section 2 lays out the model. Section 3 characterizes theoretical properties of labor supply elasticities to firms and sorting. Section 4 calibrates the model. Section 5 applies the model to study the sources of markdowns in general, their differences across firms and implications for inefficiency and redistribution. Section 6 applies the model to study the implications for optimal tax progressivity. Section 7 concludes.

2. Model

2.1. Environment

Agents. The economy consists of two types of agents, workers and firms. Firms are indexed by j and exogenously distributed across labor markets indexed by m. There is a finite number J_m of firms in each market, and a large number of markets $m \in \{1, \ldots, M\}$.[2] Firms have heterogeneous productivity levels z_{jm}. A measure $\overline{\Gamma}$ of workers is indexed by i. Workers are ex-ante heterogeneous in their skill level e^i and preferences for each firm ζ^i_{jm}. Let $\Gamma(e)$ denote the measure of workers with skill level e. Section 4 adds heterogeneity in non-labor income.

Preferences. Workers value consumption c^i, dislike hours worked h^i, and have relative idiosyncratic utilities for working at each firm $\boldsymbol{\zeta}^i = (\zeta^i_{11}, \ldots, \zeta^i_{J_M M})$. Worker i's utility from working at firm j in market m is given by,

$$U^i_{jm} = \frac{c^{i\,1-\sigma}_{jm}}{1-\sigma} - \frac{1}{\overline{\varphi}^{1/\varphi}} \frac{h^{i\,1+1/\varphi}_{jm}}{1+1/\varphi} + \zeta^i_{jm}. \tag{1}$$

We assume that the vector $\boldsymbol{\zeta}^i$ is drawn from a nested Gumbel distribution with parameter θ governing the inter-market variance of draws and η governing the intra-market variance of draws. The hours disutility shifter $\overline{\varphi}$ will be chosen to match data on average hours worked.

We nest three specifications adopted in the literature. First, absent firm specific shifters, these are simply MaCurdy (1981) preferences, and balanced growth under $\sigma = 1$. Such preferences are commonly deployed in settings that evaluate optimal tax progressivity with incomplete markets in public economics and macroeconomics.[3] Second, under log utility, and absent an hours component, labor supply is as in Lamadon et al. (2022). If infinitely many firms are in each market, then the competitive structure of labor demand is also as in Lamadon et al. (2022). Third, under log utility and Gumbel distributed ζ^i_{jm}, we can apply the aggregation results from Verboven (1996) as in Berger et al. (2022) to obtain nested-CES utility. We find that departing from log is necessary for the model to match reduced form estimates of differences in labor supply elasticities to firms of low and high wage workers.

Technology. An allocation is a share of workers $\rho_{jm}(e)$ of each worker type employed at each firm, their hours $h_{jm}(e)$, and consumption $c_{jm}(e)$. Given an allocation, output of firm jm is

$$x_{jm} = \int A\big(z_{jm}, e\big) n_{jm}(e) d\Gamma(e) \quad , \quad n_{jm}(e) = \rho_{jm}(e) h_{jm}(e). \tag{2}$$

The measure of workers of type e is $\Gamma(e)$. Of these, a share $\rho_{jm}(e)$ work at firm jm, and work $h_{jm}(e)$ hours. They produce $A(z_{jm}, e)$ units of output per hour worked , which depends on

[2] There is an arbitrarily large number of markets, such that each firm has no influence on aggregates.

[3] For example, Heathcote et al. (2017) use log preferences in their benchmark exercises.

worker skill e (Costinot and Vogel, 2010).

2.2. Worker problem

Workers face a distribution of wages per hour worked from all firms in the economy $w_{jm}(e)$ and maximize utility subject to the budget constraint: $c^i_{jm} = w_{jm}(e^i)h^i_{jm}$. When calibrating the model, we will add a tax and transfer system, and non-labor income. We use i to capture all heterogeneity across individuals, e.g. $w^i_{jm} = w_{jm}(e^i)$ and $A^i_{jm} = A\big(z_{jm}, e^i\big)$.

To make clear how workers' elasticities of labor supply are constructed, we attach the Lagrange multiplier λ^i_{jm} to the budget constraint conditional on working at firm jm:

$$\max_{j,m}\Big\{V^i_{jm} + \zeta^i_{jm}\Big\} \quad , \quad V^i_{jm} = \max_{c^i_{jm}, h^i_{jm}} u\Big(c^i_{jm}, h^i_{jm}\Big) \quad \text{s.t.} \quad c^i_{jm} = w^i_{jm}h^i_{jm} \quad \Big[\lambda^i_{jm}\Big].$$

The extensive margin decision of the household is captured by the households' *choice probability*, ρ^i_{jm}. This expression is standard given the nested-Gumbel structure of idiosyncratic preferences. Let $v^i_{jm} = \exp\{V^i_{jm}\}$, then:

$$\rho^i_{jm} = \left(\frac{v^i_{jm}}{\widetilde{v}^i_m}\right)^{\eta}\left(\frac{\widetilde{v}^i_m}{\overline{v}^i}\right)^{\theta} \quad , \quad \widetilde{v}^i_m = \left[\sum_{j\in m}\Big(v^i_{jm}\Big)^{\eta}\right]^{1/\eta} \quad , \quad \overline{v}^i = \left[\sum_{m}\Big(\widetilde{v}^i_m\Big)^{\theta}\right]^{1/\theta}. \tag{3}$$

Exponentiating V^i_{jm}, allows us to state the choice probability in a way that avoids 'log-sums,' and reflects the constant elasticity features imparted by the nested-Gumbel distribution.

The intensive margin decision of the household is made conditional on having already chosen to work at firm jm and characterized by standard first order conditions:

$$\Big(c^i_{jm}\Big)^{-\sigma} = \lambda^i_{jm} \quad , \quad \frac{1}{\overline{\varphi}^{1/\varphi}}h^{i\,1/\varphi}_{jm} = \lambda^i_{jm}w^i_{jm} \quad , \quad h^i_{jm} = \overline{\varphi}(w^i_{jm})^{\varphi}(c^i_{jm})^{-\sigma\varphi}.$$

2.3. Firm problem

Firms set wages that discriminate by worker skill e. We will maintain this assumption when we introduce non-labor income, assuming that non-labor income is not observed by the firm. The firm chooses wages of each worker type to solve:

$$\max_{\{w_{jm}(e)\}}\int A\big(z_{jm}, e\big)n_{jm}(e)d\Gamma(e) - \int w_{jm}(e)n_{jm}(e)\,d\Gamma(e) \quad \text{s.t.} \quad n_{jm}(e) = \rho_{jm}(e)h_{jm}(e).$$

Similar to the worker problem, we economize on notation by letting superscript i encapsulate skill heterogeneity (for example, the labor supply elasticity is denoted $\varepsilon^i_{jm} = \varepsilon^i_{jm}(e)$). The optimality condition for wages of skill level e^i workers at firm jm can be represented as a

markdown on the marginal revenue product of labor:

$$w^i_{jm} = \mu^i_{jm} A^i_{jm} = \frac{\varepsilon^i_{jm}}{\varepsilon^i_{jm}+1} A^i_{jm} \quad , \quad \varepsilon^i_{jm} = \left.\frac{\partial \log n^i_{jm}}{\partial \log w^i_{jm}}\right|_{\mathbf{w}^i_{-jm}} = \underbrace{\left.\frac{\partial \log \rho^i_{jm}}{\partial \log w^i_{jm}}\right|_{\mathbf{w}^i_{-jm}}}_{\text{Extensive margin: } \varepsilon^{i,\rho}_{jm}} + \underbrace{\left.\frac{\partial \log h^i_{jm}}{\partial \log w^i_{jm}}\right|_{\mathbf{w}^i_{-jm}}}_{\text{Intensive margin: } \varepsilon^{i,h}_{jm}}$$

The labor supply elasticity is computed taking competitors' wages as given and has an *extensive* and *intensive* margin component. We derive these below.

2.4. Equilibrium

The equilibrium concept is consistent with the granularity of agents in the economy. Workers are infinitesimal and hence competitive. Firms are infinitesimal in the aggregate and hence competitive, however within their own markets there are finitely many firms and hence their behavior is strategic. The conduct assumed for this strategic behavior is a Nash equilibrium—i.e. static, one-shot, *oligopsony*—where actions are in wages and payoffs are in profits. The key equilibrium consistency condition is that the labor supply curve for each type of labor to each firm $n^i_{jm}(w^i_{jm}, \mathbf{w}^i_{-jm})$ is consistent with worker optimality. Below we consider alternative behavioral assumptions, in particular that firms take wages as given (i.e. competitive).

3. Characterization - Labor supply elasticities and sorting

We characterize how workers' labor supply elasticities vary across workers and firms, and the implications of these for sorting.

3.1. Labor supply elasticities

Extensive margin elasticity. The extensive margin elasticity reflects workers' tendency to move away from firm jm and toward any other firm as firm jm cuts its wage. First, we use conditional choice probabilities to express extensive margin supply (3):

$$\rho^i_{jm} = \rho^i_{j|m}\rho^i_m, \quad \rho^i_{j|m} = \left(\frac{v^i_{jm}}{\widetilde{v}^i_m}\right)^\eta, \quad \rho^i_m = \left(\frac{\widetilde{v}^i_m}{\overline{v}^i}\right)^\theta$$

Second, we differentiate to obtain the extensive margin elasticity, which has two components:

$$\varepsilon^{\rho,i}_{jm} = \frac{\partial \log \rho^i_{jm}}{\partial \log w^i_{jm}} = \frac{\partial \log \rho^i_{jm}}{\partial \log v^i_{jm}} \times \frac{\partial \log v^i_m}{\partial \log w^i_{jm}} = \left[\eta(1-\rho^i_{j|m}) + \theta\rho^i_{j|m}\right] \times \left(\lambda^i_{jm} w^i_{jm} h^i_{jm}\right). \tag{4}$$

The first component captures how sensitive an individuals' choice is to the value represented by a firm in the market, v^i_{jm}. This nests existing *size-based* theories of market power as in Berger et al. (2022). In markets with many firms, any change in a single firm's value has a minute effect on the market value $\widetilde{v}^i_m$, and hence the elasticity nears η. This also occurs when

there are few firms, but the firm's value is small relative to other firms in the market. Either case is captured by the within-market choice probability $\rho^i_{j|m}$ going to zero. When the firm is *large* in the worker i's market in the sense of accounting for a lot of the market value and being chosen often, it ends up competing more *across* markets. This is reflected in $\rho^i_{j|m}$ going to one and the first term approaching the across market parameter θ.

The second component captures how the value itself responds to the wage. This varies across workers. The first piece is the marginal value of wealth, λ^i_{jm}. If this is high, then changes in earnings move values by a lot, off-setting any idiosyncratic utility differences across firms and encouraging workers to switch based on wages. *Ceteris paribus*, workers with a high marginal value of wealth will be more elastic across firms as they place a higher value on higher income. The second piece is income. Holding the marginal value of wealth fixed, a high after-tax income—due to a worker having high skill or the firm being a high wage firm—raises the value-at-stake for wage changes. For a high e worker, an increase in the wage raises income by more, shifting the value of working for the firm by more. *Ceteris paribus*, high income raises the stakes for further wage changes, making workers more elastic.

In general, higher earnings reduce this elasticity. Substituting in the first order condition for consumption, and writing earnings as $y^i_{jm} = w^i_{jm} h^i_{jm}$:

$$\varepsilon^{\rho,i}_{jm} = \left[\eta(1-\rho^i_{j|m}) + \theta\rho^i_{j|m}\right] \times \left(y^i_{jm}\right)^{-(\sigma-1)}. \tag{5}$$

Consider $\sigma > 1$. When a worker compares two firms with different earnings, the marginal value of the additional earnings at the high wage firm fall faster than the earnings themselves increase. A 10 percent wage increase at the high wage firm is therefore less likely to attract the worker than a 10 percent wage increase at the low wage firm. This is precisely what we mean by a lower elasticity of labor supply to the high wage firm on the extensive margin. Note how this intersects with standard wealth effects on labor supply, but rather than on hours worked, or the extensive margin of working or not, we are studying *wealth effects across firms*. When wealth effects dominate substitution effects, workers place less value on working at higher wage firms, just like they place less value on working longer hours or working at all.

While we have explained equation (5) in terms of a worker comparing different firms, it also provides a route to identifying σ via across-worker differences. Holding a firm fixed, equation (5) implies that σ determines how differentially wage sensitive different groups of workers will be based on their income y^i_{jm}.

Intensive margin elasticity. The intensive margin labor supply elasticity reflects workers' tendency to work longer hours at firm jm, conditional on having chosen to work there:

$$\varepsilon_{jm}^{h,i} = \frac{\Big(1-\sigma\Big)}{\Big(1+\frac{1}{\varphi}\Big)-\Big(1-\sigma\Big)}$$

This is the usual Marshallian labor supply elasticity. If $\sigma > 1$ wealth effects imply: $\varepsilon_{jm}^{h,i} < \varphi$.

A simple takeaway from our model is that empirical estimates of Frisch and Marshallian labor supply elasticities on hours rule out the intensive margin as determinative of wage setting by firms. Suppose $\sigma = 0$, and consider an upper bound of φ around 1. If firms set wages with only this elasticity in mind, then an *upper bound* on markdowns would be one-half: $\varphi/(1+\varphi)$. This seems unreasonable. On the other end of the spectrum, if firms are perfect substitutes ($\eta \to \infty$), then $\varepsilon_{jm}^{\rho,i}$ is infinite and wages are competitive. Hence, small deviations from competitive wages indicate extensive margin elasticities that are large, but finite, and price labor.

3.2. Sorting

Heterogeneity in extensive labor supply elasticities and complementarity in production yield rich sorting patterns of workers across firms. Here we show how these two forces shape sorting.

First, we establish a metric for sorting. Consider two workers $i \in \{L, H\}$, with $e^H > e^L$. Consider two firms $j \in \{1, 2\}$, in the same labor market, with $z_1 > z_2$. Let m_j^i be some property of worker i at firm j. *Log super-modularity* in m is measured as follows (Shimer and Smith, 2000):

$$\Upsilon_m = \underbrace{\Big(\log m_1^H + \log m_2^L\Big)}_{\text{Pair likes}} - \underbrace{\Big(\log m_2^H + \log m_1^L\Big)}_{\text{Pair dislikes}} = \log\left(\frac{m_1^H/m_2^H}{m_1^L/m_2^L}\right).$$

The first expression is useful for production: $\Upsilon_m > 0$ means production from pairing likes (the diagonal terms) exceeds production from pairing dislikes (the off-diagonal terms). The second is useful for sorting: if m maps to worker values, then $\Upsilon_m > 0$ means that the propensity for the high type worker to go to Firm 1 exceeds that of a low type worker.

A useful application is to productivity. The production function we will use in our quantification of the model will feature a single parameter ψ that controls complementarity. An example of a simple production function with this property is:

$$A(z,e) = ze \times \exp\Big\{ze\Big\} \quad , \quad \Upsilon_A = \psi\Big(z_1 - z_2\Big)\Big(e^H - e^L\Big) > 0.$$

Applied to sorting of workers across firms, we can compute Υ_ρ, the super-modularity asso-

ciated with the extensive margin of labor supply. Using workers' choice probabilities:

$$\Upsilon_\rho = \log\left(\frac{\rho_1^H/\rho_2^H}{\rho_1^L/\rho_2^L}\right) = \log\left(\frac{v_1^H/v_2^H}{v_1^L/v_2^L}\right)^\eta.$$

Suppose that wages are competitive, such that $w_j^i = A_j^i$. Then we can split this into the two forces described above:

$$\frac{\Upsilon_\rho}{\eta} = \underbrace{\Upsilon_A u'\Big(c_2^L\Big)c_2^L}_{\text{Technological complementarity}} + \underbrace{-\Big(\sigma-1\Big)u'\Big(y_1^L\Big)\Big(y_1^L - y_2^L\Big)\Delta\log y_1}_{\text{Wealth effects}} \quad , \quad \Delta\log y_1 = \frac{y_1^H - y_2^H}{y_2^H}. \tag{6}$$

The first piece reflects the complementarity in production, holding fixed any labor supply elasticity differences across workers. If both workers have the same elasticity, and there are increasing differences in firm pay ($\Upsilon_A > 0$), then there will also be increasing differences in firm employment. Hence this term is *positive*. The second piece reflects heterogeneity in extensive margin labor supply elasticities, holding fixed production complementarities. High wage workers have higher consumption, and if $\sigma > 1$, this reduces their marginal value of working at high wage firms. With labor supply less responsive to wages at high wage firms, high e workers will sort toward low wage firms. Hence this term is *negative*. Empirically, positive sorting indicates that the first piece dominates. Hence, conditional on labor supply elasticity differences across worker types, data on sorting is determinative of production complementarities.

3.3. Identification

These results give a particular interpretation to labor supply elasticity differences across workers and firms, and sorting. These are three key features of the data:

Fact 1. Conditional on a firm type, labor supply of high earning workers is less elastic

Fact 2. High productivity workers sort into high productivity firms

Fact 3. Conditional on a worker type, labor supply to high productivity firms is less elastic

The first fact pushes toward wealth effects, and thus $\sigma > 1$. In equation (5), the extensive margin elasticity declines with income only when wealth effects are sufficiently strong. However, if high earners are less wage-sensitive, then equilibrium sorting (6) implies they tend toward low wage firms. Fact two, therefore, points toward complementarities ($\Upsilon_A > 0$) that raise the wages of high earners relatively more at productive firms. This mitigates high earners' higher tendency to ignore wages and follow idiosyncratic tastes, pushing them toward high wage firms and lowering those firms' average elasticity relative to low wage firms. Any remaining difference in elasticities between low and high wage firms must be driven by the *size-based* piece of the extensive margin elasticity. This reveals how much less substitutable markets are than firms from the worker's perspective, and therefore identifies the gap between θ and η.

Appendix C.1 contrasts our approach to Volpe (2025). In Volpe (2025) sorting is entirely driven by low skill workers deriving systematically higher non-pecuniary utility from low wage firms, while labor supply to high productivity firms is *more elastic*. Appendix A.2 describes comparative statics of sorting in the model with respect to ψ and σ.

Summary. Households' labor supply elasticities on the extensive margin across firms are shaped by various forces. These include wealth effects on labor supply, and the desirability of a firm from the workers' perspective which determines labor market power. An implication is that counterfactual policies that shift the distribution of income and wealth feed back into wages via their impact on labor supply elasticities, and hence the wages firms pay. We now turn to calibrating the model.

4. Calibration

4.1. Extending the model

Since heterogeneity in the marginal wealth λ^i_{jm} is a key determinant of labor supply elasticities, we add two empirical features that shape the marginal value of wealth: (i) income taxes and transfers, (ii) non-labor income.

Taxes and transfers. Households receive transfers T, and after tax labor income $\widetilde{y}^i_{jm}$ which is an iso-elastic function of pre-tax labor income y^i_{jm}: $\widetilde{y}^i_{jm} = (1-\tau_0)y^{i\,(1-\tau_1)}_{jm}$. If $\tau_1 = 0$, then τ_0 is the constant proportional tax rate. If $\tau_1 > 0$, then taxes are *progressive*, with marginal tax rates increasing in income. Benabou (2002) and Heathcote et al. (2020a) find this functional form to be a good approximation for the U.S. tax and transfer system. Given the tax system, we compute the budget surplus / deficit G and keep this fixed across counterfactuals by changing τ_0.

Unearned income. We add unearned income a^i, which we measure using the Survey of Consumer Finances. In the SCF we isolate all working age individuals (20 to 55 years old) who earn at least \$1,000 (where all nominal variables are deflated to 2014). We equivalize net worth by the number of adults in the household and then assign unearned income a^i as 4% of net worth, where 4% is the return during our sample window (2000 to 2019) based on Auclert et al. (2021) (Panel A of Figure A.1). We split households into 20 quantiles of income, and compute the share of households in each quantile with a^i in four bins: $(-\infty, 0)$, $[0, \$5,000)$, $[\$5,000, \$20,000)$ and $[\$20,000, \infty)$.[4] This gives the conditional density of $\gamma(a^i|e^i)$, where we set $a^i \in \{\$0,\ \$2,500,\ \$12,500,\ \$20,000\}$. For example, 28 (*2*) percent of the lowest income households are in the lowest (*highest*) non-labor income bin, while 2 (*58*) percent of the highest income households are in the lowest (*highest*) bin.

Key formulas. The budget constraint for an individual is now $c^i_{jm} = (1-\tau_0)y^{i\,(1-\tau_1)}_{jm} + T + a^i$. Let $\Gamma(a, e)$ denote the joint distribution over (a, e). We assume that firms can set wages $w_{jm}(e)$

[4] These bins represent 12%, 62%, 19% and 7% of households, respectively.

that depend on e, but not a. However, firms understand the distribution $\Gamma(a, e)$ and the labor supply of different types. Hence the firm's elasticity of labor supply for a type of worker e is a weighted average of (a, e) types, conditional on e:

$$\begin{aligned} n_{jm}(e) &= \int \rho_{jm}(a,e) h_{jm}(a,e) d\Gamma(a|e) \\ \varepsilon_{jm}(e) &= \int \frac{\rho_{jm}(a,e) h_{jm}(a,e)}{\int \rho_{jm}(a',e) h_{jm}(a',e) d\Gamma(a'|e)} \Big[\varepsilon^{\rho}_{jm}(a,e) + \varepsilon^{h}_{jm}(a,e)\Big] \, d\Gamma(a|e). \end{aligned}$$

Adopting similar notation as before, we summarize worker heterogeneity by superscript i, i.e. $\rho^i_{jm} = \rho_{jm}(a, e)$. The extensive and intensive margin elasticities of workers are now:

$$\begin{aligned} \varepsilon^{\rho,i}_{jm} &= \Big[\eta(1-\rho^i_{j|m}) + \theta\rho^i_{j|m}\Big] u'\Big(\widetilde{y}^i_{jm} + a^i + T\Big)\widetilde{y}^i_{jm}\Big(1-\tau_1\Big) \quad , \\ \varepsilon^{h,i}_{jm} &= \frac{\Big(1-\sigma MPC^i_{jm}\Big)\Big(1-\tau_1\Big)}{\Big(1+\frac{1}{\varphi}\Big) - \Big(1-\sigma MPC^i_{jm}\Big)\Big(1-\tau_1\Big)} \quad , \end{aligned} \tag{7}$$

where MPC^i_{jm} is the elasticity of consumption with respect to after tax income. In our static economy, this is simply $\widetilde{y}^i_{jm}/(\widetilde{y}^i_{jm} + a^i + T)$.

These equations deliver three observations. First, higher tax progressivity reduces both elasticities. The firms' labor supply elasticity is with respect to the *pre-tax wage*. However, on the margin, a \$1 wage increase from raises workers' after-tax income by progressively less when τ_1 is higher. Hence, *ceteris paribus*, higher tax progressivity mutes the impact of pre-tax wages on workers' decisions to choose a higher paying employer or supply more hours. Second, additional non-labor income shifts elasticities in different directions. Extensive margin elasticities decreases as differences in the marginal value of wages across firms compress, requiring larger wage changes to shift labor supply. Intensive margin elasticities decrease as consumption becomes less elastic to labor income, dampening wealth effects that otherwise lower the Marshallian labor supply elasticity. Third, firms incorporate these effects into the prices of labor. If most high e households have high a, then $w_{jm}(e)$ will be set low, reflecting less elastic supply. However, the wage will be higher than if the firm could set $w_{jm}(e, a)$ as a function of assets. High a households benefit from firms lack of ability to discriminate by a, while low a households lose.

4.2. Approach

All parameters of the model are estimated via Simulated Method of Moments (SMM) except for the initial wealth distribution and institutional parameters, including the tax and transfer function, which are set outside of the SMM routine. We assume that all markets in the economy are populated by an identical number of firms J, with the same levels of productivity $z_1, \ldots, z_J$.

Average markdown and size-based market power. Parameters governing markdowns are $\{\varphi, \sigma, \eta, \theta, J\}$. We estimate the Frisch elasticity φ to target the average marginal propensity to earn (MPE) from annuitized lottery winnings in the U.S. (Golosov et al., 2024). Golosov et al. (2024) define the marginal propensity to earn as the dollar change in labor earnings for every dollar change in annuitized lottery winnings. The average lottery winner has an increase in non-labor income of approximately \$7,500 dollars per annum (Panel (a) of Figure IV). The average MPE is -0.52 (Column (1) of Table III). To implement this experiment in our model we increase a^i by \$7,500 and compute changes in labor earnings, holding equilibrium wages fixed.

Remaining preference parameters $\{\sigma, \eta, \theta\}$ govern the relative labor supply elasticities of high versus low income households, and the relative markdowns of small and large firms, respectively. Estimates of how wages and employment change at different firms and for different groups of workers following shocks to labor demand are available in Brazil (Bils et al., 2025), Norway (Volpe, 2025), and the US (Seegmiller, 2025). In all cases firms' *reduced form labor supply elasticities*—computed as the change in log employment divided by the change in log average wage following identified shocks to labor demand—are higher for low income workers. This is true across all firm types. In Brazil and the U.S., reduced form labor supply elasticities are also higher at firms with low average wages across all worker types. In Norway, Volpe (2025) finds that higher wage firms have higher reduced form labor supply elasticities.

We estimate our parameters to match the joint profile of reduced form labor supply elasticities across worker skills and firm value-added for the U.S. Seegmiller (2025) uses variation in firm-specific stock returns as instruments for the marginal revenue product of labor. Workers are split into quintiles of skill based on AKM (Abowd et al., 1999) fixed effects, with the bottom two and top two labeled low- and high-skilled. The estimated average labor supply elasticity is 5.34. The gap in labor supply elasticities between high- and low-skilled workers is 5.69. The gap in labor supply elasticities between low- and high-value-added firms is 1.69. To compute the model counterpart to these reduced form elasticities, we (i) simulate productivity shocks to individual firms to deliver wage increases consistent with the data, (ii) recompute the equilibrium in the market, (iii) compute the change in employment and wages at treated firms relative to non-treated firms.[5]

The number of firms per market J is chosen to match the payroll-weighted, wage-bill Herfindahl.[6] Since all markets are identical, national payroll-weighted wage-bill Herfindahl is equal to its market-level counterpart. We estimate $J = 16$ which yields $HHI = 0.11$ (Berger et al., 2022).

Technology and complementarity. We build on a long tradition of labor assignment models and assume that output is decreasing in the distance between worker skills and firm produc-

[5]This procedure follows Berger et al. (2026) which points out the issues associated with mapping reduced form labor supply elasticities to structural objects like ε_{jm} in concentrated markets. For full details see Appendix C.2.

[6]Let $\overline{w}_{jm}$ denote the average wage per worker of firm j in market m. Let s_{jm} denote their wage-bill share $s_{jm} = \frac{\overline{w}_{jm} n_{jm}}{\sum_{j' \in m} \overline{w}_{j'm} n_{j'm}}$. The Herfindahl is computed as $HHI_m = \sum_j s_{jm}^2$.

tivity (e.g., Marimon and Zilibotti (1999) and Lise and Postel-Vinay (2020)). We specify the following functional form for $A(z_j, e^i)$. The functional form has three benefits:

$$A\Big(z_j, e^i\Big) = z_j e^i \exp\left\{-\frac{\psi}{2}\left(\widetilde{z}_j - \widetilde{e}^i\right)^2\right\} \quad , \quad \widetilde{z}_j = \frac{\log z_j - \mathbb{E}[\log z_j]}{\mathbb{V}[\log z_j]} \quad , \quad \widetilde{e}^i = \frac{\log e^i - \mathbb{E}[\log e^i]}{\mathbb{V}[\log e^i]}.$$

First, a single parameter controls complementarity, ψ. A higher value of ψ increases complementarity by penalizing matches of workers and firms that are *off-diagonal*. Second, if we apply the log-supermodularity measure of sorting from above, then:

$$\Upsilon_A = \psi\Big(\widetilde{z}_1 - \widetilde{z}_2\Big)\Big(\widetilde{e}^H - \widetilde{e}^L\Big).$$

By normalizing z_j and e^i we ensure that sorting induced via A depends only on ψ and does not mechanically vary as variances of worker or firm productivities vary across potential sets of parameters. Third, we nest the case of 'pure AKM' or 'linear-additive wages', which obtain under $\psi = 0$ and markdowns μ that are independent of worker or firm type: $\log w_{jm}(e^i) = \log \mu + \log z_j + \log e^i$.

Heterogeneity. Worker productivity e is log normally distributed, with standard deviation σ_e and mean normalized to one. We discretize the distribution using mid points of 20 quantiles, which we combine with the distribution of $\Gamma(a|e)$ from the SCF to obtain the joint distribution $\Gamma(a, e)$. Firm productivity z is drawn from a Pareto distribution, with standard deviation σ_z (which implies the tail parameter), and a mean normalized to one.[7] Conditional on all other parameters, σ_e increases dispersion in consumption. Given the central importance of the marginal utility of consumption in the elasticity of labor supply, we make sure that this lines up with U.S. data: $var(\log c) = 0.27$ as computed by Heathcote et al. (2020b) in CEX data. To calibrate $\{\sigma_z, \psi\}$ we match the covariance structure of worker and firm fixed effects reported in the literature. We use two moments, the share of the variance of worker wages attributable to (a) firm fixed effects, and (b) their covariance. We take these from Kline et al. (2020), and are 0.13 and 0.16, respectively. To compute these in the model, we simulate a 100 period panel of 100,000 workers which we hand over to the code from Kline et al. (2020) for the estimation of a standard Abowd et al. (1999) decomposition.[8] Note that the largest share of the variance is due to worker fixed effects (0.61), while the remaining 0.10 is error.[9] These shares are similar to those estimated by Bonhomme et al. (2019) in Sweden (77% worker, 5% firm, 17% covariance).

[7] Given a number of firms J, we compute productivities $z_1, \ldots, z_J$ by evaluating the inverse CDF of the Pareto at J evenly spaced points between of 0.05 and 0.95.

[8] Kline et al. (2020) apply methods to adjust for the uncertainty over worker and firm effects induced by limited mobility bias (i.e., estimation uncertainty over types due to limited sample size).

[9] In the data, the error term is due to time varying idiosyncratic factors, measurement error and departures from a linearly additive wage structure. In the model, the first two are absent, but there are departures from linear additivity. We find that these departures from linear additivity account for around half of the variance of the error in the data.

Parameter		Value	Moment	Data	Model
Government transfer	T	\$2,251	Ferriere et al. (2023)		
Tax function - Progressivity	τ_1	0.06	Ferriere et al. (2023)		
Tax function - Shifter	τ_0	-0.35	Ratio: Income tax / Payroll	0.29	0.30
Productivity shifter	$\overline{Z}$	44.56	Average earnings	\$45,207	\$45,459
Labor disutility shifter	$\overline{\varphi}$	15,351	Average hours	1,777	1,785
Worker prod. dispersion	σ_e	0.73	Variance of log consumption	0.27	0.26
Firm prod. dispersion	σ_z	0.50	Share of log wage variance - Firm effects	0.13	0.12
Complementarity coefficient	ψ	0.09	Share of log wage variance - Covariance	0.16	0.16
			Share of log wage variance - Error$^+$	0.10	0.05
			Share of log wage variance - Worker effects$^+$	0.61	0.67
Within-market dispersion	η	2.35×10^7	Average elasticity*	5.34	4.73
Across-market dispersion	θ	$0.15 \times \eta$	Elasticity gap: Low - High wage firms*	1.69	1.71
Relative risk aversion	σ	2.48	Elasticity gap: Low - High wage workers*	5.69	5.90
Frisch elasticity	φ	0.09	MPE after \$7,500 windfall gain	-0.52	-0.52
Number of firms	J	16	HHI - Payroll	0.11	0.12

Table 1: Baseline calibration

Note: (i) $*$ denotes from (Seegmiller, 2025), (ii) $+$ denotes not targeted, (iii) firm productivity is Pareto, however we report the standard deviation of log productivity instead of the tail parameter, which is 1.58.

Earnings and hours. Two parameters control scale in the economy: hours disutility ($\overline{\varphi}$), and a multiplicative constant that we add to firm productivity ($\overline{Z}$). We choose $\overline{Z}$ to match average earnings of \$45,207 (in 2014 dollars), obtained from the SCF. We choose $\overline{\varphi}$ to match average annual worker hours of 1,777, obtained from the OECD (see Berger et al., 2025).

Taxes and transfers. We use the transfer-augmented Heathcote et al. (2017) tax function from Ferriere et al. (2023). Ferriere et al. (2023) estimate $\tau_1 = 0.06$, and $T = \$2,251$, this implies a ratio of transfers to income of 4.98%.[10] We set τ_0 to match the ratio of labor income taxes to total payroll of 29 percent.[11]

4.3. Results - Estimation

We highlight five lessons from the calibration itself which we connect to our discussion of identification (Section 3.3). Table 1 provides parameter estimates and moments in model and data.[12]

First, only moderate wealth effects are required to rationalize the gap between reduced form estimates of labor supply elasticities to firms of low versus high wage workers ($\sigma = 2.48$).[13] This

[10] We directly use the Ferriere et al. (2023) estimates of progressivity from their Table B.3. However, they use a household level definition of transfers. We thus rescale T so that transfers as a ratio of household income (their $\$4,492/\$90,185 = 4.98\%$) is similar in our model economy ($\$2,251/\$50,436 = 4.46\%$).

[11] Labor income taxes are 19% of U.S. GDP. To get their share of payroll we divide by 0.65 labor share.

[12] The large value of η is not comparable to existing estimates because it is scaled by marginal utility. Returning to the case of $\widetilde{y}^i_{jm} = y^i_{jm} = c^i_{jm}$, the extensive margin elasticity is: $\varepsilon^{\rho,i}_{jm} = [\eta(1-\rho^i_{j|m}) + \theta\rho^i_{j|m}] \times (y^i_{jm})^{-(\sigma-1)}$. Average per worker earnings are \$45,207. With $\sigma = 2.48$ the second term is of order 10^{-6}. To deliver markdowns consistent with the data, the first term, $[\eta\rho^i_{j|m} + \theta(1-\rho^i_{j|m})]$, must be of a similar inverse magnitude.

[13] Appendix Table A1 provides the reduced form labor supply elasticity estimates, rather than the gap. While

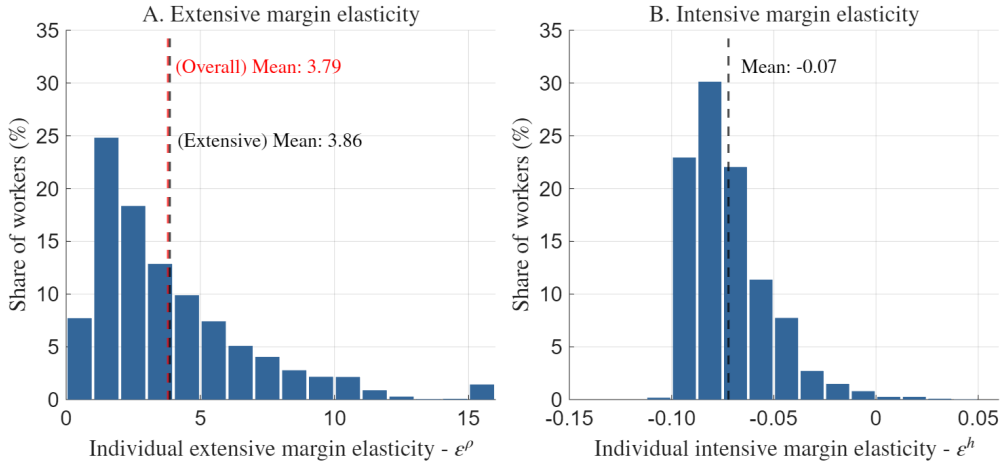


Figure 1: Distribution of workers' extensive and intensive margin labor supply elasticities

falls within ranges of standard estimates.

Second, there are notable complementarities ($\psi = 0.09$). With $\sigma > 1$, high wage workers have a lower marginal value of working at a high wage firm, a force toward negative sorting. Matching the observed covariance of firm and worker fixed effects requires positive sorting. This is achieved via complementarities in production.[14]

Third, it is not the case that high wage firms are productive simply because they are a collections of high productivity workers ($\sigma_z = 0.50$). Despite a small share of variation in wages being accounted for by firm fixed effects, heterogeneity in firm productivity is nonetheless large and in the range of estimates from administrative firm level data (Decker et al., 2013). We share the conclusion of Borovičková and Shimer (2024) that firm-wage-variance decompositions can not be used to make prima-facie conclusions about heterogeneity in worker and firm productivity.

Fourth, significant size-based market power is required ($\theta = 0.15 \times \eta$). Since $\sigma > 1$, an individual worker is less responsive to wage increases at high wage firms, this delivers a low elasticity at high wage firms. Despite this, a value of $\theta < \eta$ is still required to line up the labor supply elasticities of high wage firms with their counterparts in the data.

More broadly, these provide our first main quantitative result. A neoclassical model of labor supply, extended to extensive margin decisions over which firm to work at, complementarities in production, and oligopsonistic wage setting, can neatly rationalize headline statistics obtained from important empirical studies of firm and worker heterogeneity in labor economics. This is achieved under a set of relatively standard parameter values and a transparent identification argument. We view this as a strong outcome for the theory.

we targeted the gaps across worker and firm types, the model is also consistent with a larger decline across firm types for low wage workers, relative to high wage workers.

[14]To give a sense of magnitudes, we can compute the supermodularity measure used in Section 3 for a given pair of workers and firms. For workers and firms at the 25th and 75th percentiles of e and z respectively, the high skill worker is nearly twice as productive at the high z firm than the low skill worker is: $A(e^H, z^H)/A(e^H, z^L) = 1.7 \times A(e^L, z^H)/A(e^L, z^L)$. For workers and firms at the 10th and 90th percentiles, this factor increases to more than nine.

	A. MPE			B. Δ Pr(move down | any move)		
Earnings quartile	All	Low	High	All	Low	High
Data	-0.52	-0.31	-0.67	7.0	6.0	3.2
Model	-0.52	-0.26	-0.78	9.1	8.7	6.8

Table 2: Heterogeneity in responses to unearned income

Note: Model moments computed in response to $7,500 increase in unearned income in partial equilibrium. Data moments taken from Golosov et al. (2024). See details on computation of conditional downward job switching rate in Appendix C.3.

Finally, from the calibrated model we can measure structural elasticities of labor supply of each worker on the extensive and intensive margin, which, once employment weighted yield an average elasticity of 3.79. Figure 1 shows that this is entirely shaped by the extensive margin elasticity. Intensive margin elasticities sit tightly within the range of Marshallian labor supply elasticities used by the Congressional Budget Office, from -0.10 to 0.20. Extensive margin elasticities are large and vary substantially across workers. Fundamentally this paper is about shifting attention in the canonical model of labor from an inelastic piece of labor supply, to a highly elastic piece.

4.4. Additional statistics

Extensive margin elasticities. Table 2 shows that the model generates responses to unearned income that line up with recent estimates from lottery winnings in the U.S. (Golosov et al., 2024). While the average MPE was targeted, Panel A shows that the model generates the right gradient of MPEs with respect to income. As discussed above, intensive margin elasticities are small and in line with the data due to the low Frisch elasticity, so the level and gradient of MPEs are delivered by workers switching employers.

Panel B provides direct evidence on the extensive margin. Golosov et al. (2024) compute the probability of downward and upward job moves among their sample of lottery winners. Since our model is static and has no notion of adjustment costs of switching employers, we compute the probability of a downward move to a firm with a lower average wage conditional on any move (down/(down+up)).[15] The model matches the average impact of unearned income on the direction of job switches, and qualitatively matches the gradient with respect to income. Higher income workers are less likely to move down the ladder in response to windfalls since they largely follow their tastes.

Panels A and B are simple to reconcile. In the model, high wage workers are less likely to move to lower wage firms in response to unearned income, but, due to complementarity and greater dispersion of firm productivity in the right tail, they experience larger reductions in earnings when they do move. This is verified in Figure 2A. MPEs are larger for the rich, despite smaller increases in the probability of moving to a lower wage firm.

[15] Appendix C.3 provides details on the computation of conditional job-to-job moves.

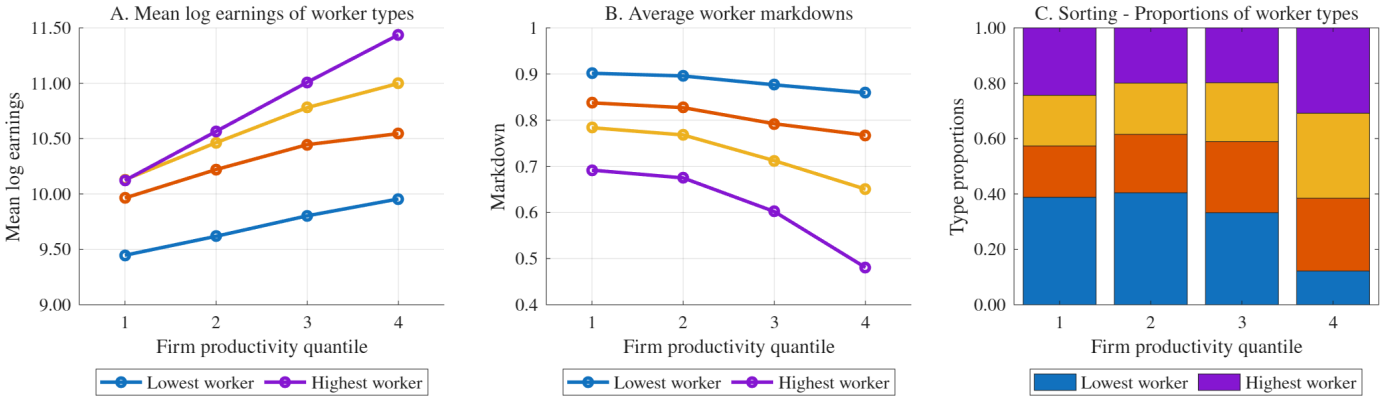


Figure 2: Wages, Markdowns and Sorting

Wages by types and sorting. An important question is whether the presence of complementarities causes wages to depart from log-additivity in a way that is inconsistent with recent empirical findings. One might think that with $\psi > 0$, the wages a worker might earn when comparing firms would be 'hump-shaped', peaking at their closest match. Figure 2A shows that this is not the case, matching the main qualitative features of Figure 2 in Bonhomme et al. (2019). Importantly, for all workers, earnings are higher at higher productivity firms. Recall that $A(z, e) = ze \times \text{Penalty}_{\psi}(z, e)$, such that if the Penalty term is 1, we maintain log additivity in the marginal product of labor. Figure 2A shows that the ze term dominates in shaping wages. Meanwhile, the *Penalty* term exerts sufficient complementarities to undo the negative sorting via wealth effects and deliver positive sorting in equilibrium. Figure 2C shows that the Penalty function delivers the empirical property that relatively few low skill workers work at high z firms. We conclude that these features of the data can be explained with mechanisms in addition to selection in the presence of search frictions (Borovičková and Shimer, 2024) or sorting on amenities (Lamadon et al., 2025; Volpe, 2025).

Summary. A lightly parameterized extension of neoclassical theories of labor supply can knit together a set of important facts in the recent empirical literature studying worker and firm heterogeneity in labor markets. We now proceed in two steps. Section 5 describes the quantitative implications of the model for wage setting and markdowns. This is a key step in understanding the implication of the model for tax progressivity in Section 6.

5. Wage setting behavior

Before turning to optimal taxation, we first quantify the model's implications for wage setting and markdowns. We begin by describing the sources of aggregate markdowns, isolating variable and fixed components. We then disaggregate markdowns across firms and workers. These forces determine optimal tax policy.

5.1. Aggregate sources of markdowns

We decompose the *aggregate markdown* μ, which we define as the earnings weighted average of worker-specific markdowns $\mu_j(a, e)$. For clarity, we drop the additional m subscript, and use i to capture heterogeneity in a^i and e^i. In the calibrated model, μ is 0.64.

$$\mu = \sum_{i,j} \omega_j^i \mu_j^i \quad , \quad \omega_j^i = \frac{y_j^i \rho_j^i \gamma(a^i, e^i)}{\sum_{i',j'} y_{j'}^{i'} \rho_{j'}^{i'} \gamma(a^{i'}, e^{i'})} \quad , \quad \mu_j^i = \frac{\varepsilon_j^{\rho,i} + \varepsilon_j^{h,i}}{1 + \varepsilon_j^{\rho,i} + \varepsilon_j^{h,i}} \tag{8}$$

From the key equations (7) for $\varepsilon_j^{\rho,i}$ and $\varepsilon_j^{h,i}$, we isolate four sources of markdowns: (1) tax progressivity, (2) oligopsony, (3) heterogeneity, and (4) standard monopsony. Each maps into components of the extensive and intensive elasticities:

$$\varepsilon_j^{\rho,i} = \underbrace{\left(1 - \tau_1\right)}_{\text{(1) Tax progressivity}} \underbrace{\left[1 - \rho_j^i \left(\frac{\eta - \theta}{\eta}\right)\right]}_{\text{(2) Oligopsony}} \underbrace{u'\left(c_j^i\right)\widetilde{y}_j^i}_{\text{(3) Heterogeneity}} \underbrace{\eta}_{\text{(4) Monopsony}} \tag{9}$$

$$\varepsilon_j^{h,i} = \frac{\overbrace{\left(1 - \tau_1\right)}^{\text{(1) Tax progressivity}} \overbrace{\left(1 - \sigma MPC_j^i\right)}^{\text{(3) Heterogeneity}}}{\left(1 + \frac{1}{\varphi}\right) - \underbrace{\left(1 - \tau_1\right)}_{\text{(1) Tax progressivity}} \underbrace{\left(1 - \sigma MPC_j^i\right)}_{\text{(3) Heterogeneity}}}. \tag{10}$$

To isolate their contribution to markdowns, we remove sources one by one. We leave the earnings weights fixed when recomputing the aggregate markdown in equation (8). Table 3 reports our results, starting from an aggregate markdown $(1 - \mu)$ of 0.36.

First, we isolate the role of tax progressivity, which enters both elasticities through $(1 - \tau_1)$. A progressive schedule ($\tau_1 > 0$) shrinks the return to earning more, lowering extensive- and intensive-margin elasticities and raising markdowns. Under flat labor taxes ($\tau_1 = 0$) the aggregate markdown falls from 0.35 to 0.34 (column 1), which accounts for 3.6 percent of the aggregate markdown (column 2).

Second, we isolate the role of oligopsony by computing markdowns under atomistic firm behavior. Under $\theta = \eta$, there is no strategic component of markdowns, and the aggregate markdown falls from 0.34 to 0.32, accounting for 6.7 percent of the markdown.

Third, we isolate the role of worker heterogeneity. We do so in two stages. With $\sigma > 1$, a worker with a fixed (a^i, e^i) will be less responsive to an additional \$10,000 of earnings at a \$100,000 job, than to an additional \$5,000 of earnings at a \$50,000 job. To remove this we replace $u'(c_j^i)\widetilde{y}_j^i$ and MPC_j^i with their average value *Within-worker, Across-firms*, weighted by ρ_j^i. The aggregate markdown falls from 0.32 to 0.30, a drop in magnitude comparable to oligopsony power. Removing the *Within-worker, Across-firm* effect still leaves differences in average elasticities *Across-workers*. A worker that earns \$50,000 on average across jobs will be more elas-

Source of markdowns removed	Markdown $(1-\mu)$ (1)	Share (%) (2)	Cumulative share (%) (3)
Baseline	0.35	—	—
1. No tax progressivity - $\tau_1 = 0$	0.34	3.6	3.6
2. No oligopsony power - $\theta = \eta$	0.32	6.7	10.3
3. No within-worker across-firm elasticity gaps - $u'(c_j^i)\widetilde{y}_j^i = \mathbb{E}\left[u'(c_k^i)\widetilde{y}_k^i \middle| i\right]$	0.30	6.0	16.3
4. No across-worker elasticity gaps - $u'(c_j^i)\widetilde{y}_j^i = \mathbb{E}\left[u'(c_k^i)\widetilde{y}_k^i\right]$	0.18	31.5	47.8
5. No monopsony $\eta = \infty$	0.00	52.2	100

Table 3: Decomposition of average markdowns

Note: This table presents a cumulative decomposition of the payroll-weighted markdown. Each row represents the marginal effect of removing the specified factor while maintaining the removal of all preceding factors. Column (1) reports the value of the markdown $(1-\mu)$. Column (2) displays the change in the markdown value between the current and previous row, expressed as a share of the baseline markdown. Column (3) shows the cumulative change in the markdown, expressed as a share of the baseline.

tic than a worker that earns on average \$100,000 across jobs. To remove this we replace $u'(c_j^i)\widetilde{y}_j^i$ and MPC_j^i with their average *Across workers*. The aggregate markdown sharply falls from 0.30 to 0.18, accounting for 31.5 percent of the markdown. Combined, these variable markdown terms combined account for nearly 50 percent of the markdown (column 3). The final half is accounted for by monopsony, removed by eliminating dispersion in idiosyncratic tastes $(\lim_{\eta\to\infty})$.

Summary. Nearly half of markdowns are "variable", driven by sources of both worker and firm heterogeneity. Of the piece that is variable, 14 percent is attributable to oligopsony, while the majority is attributable to worker heterogeneity. That markdowns strongly depend on individuals' heterogeneous wage elasticities implies that changes in marginal tax rates will heterogeneously impact wages and allocations of workers across firms.

5.2. Across-firm sources of markdowns

To unpack sources of average markdowns across firms into the same oligopsony and worker heterogeneity channels we split equation (7) for $\varepsilon_{jm}^{\rho,i}$ as follows:

$$\varepsilon_j = \sum_i \omega_j^i \underbrace{\left[\eta\left(1-\rho_j^i\right)+\theta\rho_j^i\right]}_{\text{A. Oligopsony}} \underbrace{\frac{u'\left(c_j^i\right)\widetilde{y}_j^i}{\mathbb{E}_{\rho^i}\left[u'\left(c_k^i\right)\widetilde{y}_k^i \middle| i\right]}}_{\text{B. Within-worker, Across-firm}} \underbrace{\mathbb{E}_{\rho^i}\left[u'\left(c_k^i\right)\widetilde{y}_k^i \middle| i\right]}_{\text{C. Across-worker (composition)}}, \quad \omega_j^i = \frac{\rho_j^i h_j^i \gamma(a^i, e^i)}{\sum_l \rho_j^l h_j^l \gamma(a^l, e^l)} \tag{11}$$

We rank firms by their employment, and then plot this decomposition in log deviations from the smallest firm (Figure 3).[16]

Both the oligopsony and consumption heterogeneity components contribute to declining

[16] In practice we implement the following approximate decomposition $\widetilde{\varepsilon}_j := \varepsilon_j^A \varepsilon_j^B \varepsilon_j^C$ where each component is the hours weighted average of the relevant terms in (11)

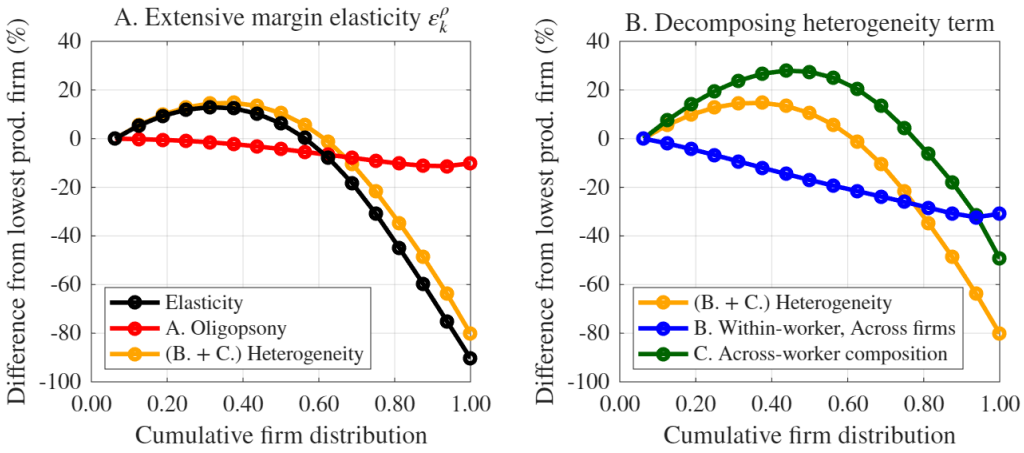


Figure 3: Decomposition of labor supply elasticities across firms

elasticities of labor supply by firm size. From the smallest to the largest firm, the combined heterogeneity terms account for 92 percent of the decline. The *Within-worker, Across-firm* effect accounts for 34 percent of the decline. A single worker is less elastic at higher paying firms due to reduced wage sensitivity at higher pay. This channel imparts larger wage setting power to high wage firms than their oligopsonistic presence in the labor market (8 percent). The *Across-worker composition effect* accounts for 58 percent of the decline. Most of the differences between large and small firms in terms of their ability to mark down wages comes from the composition of employment: sorting delivers high productivity firms disproportionately more high skill, high asset, less elastic workers.

Summary. Wide markdowns at large firms are shaped primarily by the worker heterogeneity channel. The oligopsony channel amplifies this, but plays a smaller role. These wide markdowns at large firms are endogenous and respond both directly and indirectly to tax policy through changes in the marginal value of wealth and sorting.

5.3. Across-worker sources of markdowns

Lastly, we measure the effects of markdowns on the distribution of worker wages. Table 4 compares the baseline economy (column 1) to a competitive counterfactual (column 2) in which we impose $\mu_j(a, e) = 1$ so that every worker is paid their marginal revenue product. The counterfactual is solved in general equilibrium, so workers may reallocate across firms.

Labor market power substantially compresses inequality in wages, hours and consumption. Competition pushes up wages the most at the highest productivity firms and for the highest skill workers whose wages were marked down due to a low marginal value of additional pay (recall Figure 2B). Hence, in partial equilibrium, moving to competitive wages increases wages at the top of the distribution. In general equilibrium high-e workers respond to these higher wages, reallocating towards high-z firms. Sorting rises by more than 40 percent, further expanding wages in the tail. The standard deviation of log earnings rises from 0.60 to 0.77 and

	Baseline (1)	Competitive (2)
Inequality [standard deviation of log]		
Hourly wage	0.60	0.77
Earnings	0.60	0.77
Consumption	0.51	0.65
Variance of log wages		
Overall	0.36	0.59
Across firms	0.13	0.20
Within firms	0.23	0.40
Other measures		
Covariance of firm and worker fixed effects (Sorting)	0.029	0.042

Table 4: Market structure and worker wage dispersion

Note: This table compares the effects of firms' wage setting structure on worker inequality. Column (1) is the baseline. In Column (2), firms set a competitive wage equal to marginal product. Each specification is evaluated in general equilibrium without recalibration.

log consumption from 0.51 to 0.65. The model is consistent with the empirical within-between firm log wage variance decompositions that attribute two-thirds to within and one-third between. Under competitive wage setting both increase.

Summary. Eliminating markdowns, the competitive allocation increases wages for high skill workers at high productivity firms, increasing sorting and productivity. With high efficiency comes higher inequality. Raising tax progressivity will do the opposite, reducing inequality, but also reducing sorting which moves the economy further from efficiency.

6. Positive and normative implications for tax progressivity

Variation in extensive margin labor supply elasticities are key to understanding wage setting and sorting patterns. Our key equations in (7) also showed how higher progressivity directly reduces extensive and intensive margin elasticities of labor supply to the firm. This raises the question, what are the implications of our model for the optimal shape of income taxes?

This section has four sets of results. (1) Higher tax progressivity has novel, negative, effects on markdowns and productivity. (2) Empirically, these effects generate responses of earnings to increases in marginal tax rates that are consistent with leading studies (Saez et al., 2012). (3) Analytically, these effects reduce optimal marginal tax rates, and (4) Quantitatively, these reductions are large relative to an economy with homogenous, competitive firms.

6.1. Positive implications of higher tax progressivity

To understand effects on inequality, markups, sorting, wages, output and productivity, we vary τ_1 between 0 (flat taxes) and 0.20. We change τ_0 to keep the average tax rate constant at 0.29 (Table 1). We plot results by marginal tax rate at P95 income, which increases from 29 percent

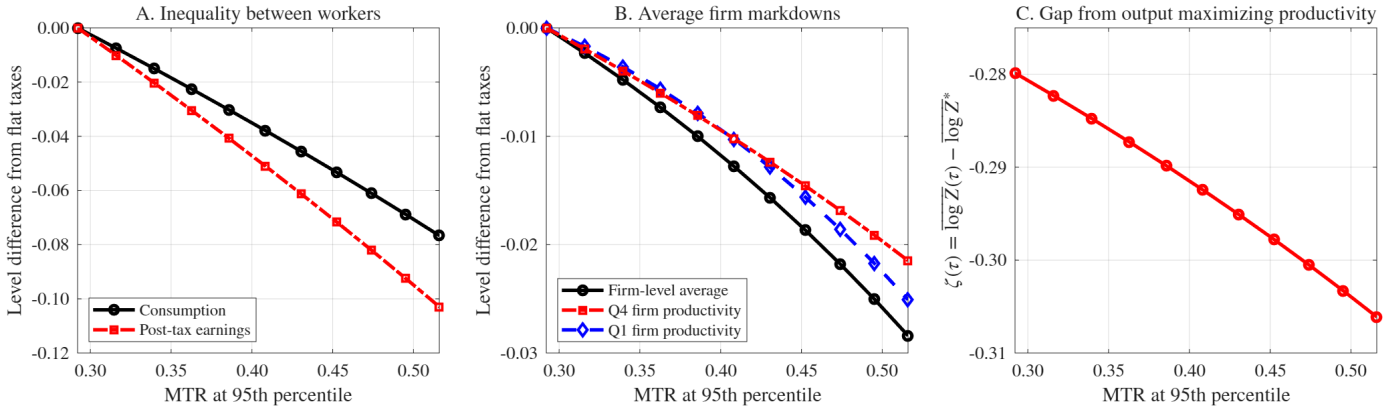


Figure 4: Tax progressivity in general equilibrium - Inequality, markdowns and sorting

to 52 percent as τ_1 increases.[17]

Inequality. Higher progressivity has the expected effects on inequality. The standard deviation of after tax earnings is reduced by nearly 11 percentage points—around one quarter of its level—as τ_1 increases from 0 to 0.20 (Figure 4A). Consumption inequality falls by slightly less due to the presence of transfers and non-labor income.

Markdowns. Our framework links firm wage setting to tax policy via the interaction of tax policy and workers' elasticities of labor supply. Higher progressivity reduces markdowns as workers become less elastic with respect to pre-tax wages (Figure 4B). On average, markdowns widen by nearly 3 percentage points. Effects are larger at high productivity firms due to non-linearities in the interaction between progressivity and labor supply elasticities.[18]

Sorting. As labor becomes less elastic, workers respond less to wages at high wage firms, and sort more toward lower productivity firms, reducing average productivity (Figure 4C). Let Z^* be average log labor productivity across workers under an allocation where all workers of skill e work at the firm $z^*(e)$ that maximizes their labor productivity.[19] With constant marginal tax rates, average productivity is 28.0 log points less than the output maximizing allocation. Some workers have idiosyncratic preferences for low productivity firms, hence productivity is distorted downwards. The distortion expands to 30.5 log points as progressivity increases. Other measures tell the same story. The share of workers at their productivity maximizing firm falls from 16 percent to 14 percent, and the share of workers at firms with above median value-added per worker decreases from 82 percent to 78 percent. These declines are larger for low

[17] In the model's baseline under $\tau_1 = 0.06$ the marginal tax rate at the P90 of income is 36.5 percent. In the U.S. the marginal tax rate at the P90 is around 39 percent. In 2025, the Federal marginal rate is 32% between the P94 ($213,050) and P97 ($266,725). Adding 2.35% Medicare plus a middle-of-the-range state-tax of 5% delivers 39.35%.

[18] An increase in τ_1 delivers a larger reduction in markdowns when $\varepsilon_{jm}^{i,\rho}$ is smaller. Suppose $\mu = (1-\tau_1)\varepsilon/(1+(1-\tau_1)\varepsilon)$. If $(1-\tau_1)\varepsilon > 1$, so $\mu > 0.5$, then $d\mu/d\tau$ is negative in sign and larger in magnitude for smaller ε.

[19] Average log labor productivity is: $\overline{Z}(\tau) = \sum_{i,j} \rho_j^i \gamma^i \log A(z_j, e^i)$. The measure Z^* is $Z^* = \sum_i \gamma^i \log A(z_j^*(e^i), e^i)$, where $z_j^*(e^i) = \arg\max_{z_j} A(z_j, e^i)$.

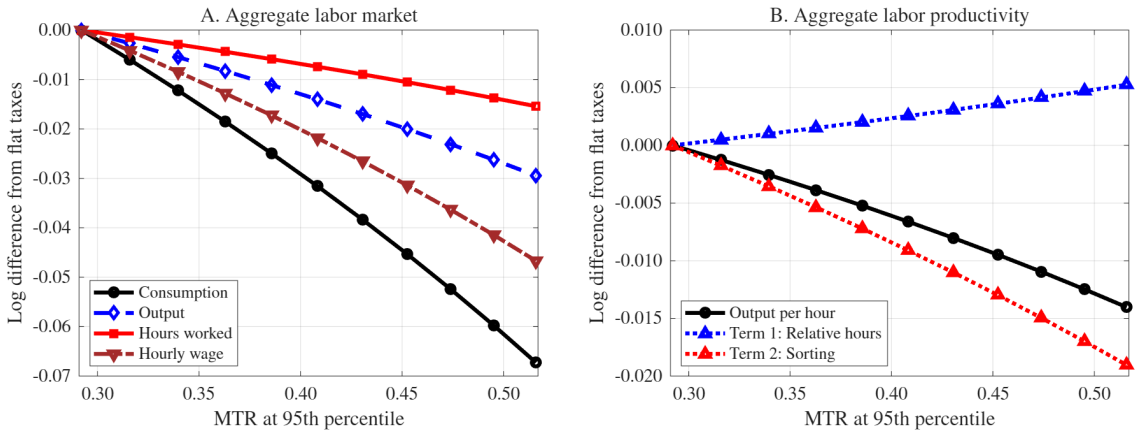


Figure 5: Tax progressivity in general equilibrium - Aggregates and productivity

skill workers, who become less elastic due to higher marginal taxes, but also richer and hence less elastic due to increased redistribution via the level of taxes.[20] Overall, lower sorting reduces aggregate productivity, which we turn to next, presenting a sharp trade-off between inequality and efficiency.

Aggregate wages, hours, output and productivity. Average hourly wages fall by more than 4.8 percent, while total hours decline by 1.5 percent (Figure 5A). Declining hours account for less than half of the 3 percent decline in output, with the remainder accounted for by a decline in aggregate labor productivity. In a standard theory of labor supply, an increase in tax progressivity leads labor supply to contract: wages increase, while equilibrium employment, output and consumption decline. In our model, higher progressivity widens markdowns, effectively contracting labor demand: wages decrease, while equilibrium employment, output and consumption decline. Are these numbers large or small? In quantitative models of optimal labor income taxes, the welfare benefits of optimizing taxes are often in the range of 1 to 3 percent of consumption. Hence, these losses have the potential to offset such gains.

The decline in aggregate productivity is due to the novel mechanisms in our model. Aggregate output and hours are:

$$Y = \sum_{i,j} \rho_j^i h_j^i A_j^i \gamma(a^i, e^i) \quad , \quad H = \sum_{i,j} \rho_j^i h_j^i \gamma(a^i, e^i).$$

We have following exact decomposition across policies, where $\widehat{X}$ is percentage change in X:

$$\widehat{Y/H} = \underbrace{\sum_{i,j} \omega_j^i \left[\widehat{h}_j^i - \widehat{H}\right]}_{\text{1. Relative hours (intensive margin)}} + \underbrace{\sum_{i,j} \omega_j^i \widehat{\rho}_j^i}_{\text{2. Sorting (extensive margin)}} + \underbrace{\sum_{i,j} \omega_j^i \widehat{h}_j^i \widehat{\rho}_j^i}_{\text{3. Interaction}} \quad , \quad \underbrace{\omega_j^i = \frac{y_j^i}{Y}}_{\text{Output share of match}}$$

[20] For below (above) median productivity workers, the share of workers at their productivity maximizing firm falls by 3 ppt (1 ppt), and the share of workers at firms with above median value-added per worker falls by 5 ppt (3 ppt).

In standard theories of taxation in the presence of worker heterogeneity and hours or effort choices (Mirrlees, 1971; Diamond and Saez, 2011), tax changes reduce productivity via the first term. If the hours of the highest output workers fall relative to overall hours, then labor productivity falls. The second term is novel, this is negative when workers reallocate from more productive to less productive matches.

The sorting term is 'missing' from standard theories of optimal taxation and is responsible for the entire decline in aggregate labor productivity (Figure 5B). Total labor productivity falls substantially: -1.4 percent. The intensive margin term actually contributes positively: $+0.5$ percent. By reducing consumption inequality, higher progressivity weakens the negative wealth effects on labor supply at the top of the distribution which contains the most productive workers. As their hours expand, productivity increases. The extensive margin sorting term accounts for more than the entire decline: -1.9 percent.

Higher tax progressivity can have significant, negative effects on output and productivity through standard channels of reductions in hours, and a new channel that reflects sorting of heterogeneous workers across heterogeneous firms. This suggests that optimal labor income taxes may be lower than implied by a theory that abstracts from these channels. Before getting to optimal taxes, we show that these features are empirically relevant. They deliver large declines in earnings following increases in marginal tax rates, consistent with the data.

6.2. Implications for the *Elasticity of Taxable Income*

Our theory resolves an important puzzle in public finance: standard models struggle to deliver large negative effects of increases in marginal tax rates on earnings, consistent with empirical estimates (see discussion in Keane and Rogerson, 2012; Chetty et al., 2011). In their survey, Saez et al. (2012) put the *elasticity of taxable income* between 0.12 to 0.40, while estimates in Kleven and Schultz (2014) are between 0.20 and 0.30 for a large tax reform in Denmark. The ETI is measured as

$$ETI = \frac{\Delta \log \text{Average Pre-tax Labor Earnings}}{\Delta \log(1-\overline{\tau})}$$

where $\overline{\tau}$ is the earnings-weighted average marginal tax rate.

Two issues in standard models lead to a small ETI. In partial equilibrium, models that are consistent with low estimates of labor supply elasticities produce small reductions in hours. In general equilibrium, higher taxes shift labor supply which pushes wages and employment in opposite directions, generating offsetting effects on earnings.

Our model addresses both issues. First, we maintain a low Frisch elasticity of labor supply (0.09), however we feature a high elasticity of labor supply on the extensive margin across firms. Disproportionately taxing wages in high productivity matches, pushes labor towards lower productivity matches, reducing earnings. Second, higher taxes reduce labor supply elasticities, widening markdowns, which effectively reduces labor demand: equilibrium wages decrease, amplifying the decline in earnings (Figure 5).

	Homogenous	Partial equilbrium		General
	firms	Fixed ρ, w	Fixed w	equilibrium
	(1)	(2)	(3)	(4)
A. Elasticity of Taxable Income				
ETI	0.07	0.05	0.16	0.24
B. Decomposition - ETI components				
Hours	0.07	0.05	0.05	0.05
Sorting	—	—	0.10	0.07
Wage	—	—	—	0.12
C. Decomposition - Shares				
Hours	100%	100%	35%	20%
Sorting	—	—	65%	29%
Wages	—	—	—	51%
— Reallocation				-2%
— Changing elasticities				53%
— ... on intensive margin				-10%
— ... on extensive margin				62%
— Direct effect				146%
— Heterogeneity effect				-82%
— Oligopsony effect				-3%
D. Changes in output and inequality				
$\Delta \log(Y)$	-0.012	-0.008	-0.032	-0.021
$\Delta \log(Y/H)$ (Productivity)	0.004	0.005	-0.020	-0.010
$\Delta \,\mathrm{StdDev}(\log c)$	-0.057	-0.059	-0.059	-0.054

Table 5: Decomposing the change in earnings following an increase in tax progressivity

Resolving both issues delivers an Elasticity of Taxable Income (ETI), in the range described by Saez et al. (2012) (0.12 to 0.40). While maintaining budget balance via τ_0, increasing τ_1 from 0.06 to 0.20 increases the average marginal tax rate from 34% to 45%. Total earnings decline by 4.3 percent. Hence the ETI is 0.24.

Quantitatively, the novel channels introduced in this paper dominate: sorting and lower wages, rather than a reduction in hours. In Table 5, column (4), we compute the share of the general equilibrium change in aggregate earnings due to hours, sorting and wages:

$$\begin{aligned} Earnings &= \sum_{a,e,j} \rho_j(a,e) w_j(e) h_j(a,e) \Gamma(a,e) \\ \Delta \log \text{Earnings} &\approx \sum_{a,e,j} \omega_j(a,e) \Big[\underbrace{\Delta \log \rho_j(a,e)}_{\text{Sorting: +29\%}} + \underbrace{\Delta \log w_j(e)}_{\text{Wages: +52\%}} + \underbrace{\Delta \log h_j(a,e)}_{\text{Hours: +20\%}} \Big]. \end{aligned}$$

Sorting and endogenous wages account for 80 percent of the ETI. In Appendix B.1 we decompose the Sorting term into the direct effect of $(1-\tau_1)$ on extensive margin elasticities in (7) and

a residual due to heterogeneity. The direct effect accounts for 76 percent, which informs how we simplify the model to study optimal taxes in Section 6.3.[21]

To understand these changes, Table 5 builds toward the full ETI. In column (1) we specialize our model to the nested neoclassical model with a continuum of identical firms. Competitive wages are equal to marginal products regardless of tax policy.[22] Increasing τ_1 only impacts hours, delivering a small ETI: 0.07. Column (2) shows that we obtain comparable results if we take the full change in hours in our baseline model, and hold wages and allocations of workers fixed across firms when computing the change in total earnings. This confirms that the large ETI is not obtained by amplifying the hours channel.

The first novel feature is that tax policy induces changes in sorting of workers across firms. The ETI triples if we keep wages fixed at their level under $\tau_1 = 0.06$, but allow workers to change firms and hours (column 3). Workers become less elastic, moving to lower wage firms which reduces total earnings. These results are consistent with the main empirical findings in Kleven et al. (2025): a small ETI when estimated on *stayers*, but much larger when computed on *switchers*.

The second novel feature is that tax policy reshapes labor supply elasticities, which lead to repricing of wages. Since marginal products are unchanged, this comes entirely from widening markdowns: $\Delta \log w_j(e) = \Delta \log \mu_j(e)$. The ETI increases by 50 percent when equilibrium markdowns change. Following Table 3, decomposing markdowns in three steps to reveal worker that heterogeneity is key to understanding these effects.

First, all widening of markdowns is due to declining elasticities rather than composition.

$$\begin{aligned} \mu_j(e) &= \frac{\varepsilon_j(e)}{\varepsilon_j(e)+1} \quad , \quad \varepsilon_j(e) = \sum_a \omega_j^h(a,e)\varepsilon_j(a,e) \\ \underbrace{\Delta \log \mu_j(e)}_{\text{Wages: +52\%}} &\approx \frac{1}{1+\varepsilon_j(e)} \Bigg[\underbrace{\sum_a \omega_j^\varepsilon(a,e)\Delta \log \omega_j^h(a,e)}_{\text{Within-}(e,j)\text{ composition across } a\text{: -2\%}} + \underbrace{\sum_a \omega_j^\varepsilon(a,e)\Delta \log \varepsilon_j(a,e)}_{\text{Changing elasticities: +54\%}} \Bigg] \end{aligned}$$

Changes in the within-(e,j) composition of workers by assets increases the average elasticity slightly, nudging wages higher. The main effect, however, is a decline in workers' elasticities.

Second, as we have emphasized throughout, elasticities move mostly on the *extensive margin*, which is the largest share of the overall elasticity:

$$\underbrace{\Delta \log \varepsilon_j(a,e)}_{\text{Changing elasticities: +54\%}} \approx \underbrace{\left(\frac{\varepsilon_j^\rho(a,e)}{\varepsilon_j(a,e)}\right)\Delta \log \varepsilon_j^\rho(a,e)}_{\text{Extensive margin: +64\%}} + \underbrace{\left(\frac{\varepsilon_j^h(a,e)}{\varepsilon_j(a,e)}\right)\Delta \log \varepsilon_j^h(a,e)}_{\text{Intensive margin: -10\%}}$$

[21] In column (3) and (4), 113% and 76% is due to the direct effect, respectively. See Appendix B.1.

[22] Parameters $\overline{Z}$ and $\overline{\varphi}$ are chosen to keep average earnings and hours the same as the full model under the baseline value of $\tau_1 = 0.06$. The remaining parameters are φ, σ, σ_e, which are taken from Table 1.

Third, the decline in the average extensive margin elasticity reflects large offsetting forces due to the core interaction between household heterogeneity and wage setting:

$$\underbrace{\Delta \log \varepsilon_j^{\rho}(a,e)}_{\text{Extensive margin: +64\%}} = \underbrace{\Delta \log (1-\tau)}_{\text{Direct effect: +146\%}} + \underbrace{\Delta \log u'(c_j(a,e))\widetilde{y}_j(a,e)}_{\text{Heterogeneity effect: -81\%}} + \underbrace{\Delta \log \Big(\theta \rho_j(a,e) + \eta(1-\rho_j(a,e))\Big)}_{\text{Oligopsony effect: -1\%}}$$

On the one hand, the *Direct effect* of higher progressivity is to make workers less elastic on the extensive margin, this is the leading term in our key equation (7) for $\varepsilon_j^{\rho}(a,e)$. This is large, encoded in wages by firms, and by itself would account for more than the entire decline in earnings. On the other hand, higher progressivity compresses consumption $c_j(a,e)$ and after tax earnings $\widetilde{y}_j(a,e)$. As the rich become more elastic and poor become less elastic, the earnings weighted elasticity increases, leading to higher wages. This *Heterogeneity effect* would be absent in a model without across-worker elasticity differences (e.g. log $u(c)$ and $a = 0$). Setting this to zero in our model leads to an ETI of 0.44, outside the range of empirical estimates.[23] Hence the heterogeneity channel is key to the aggregate response of earnings.

Summarizing our positive results, we find that sorting and endogenous wages have important implications for the effects of marginal tax rates on quantities and prices (Table 5D). Relative to a nested model with homogeneous firms, increasing progressivity leads to (a) larger declines in earnings (ETI), (b) output, and (c) labor productivity, while yielding similar declines in consumption inequality. These effects are driven by sorting and endogenous reductions in wages. They suggest that optimal progressivity may be lower when worker heterogeneity shapes labor supply elasticities and firms have labor market power. We now formalize and quantify this intuition.

6.3. Implications for *Optimal Tax Progressivity*

To speak to a wider literature, we consider a specialization of our model that (i) delivers analytical results, and (ii) reduces to the canonical frameworks in Benabou (2002) and Heathcote et al. (2020b) under the additional restriction of identical firms ($\sigma_z = 0$) and competitive wage-setting ($\mu_j^i = 0$). This captures the *Hours* effects and direct effects of taxes on *Sorting* and *Wages* from Table 5. Analytically, we derive optimal progressivity formulas that separate these effects. Quantitatively, we show that firm heterogeneity sharply reduces optimal progressivity and its associated welfare gains.

We specialize the economy to: (a) log utility in consumption ($\sigma = 1$), (b) no complementarities in production ($\psi = 0$), (c) log normally distributed firm productivity with standard deviation σ_z, (d) only labor income ($a^i = 0,\ T = 0$), (e) a continuum of firms ($J \to \infty$). Applying these to our main equations (7) for elasticities of labor supply, wages are $w(e,z) = \mu(\tau)ez$, where $\mu(\tau)$ is based on an elasticity $\varepsilon = (1-\tau_1)\eta$ that reflects only taste dispersion and tax progressivity on the extensive margin, with no intensive margin response due to offsetting income

[23] To obtain 0.43, we take the dampening of -81% and add it back on to the overall ETI (0.24): $1.81 \times 0.24 = 0.44$

and substitution effects. This preserves the *Direct effect* of Table 5. Appendix B.1 provides all derivations.

Sorting distortion. As in the full model, higher progressivity makes labor less elastic across firms with respect to pre-tax wages, reducing employment at high wage firms. In the restricted model we can characterize exactly how the equilibrium distribution of *employment weighted* firm productivity is tilted toward higher productivity firms:

$$\log z_j \mid \text{Worker sorting} \sim N\left(\mu_z + (1-\tau_1)\eta\sigma_z^2,\ \sigma_z^2\right) \tag{12}$$

By making workers less elastic with respect to wages, higher τ_1 partially undoes this sorting, reducing output via lower aggregate labor productivity. This preserves the *Sorting effect* on wages and impacts on output from Table 5. These effects erode the tax base.

Optimal taxes. We compute optimal tax progressivity τ_1^* of a utilitarian Government planner. The Government keeps average tax rates constant, returning additional revenue to households via the intercept τ_0. The Government also taxes profits at a rate $\gamma \in [0,1]$. Appendix B.1 shows that optimal progressivity τ_1^* equates the marginal benefit to the sum of distinct marginal costs:

$$\underbrace{\left(1-\tau_1^*\right)\left(\sigma_e^2+\sigma_z^2\right)}_{\text{Mechanical revenue benefit}} = \underbrace{\frac{(1-\gamma)\,\eta}{[\gamma+\eta(1-\tau_1^*)]\,[1+\eta(1-\tau_1^*)]}}_{\text{Leakage: } \omega(\tau_1^*)\partial\log\mu\left(\tau_1^*\right)/\partial\tau_1^*} + \underbrace{\tau_1^*\left(\frac{\varphi}{\varphi+1}\right)\left(\frac{1}{1-\tau_1^*}\right)}_{\text{Hours: } -\tau_1^*\partial\log h\left(\tau_1^*\right)/\partial\tau} + \underbrace{\tau_1^*\left(\eta\sigma_z^2\right)}_{\text{Sorting: } -\tau_1^* d\mathbb{E}_\rho[\log z_j]/d\tau_1^*} \tag{13}$$

Setting $\sigma_z = 0$ and removing markdowns delivers the case in Ferriere et al. (2023) with only the hours term on the right and inequality term on the left.[24] Figure 6 plots components of (13).

Benefits. The benefits of progressivity are mechanical, and depend on the variance of log match output: $\sigma_e^2 + \sigma_z^2$. More dispersion in worker skills increases revenue-net-of-payments from increasing progressivity. This can be visualized in Figure 6 as an *upward shift* in the black (benefit) curve, raising optimal τ_1^*.

Leakage. The *Leakage distortion* arises because markdowns endogenously widen with higher progressivity. This represents lost resources to the Government so long as all profits are not taxed ($\gamma < 1$). For transparency, and the fact that optimal profit taxes in our model are indeed 100%, we focus on the *Hours* and *Sorting* terms which are independent of γ, and are the only surviving terms under $\gamma = 1$. Appendix A.4 considers the case of $\gamma = 0$.

Hours. Raising progressivity reduces hours, and this behavioral effect pushes toward lower τ_1^*. If labor is inelastic on the intensive margin, via a lower φ, then this cost is low. This would *rotate outwards* the *Hours only* (green) curve in Figure 6, increasing τ_1^*. Relating to the previous

[24] More precisely, this formula can be obtained by differentiating the expression for welfare as a function of τ in equation (6) of Ferriere et al. (2023), under an assumption of zero Government spending.

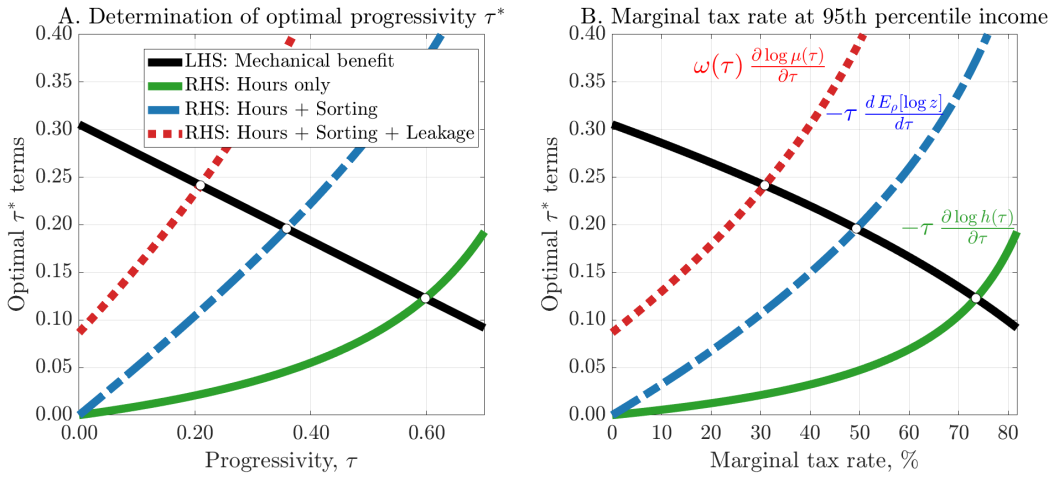


Figure 6: Determination of optimal tax progressivity

Notes: Panel A uses equation (13) to present the determination of optimal tax progressivity τ_1^*. The black line represents the marginal benefit of progressivity, which is decreasing in progressivity. The remaining lines represent the cumulative marginal costs of progressivity, which are increasing in progressivity. The green line represents only the *Hours* term from equation (13). The blue line is the sum of the *Hours* + *Sorting* terms. The red line is the sum of the *Hours* + *Sorting* + *Leakage* terms. Here we use the calibration described in the text, with $\varphi = 0.09$, corresponding to Row 1 of Table 6. Panel B presents the implied marginal tax rates at the 95th percentile of pre-tax income.

section, a low ETI based only on hours generates high optimal marginal tax rates (Table 5).

Sorting. Analytically, firm heterogeneity introduces two offsetting effects. The benefit is mechanical: more dispersion in wages yields a higher return from raising progressivity. The cost is novel, and depends on how the employment weighted productivity distribution (12) responds to τ_1. Raising progressivity reduces productivity, and this behavioral effect pushes toward lower τ_1^*. If labor is elastic on the extensive margin, via a higher η, or incurs large losses in output when sorting worsens, via a higher σ_z, then this cost is high. This would *rotate inwards* the *Hours + Sorting* (blue) curve in Figure 6, decreasing τ_1^*. Relating to the previous section, a high ETI based on sorting generates low optimal marginal tax rates.

Quantitative results. We can use equation (13) to quantify the forces behind τ_1^* consistent with our approach so far in the paper. This requires values for $\{\eta, \sigma_e, \sigma_z, \varphi\}$. Under $\tau_1 = 0.06$ (Table 1), $\eta = 4.11$ delivers an average extensive margin elasticity of 3.86 (Figure 1), $(\sigma_e^2 + \sigma_z^2) = 0.31$ delivers a variance of log consumption of 0.27 (Table 1), and $\sigma_z^2 = 0.10$ implies that a third of wage inequality is across-firms (Table 4). We vary $\varphi \in \{0.09, 0.50, 0.90\}$. Exploiting (13), we first compute τ_1^* with only the *Hours* term on right, then with *Hours and Sorting*. Table 6 provides optimal τ^*, the implied marginal tax rate at P95 income—we call this *'the marginal tax rate'* and use this to state all results—and the consumption equivalent welfare gain relative to τ_1 of 0.06.

Firm heterogeneity is quantitatively important for optimal taxes. Table 6 shows that when the negative sorting effect is incorporated, the optimal marginal tax rate falls by between 23 and 32 percent, depending on the value of φ. The marginal tax rate itself drops by between 10 and 24 percentage points. Incorporating sorting also cuts the welfare gains from optimizing

	A. Optimal τ^*				B. MTR* at P95 (%)				C. Welfare gain (%)			
Frisch (φ)	$Hours$	$+Sorting$	Δ	Ratio	$Hours$	$+Sorting$	Δ	Ratio	$Hours$	$+Sorting$	Δ	Ratio
0.09	0.60	0.36	-0.24	0.40	73.5	49.3	-24.1	0.33	8.83	3.99	-4.84	0.55
0.50	0.37	0.26	-0.11	0.29	50.3	37.4	-12.9	0.26	4.53	2.52	-2.01	0.44
0.90	0.31	0.23	-0.08	0.26	43.4	33.3	-10.1	0.23	3.48	2.03	-1.45	0.42

Table 6: Decomposing optimal tax progressivity

Notes: A. Reports the optimal level of tax progressivity, obtained by solving equation (13). We assume that σ_z^2 is such that one-third of the variance of wages is across firms, and σ_e^2 is such that the variance of log consumption is 0.27 (Table 1). B. Reports the implied marginal tax rate at the 95th percentile of pre-tax income. C. Reports the welfare gain associated with moving from $\tau_1 = 0.06$ (Table 1) to τ_1^*, stated in consumption equivalent units.

progressivity by 41 to 54 percent relative to the gains from optimization under only hours.[25] Adding *Leakage* reduces optimal progressivity, but the contribution of sorting relative to hours remains similar (Appendix Table A2).

Summary. We conclude that firm heterogeneity and labor market power provide a novel rationale for lower tax progressivity. We have shown this theoretically and quantitatively. A key obstacle to computing optimal tax progressivity in the full model is deciding how the tax system treats profits. In particular, if profits are not fully taxed ($\gamma < 1$), then how are they distributed among households. We leave this to future work.

7. Conclusion

In this paper we have provided a simple integration of key building blocks in the quantitative study of labor markets. First, the neoclassical model of consumption demand and hours supply (MaCurdy, 1981). Second, the discrete choice model of labor supply across firms (McFadden, 1974; Card et al., 2018). Third, the imperfect competition model of labor demand with a finite number of strategically engaged firms (Berger et al., 2022). Fourth, general equilibrium under production complementarities between firm and worker types. We purposefully avoided additions to each setting, and demonstrated that they fit together to provide new theoretical and quantitative insights into the functioning of labor markets.

From a positive perspective, a key result is that the interaction of these ingredients does not lead to counterfactual implications with regard to recent empirical results on firm and worker heterogeneity. First, standard decompositions of wage inequality into worker and firm fixed effects and their covariance. Second, cross-sectional dispersion in worker data on consumption, wages, earnings, hours, and firm data on employment and wages. Third, heterogeneity in reduced form elasticities of labor supply of different groups of workers to different types of

[25] Note that we have also removed transfers. Increasing the baseline τ_1 to 0.18 would be consistent with estimates of the tax system that load transfers back into net income. Under this baseline the effects of incorporating sorting are larger. Table A3 shows that when the negative sorting effect is incorporated, the optimal marginal tax rate falls by between 27 and 36 percent, depending on the value of φ. The marginal tax rate itself drops by between 14 and 28 percentage points. Incorporating sorting also cuts the welfare gains from optimizing progressivity: 80 to 82 percent less than the gains from optimization under only hours.

firms. Fourth, estimates of how labor supply responds to shocks to non-labor income. Our view is that it is a virtue of existing theoretical frameworks designed to handle each piece of these empirics that when combined together neither framework has counterfactual implications for the remaining data. This provides a unified quantitative theory.

From the perspective of counterfactual analysis, the model provides important results for sources of markdowns and effects of income tax policy. First, in general, more than half of markdowns are variable, and hence may respond to changes in policy. Second, across firms, the dominant source of markdown variation is the type of workers firms hire and the role that high pay plays in shaping workers' extensive margin elasticity of demand. Third, by focusing on a highly elastic margin of labor supply—across firms, rather than hours or participation—the model has new implications for the effect of income tax policy on aggregate productivity, wages and earnings.

Our paper provides a quantitative apparatus for the study of additional policies that interact with heterogeneity in workers' marginal value of wealth, and hence will lead firms to rethink how wages are set. As a leading example for future work, *Universal Basic Income* (UBI) has been studied in rich quantitative lifecycle models of consumption and savings in Daruich and Fernández (2024). Labor markets in that paper are competitive. Intuitively, however, the policy would change how elastically different workers' labor is supplied, something that firms would internalize in imperfectly competitive settings. Loosely, the idea that workers would be *less attached* to the labor market under UBI, would show up as less elastic labor, and hence lower wages. Other examples are unemployment insurance as in Acemoglu and Shimer (2000), provisions for worker retirement, minimum wages and other policies that shape how workers interact with the labor market via their marginal value of a dollar of earnings.

References

ABOWD, J. M., F. KRAMARZ, AND D. N. MARGOLIS (1999): "High wage workers and high wage firms," *Econometrica*, 67, 251–333.

ACEMOGLU, D. AND R. SHIMER (2000): "Productivity gains from unemployment insurance," *European Economic Review*, 44, 1195–1224.

ANDOLFATTO, D. (1996): "Business cycles and labor-market search," *American Economic Review*, 112–132.

ARSENEAU, D. M. AND S. K. CHUGH (2012): "Tax smoothing in frictional labor markets," *Journal of Political Economy*, 120, 926–985.

——— (2023): "Tax Smoothing in Frictional Labor Markets: A Reply," *Journal of Political Economy*, 131, 1372–1382.

AUCLERT, A., H. MALMBERG, F. MARTENET, AND M. ROGNLIE (2021): "Demographics, wealth, and global imbalances in the twenty-first century," NBER Working Paper 29161, National Bureau of Economic Research.

BAGGER, J., E. R. MOEN, AND R. M. VEJLIN (2021): "Equilibrium worker-firm allocations and the deadweight losses of taxation," Working paper.

BENABOU, R. (2002): "Tax and education policy in a heterogeneous-agent economy: What levels of redistribution maximize growth and efficiency?" *Econometrica*, 70, 481–517.

BERGER, D., K. HERKENHOFF, AND S. MONGEY (2025): "Minimum Wages, Efficiency, and Welfare," *Econometrica*, 93, 265–301.

BERGER, D., K. HERKENHOFF, S. MONGEY, AND N. MOUSAVI (2024): "Monopsony amplifies distortions from progressive taxes," in *AEA Papers and Proceedings*, vol. 114, 555–560.

BERGER, D. W., K. F. HERKENHOFF, AND S. MONGEY (2022): "Labor Market Power," *American Economic Review*, 112, 1147–93.

——— (2026): "Labor Market Power: From Micro Evidence to Macro Consequences," *Journal of Economic Perspectives*, 40, 93–114.

BILS, M., B. KAYMAK, AND K.-J. WU (2025): "Robinson meets Roy: Monopsony power and comparative advantage," NBER Working Paper 34165, National Bureau of Economic Research.

BONHOMME, S., T. LAMADON, AND E. MANRESA (2019): "A distributional framework for matched employer employee data," *Econometrica*, 87, 699–739.

BOROVIČKOVÁ, K. AND R. SHIMER (2024): "Assortative Matching and Wages: The Role of Selection," NBER Working Paper 33184, National Bureau of Economic Research.

CARD, D., A. R. CARDOSO, J. HEINING, AND P. KLINE (2018): "Firms and labor market inequality: Evidence and some theory," *Journal of Labor Economics*, 36, 13–70.

CHETTY, R., A. GUREN, D. MANOLI, AND A. WEBER (2011): "Are Micro and Macro Labor Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins," *American Economic Review*, 101, 471–475.

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

DARUICH, D. AND R. FERNÁNDEZ (2024): "Universal Basic Income: A Dynamic Assessment," *American Economic Review*, 114, 38–88.

DECKER, R., J. HALTIWANGER, R. JARMIN, AND J. MIRANDA (2013): "The Secular Decline in Business Dynamism in the US," *Manuscript, University of Maryland*.

DIAMOND, P. AND E. SAEZ (2011): "The Case for a Progressive Tax: From Basic Research to Policy Recommendations," *Journal of Economic Perspectives*, 25, 165–190.

FERRIERE, A., P. GRÜBENER, G. NAVARRO, AND O. VARDISHVILI (2023): "On the optimal design of transfers and income tax progressivity," *Journal of Political Economy Macroeconomics*, 1, 276–333.

GOLOSOV, M., M. GRABER, M. MOGSTAD, AND D. NOVGORODSKY (2024): "How Americans respond to idiosyncratic and exogenous changes in household wealth and unearned income," *The Quarterly Journal of Economics*, 139, 1321–1395.

HEATHCOTE, J., K. STORESLETTEN, AND G. L. VIOLANTE (2014): "Consumption and labor supply with partial insurance: An analytical framework," *American Economic Review*, 104, 2075–2126.

——— (2017): "Optimal Tax Progressivity: An Analytical Framework," *The Quarterly Journal of Economics*, 132, 1693–1754.

——— (2020a): "Optimal progressivity with age-dependent taxation," *Journal of Public Economics*, 189.

——— (2020b): "Presidential address 2019: How should tax progressivity respond to rising income inequality?" *Journal of the European Economic Association*, 18, 2715–2754.

HUMMEL, A. J. (2023): "Tax curvature," *Journal of Public Economics*, 224, 104921.

JAROSCH, G., J. S. NIMCZIK, AND I. SORKIN (2024): "Granular search, market structure, and wages," *Review of Economic Studies*, 91, 3569–3607.

KEANE, M. AND R. ROGERSON (2012): "Micro and macro labor supply elasticities: A reassessment of conventional wisdom," *Journal of Economic Literature*, 50, 464–76.

KLEVEN, H., C. T. KREINER, K. LARSEN, AND J. E. SØGAARD (2025): "Micro versus Macro Labor Supply Elasticities: The Role of Dynamic Returns to Effort," *American Economic Review*, 115, 2849–2890.

KLEVEN, H. J. AND E. A. SCHULTZ (2014): "Estimating Taxable Income Responses Using Danish Tax Reforms," *American Economic Journal: Applied Economics*, 6, 271–301.

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

KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2020): "Leave-out estimation of variance components," *Econometrica*, 88, 1859–1898.

LAMADON, T., J. LISE, C. MEGHIR, AND J.-M. ROBIN (2025): "Labor Market Matching, Wages, and Amenities," NBER Working Paper 32687, National Bureau of Economic Research.

LAMADON, T., M. MOGSTAD, AND B. SETZLER (2022): "Imperfect competition, compensating differentials, and rent sharing in the US labor market," *American Economic Review*, 112, 169–212.

LISE, J. AND F. POSTEL-VINAY (2020): "Multidimensional skills, sorting, and human capital accumulation," *American Economic Review*, 110, 2328–2376.

LOBEL, F. (2021): "Who benefits from payroll tax cuts? Market power, tax incidence and efficiency," *Market Power, Tax Incidence and Efficiency (May 28, 2021)*.

LORENZINI, L. (2026): "Who Competes for Whom? Monopsony in Ability-Segregated Labor Markets," *Working paper*.

MACURDY, T. E. (1981): "An empirical model of labor supply in a life-cycle setting," *Journal of Political Economy*, 89, 1059–1085.

MARIMON, R. AND F. ZILIBOTTI (1999): "Unemployment vs. mismatch of talents: Reconsidering unemployment benefits," *The Economic Journal*, 109, 266–291.

MCFADDEN, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, New York: Academic Press, 105–142.

MIRRLEES, J. A. (1971): "An Exploration in the Theory of Optimum Income Taxation," *The Review of Economic Studies*, 38, 175–208.

MOUSAVI, N. (2021): "Optimal Labor Income Tax, Incomplete Markets, Labor Market Power," .

SACHS, D., A. TSYVINSKI, AND N. WERQUIN (2020): "Nonlinear Tax Incidence and Optimal Taxation in General Equilibrium," *Econometrica*, 88, 469–493.

SAEZ, E., J. SLEMROD, AND S. H. GIERTZ (2012): "The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review," *Journal of Economic Literature*, 50, 3–50.

SCHEUER, F. AND I. WERNING (2017): "The taxation of superstars," *The Quarterly Journal of Economics*, 132, 211–270.

SEEGMILLER, B. (2025): "Valuing labor market power: The role of productivity advantages," Working paper.

SHIMER, R. AND L. SMITH (2000): "Assortative matching and search," *Econometrica*, 68, 343–369.

VERBOVEN, F. (1996): "The nested logit model and representative consumer theory," *Economics Letters*, 50, 57–63.

VOLPE, O. (2025): "Job Preferences, Labor Market Power, and Inequality," Working paper.

YEH, C., C. MACALUSO, AND B. HERSHBEIN (2022): "Monopsony in the US labor market," *American Economic Review*, 112, 2099–2138.

APPENDIX

This Appendix is organized as follows. Section A provides additional figures and tables. Section B provides derivations of mathematical formulas in the main text. Section C provides additional discussion referred to in the main text.

A. Additional figures and tables

A.1. Additional statistics on reduced form labor supply elasticities

In Table 1 we target the gap in reduced form labor supply elasticities (i) across workers, (ii) across firms. Table A1 gives the levels that underlie these gaps. It is clear that, conditional on worker type, elasticities are falling as we move up through the firm productivity distribution. It is also clear that, conditional on firm type, elasticities are falling as we move up through the worker wage distribution. Consistent with the data, the decline across firms is larger for lower wage workers than for high wage workers.

	A. Data - Seegmiller (2025)		**B. Model**	
	Low z	High z	Low z	High z
Low skill workers	8.44	5.72	8.96	6.38
High skill workers	1.72	1.07	2.19	1.36

Table A1: Levels of reduced form labor supply elasticities

A.2. Comparative statics of sorting

To further illustrate key properties of the model, we consider two simple counterfactuals and their implications for sorting.[26]

No complementarities. Figure A1(A) describes sorting under $\psi = 0$. Absent production complementarities, wealth effects dominate sorting. With $\sigma > 1$, high productivity workers—who also have higher assets on average—care less about wages and hence, in relative terms, are employed more in low wage firms. Meanwhile, low productivity workers are employed more in high wage firms. This is counterfactual with respect to the empirical covariance in worker and firm fixed effects. The *'Share of log wage variance - Covariance of worker and firm fixed effects'* from Table 1 flips from 0.16 to -0.76.

Flipped wealth effects. Figure A1(B) describes sorting under $\sigma = 0.5$. High wage workers produce more at high productivity firms and, under $\sigma < 1$, care more about wages when choosing where to work. Sorting become extreme. This is counterfactual with respect to the empirical gap in labor supply elasticities to firms of low skill and high skill workers. The *'Elasticity gap: Low - high wage workers'* from Table 1 flips from 5.69 to -1.90.

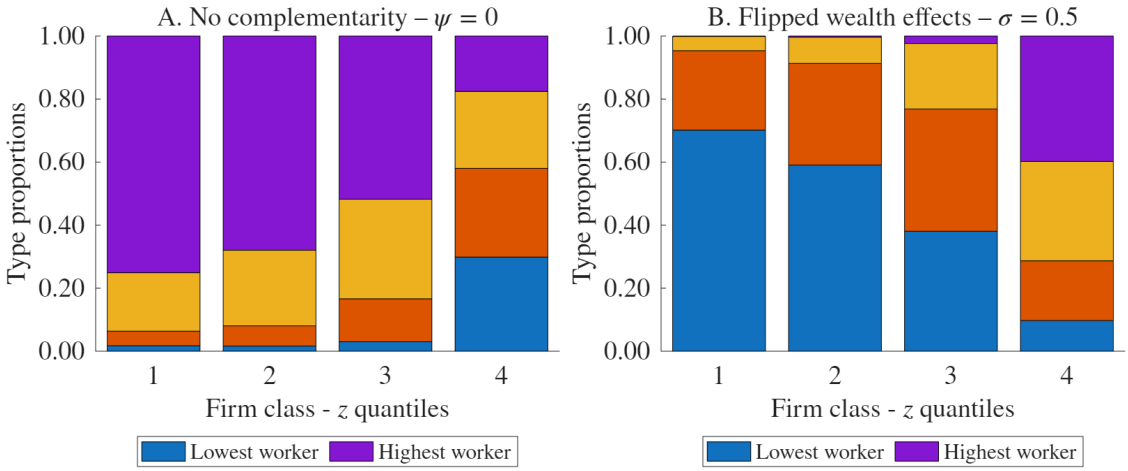


Figure A1: Sorting - Counterfactual patterns of sorting under $\psi = 0$ and $\sigma = 0.5$

[26] In each case $\overline{Z}$ and $\overline{\varphi}$ are recalibrated to keep average hours and wages constant.

A.3. Employment and hourly wages across firms

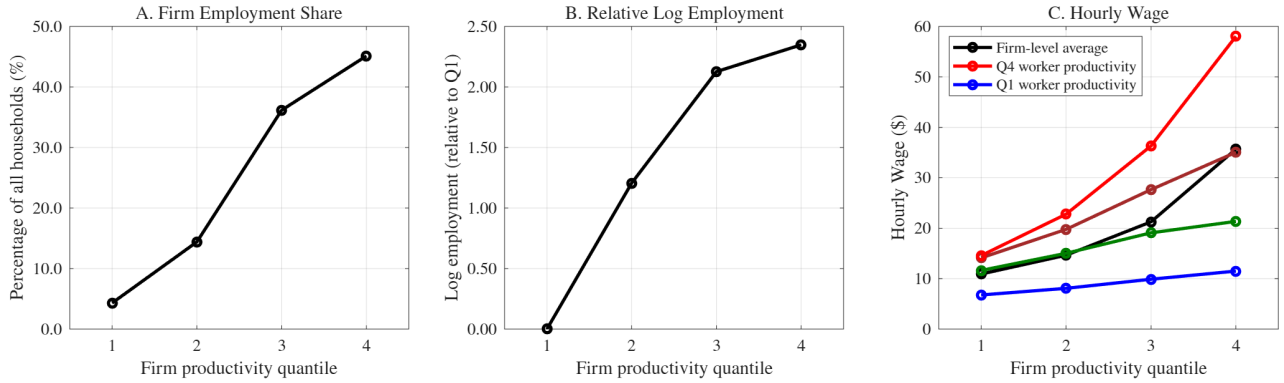


Figure A2: Employment and Hourly Wage

A.4. Optimal taxes under alternative assumptions on profit taxes

Table 6 did not present optimal taxes that incorporated the leakage term. Table A2 presents results in which both the *Hours* and *Hours + Sorting* columns also include *Leakage* under $\gamma = 0$.

	A. Optimal τ^*				**B. MTR* at P95 (%)**				**C. Welfare gain (%)**			
Frisch (φ)	$Hours$	$+Sorting$	Δ	Ratio	$Hours$	$+Sorting$	Δ	Ratio	$Hours$	$+Sorting$	Δ	Ratio
0.09	0.19	0.13	-0.06	0.33	27.6	19.1	-8.5	0.31	0.67	0.26	-0.42	0.62
0.50	0.14	0.10	-0.04	0.27	21.1	15.7	-5.4	0.26	0.35	0.13	-0.22	0.63
0.90	0.12	0.09	-0.03	0.25	18.7	14.3	-4.4	0.24	0.25	0.09	-0.16	0.65

Table A2: Decomposing optimal tax progressivity: Under no tax on profits ($\gamma = 0$)

Notes: A. Reports the optimal level of tax progressivity, obtained by solving equation (13), including *Leakage* in all cases. We assume that σ_z^2 is such that one-third of the variance of wages is across firms, and σ_e^2 is such that the variance of log consumption is 0.27 (Table 1). B. Reports the implied marginal tax rate at the 95th percentile of pre-tax income. C. Reports the welfare gain associated with moving from $\tau_1 = 0.06$ (Table 1) to τ_1^*, stated in consumption equivalent units.

A.5. Optimal taxes under different baseline τ_1

Table 6 used parameters calibrated under a baseline value of τ_1 of 0.06. Table A3 uses parameters calibrated under a baseline value of τ_1 of 0.18. This impacts the values of η, σ_e and σ_z.

	A. Optimal τ^*				**B. MTR* at P95 (%)**				**C. Welfare gain (%)**			
Frisch (φ)	$Hours$	$+Sorting$	Δ	Ratio	$Hours$	$+Sorting$	Δ	Ratio	$Hours$	$+Sorting$	Δ	Ratio
0.09	0.64	0.35	-0.29	0.46	77.8	48.9	-28.9	0.37	8.22	1.68	-6.54	0.80
0.50	0.41	0.27	-0.14	0.35	56.6	39.5	-17.1	0.30	3.26	0.65	-2.61	0.80
0.90	0.35	0.24	-0.11	0.32	49.8	35.9	-13.9	0.28	2.09	0.35	-1.73	0.83

Table A3: Decomposing optimal tax progressivity: Baseline $\tau_1 = 0.18$

Notes: A. Reports the optimal level of tax progressivity, obtained by solving equation (13). We assume that σ_z^2 is such that one-third of the variance of wages is across firms, and σ_e^2 is such that the variance of log consumption is 0.27 (Table 1). B. Reports the implied marginal tax rate at the 95th percentile of pre-tax income. C. Reports the welfare gain associated with moving from $\tau_1 = 0.06$ (Table 1) to τ_1^*, stated in consumption equivalent units.

B. Additional derivations

B.1. Derivation of optimal tax formulas

- We consider the following specialization of the full model presented in the text. This is obtained by limits of existing model, hence is a strict specialization of our full model.

  1. Preferences are log in consumption ($\sigma = 1$)
  2. No non-labor income: (i) $a^i = 0$, (ii) transfers are zero ($T = 0$). In the notation of the paper, consumption c_j^i is equal to after-tax labor income $\widetilde{y}_j^i = (1 - \tau_0) \left(w_j (e^i) h_j^i\right)^{1-\tau_1}$.
     - To simplify on notation, we will use $\lambda = (1 - \tau_0)$, and write $\tau = \tau_1$.
     - In equilibrium, we require that the tax system is revenue neutral. This implies that $\lambda(\tau)$ is a function of τ.
  3. Log-modular production ($\psi = 0$)
  4. Log normal distribution of firm productivity $\log z \sim N(\mu_z, \sigma_z^2)$
  5. Continuum of firms in each market ($M \to \infty$) .

- Note that firms are still monopolistically competitive within their labor market, they are simply small relative to the market

- **Assumption** - *We assume that $\gamma \in [0, 1]$ fraction of firm profits are taxed.*

- Optimally, a planner would set γ to 1, taxing all of profits. However, we entertain $\gamma < 1$ to show how this impacts results.

B.1.A. Firms

- Firm profits for each type of worker are as follows:

$$\begin{aligned}
\pi_j(e) &= \max_{w_j(e)} \; z_j e \times H(e, w_j) - w_j(e) H(e, w_j) \\
H(e, w_j) &= \rho_j(e) h_j(e) \\
\rho_j(e) &= \frac{w_j(e)^{\eta(1-\tau)}}{\int w_k(e)^{\eta(1-\tau)} dk} \\
h_j(e) &= \overline{\varphi}^{\frac{1}{\varphi+1}} (1 - \tau)^{\frac{\varphi}{\varphi+1}}
\end{aligned}$$

- The extensive $\rho_j(e)$ and intensive $h_j(e)$ labor supply conditions come from household optimality, which we derive below.. These can be obtained from the equations in the text under $\sigma = 1$, and $c_j^i = \widetilde{y}_j^i$.

- For type e labor, the firm faces a labor supply elasticity:

$$\varepsilon_j(e) = \varepsilon_j^{\rho}(e) + \varepsilon_j^{h}(e)$$
$$\varepsilon_j(e) = \eta(1-\tau) + 0$$

- First, the firm is atomistic within the market, and hence there is no *oligoposony term* in the elasticity. Second, with offsetting income and substitution effects, hours are independent of wages.

- The firm therefore set wages

$$w_j(e) = \mu_j(z_j e) \quad , \mu_j = \frac{\varepsilon_j}{\varepsilon_j + 1} \quad , \quad \varepsilon_j = \eta(1-\tau)$$

- Define the following functions:

$$h(\tau) = \overline{\varphi}^{\frac{1}{\varphi+1}}(1-\tau)^{\frac{\varphi}{\varphi+1}} \quad , \quad \mu(\tau) = \frac{(1-\tau)\eta}{(1-\tau)\eta + 1}$$

- Two observations are important: (i) the markdown is independent across firms and workers, but still encodes the *Direct Effect* that we found was quantitatively responsible for the large changes in earnings via reduced sorting when we decomposed the ETI, (ii) hours are independent across firms and workers.

B.1.B. Household

- We draw on the representation of discrete choice problems used in Mongey and Waugh (2025). This represents the problem as a direct optimization problem over the choice probability ρ_j. As shown in Mongey and Waugh (2025), writing the problem in this way allows for clear applications of the envelope theorem, which facilitate standard welfare analysis.

- Welfare for an individual with skill e is

$$W(e,\tau) = \max_{\rho_j, c_j, h_j} \sum_j \rho_j \left[\log c_j - \frac{1}{\overline{\varphi}^{1/\varphi}} \frac{h_j^{1+1/\varphi}}{1+1/\varphi} - \frac{1}{\eta}\log\rho_j\right] + \sum_j \Lambda_j \left[\lambda(\tau)(w_j(e)h_j)^{1-\tau} - c_j\right] + \psi\left[1 - \sum_j \rho_j\right]$$

- The first order condition for consumption is

$$\rho_j = \Lambda_j c_j$$

- The first order condition for hours is as follows, and delivers the intensive margin supply curve given above:

$$\begin{aligned}
\rho_j \frac{1}{\overline{\varphi}^{1/\varphi}} h_j^{\frac{1}{\varphi}} &= \Lambda_j (1-\tau) \lambda(\tau) (w_j(e) h_j)^{-\tau} w_j(e) \\
\rho_j \frac{1}{\overline{\varphi}^{1/\varphi}} h_j^{1+\frac{1}{\varphi}} &= \Lambda_j (1-\tau) \left[\lambda(\tau) (w_j(e) h_j)^{1-\tau}\right] \\
\frac{1}{\overline{\varphi}^{1/\varphi}} h_j^{1+\frac{1}{\varphi}} &= \left(\frac{\Lambda_j}{\rho_j}\right) (1-\tau) [c_j] \\
h_j(\tau) &= \overline{\varphi}^{\frac{1/\varphi}{1+1/\varphi}} (1-\tau)^{\frac{1}{1+1/\varphi}}
\end{aligned}$$

- Equating first order conditions for ρ_j and ρ_k:

$$\begin{aligned}
\log c_j - \frac{1}{\overline{\varphi}^{1/\varphi}} \frac{h_j^{1+1/\varphi}}{1+1/\varphi} - \frac{1}{\eta}\log\rho_j - \frac{1}{\eta} - \psi &= \log c_k - \frac{1}{\overline{\varphi}^{1/\varphi}} \frac{h_k^{1+1/\varphi}}{1+1/\varphi} \frac{1}{\eta}\log\rho_k - \frac{1}{\eta} - \psi \\
\log c_j - \frac{1}{\eta}\log\rho_j &= \log c_k - \frac{1}{\eta}\log\rho_k \\
\frac{\rho_j}{\rho_k} = \left(\frac{c_j}{c_k}\right)^{\eta} &= \left(\frac{\lambda(\tau)(w_j(e) h(\tau))^{1-\tau}}{\lambda(\tau)(w_k(e) h(\tau))^{1-\tau}}\right)^{\eta} = \left(\frac{w_j(e)}{w_k(e)}\right)^{\eta(1-\tau)}
\end{aligned}$$

  Using $\sum_k \rho_k(e) = 1$, we obtain

$$\rho_j(e) = \frac{w_j(e)^{\eta(1-\tau)}}{\sum_k w_k(e)^{\eta(1-\tau)}}$$

  This gives the extensive margin supply curve taken as given by the firm.

- Combining the choice probabilities of the worker, and the equilibrium wages of the firm:

$$\begin{aligned}
\rho_j(e) &= \frac{(\mu(\tau) z_j e)^{\eta(1-\tau)}}{\sum_k (\mu(\tau) z_k e)^{\eta(1-\tau)}} \\
\rho_j(e) &= \frac{z_j^{\eta(1-\tau)}}{\sum_k z_k^{\eta(1-\tau)}}
\end{aligned}$$

B.1.C. Planner's optimality condition

- The planner chooses τ to maximize welfare. Under the unit measure of individuals $\Gamma(e)$:

$$W(\tau) = \mathbb{E}[W(\tau, e)]$$

- Our approach is to take the derivative of $W(\tau, e)$ directly, exploiting envelope conditions.

- Recall individual utility, where we now substitute in the equilibrium wage, which the planner understands when choosing τ:

$$W(e,\tau) = \max_{\rho_j, c_j, h_j} \sum_j \rho_j \left[\log c_j - \frac{1}{\overline{\varphi}^{1/\varphi}} \frac{h_j^{1+1/\varphi}}{1+1/\varphi} - \frac{1}{\eta} \log \rho_j \right] + \sum_j \Lambda_j \left[\lambda(\tau) \left(\mu(\tau) z_j e h_j \right)^{1-\tau} - c_j \right] + \psi \left[1 - \sum_j \rho_j \right]$$

- We take the derivative with respect to τ and then evaluate $\rho_j, c_j, h_j, \Lambda_j$ at the optimal policies. Let $y_j(\tau, e) = \mu(\tau) z_j e h(\tau)$

$$W_\tau(e,\tau) = \sum_j \Lambda_j(e) \left[\frac{\partial \lambda(\tau)}{\partial \tau} y_j(\tau,e)^{1-\tau} - \lambda(\tau) y_j(\tau,e)^{1-\tau} \left(\log y_j(\tau,e) - (1-\tau) \frac{\partial \log \mu(\tau)}{\partial \tau} \right) \right]$$

- Using $\Lambda_j(e) = \rho_j(e)/c_j(e)$ and $c_j(e) = \lambda(\tau) y_j(\tau,e)^{1-\tau}$:

$$W_\tau(e,\tau) = \sum_j \rho_j(e) \left[\frac{\partial \log \lambda(\tau)}{\partial \tau} - \log y_j(\tau,e) + (1-\tau) \frac{\partial \log \mu(\tau)}{\partial \tau} \right]$$

$$W_\tau(e,\tau) = \frac{\partial \log \lambda(\tau)}{\partial \tau} - \left(\log \mu(\tau) + \left[\sum_j \rho_j(e) \log z_j \right] + \log + \log h(\tau) \right) + (1-\tau) \frac{\partial \log \mu(\tau)}{\partial \tau}$$

- Note that $\rho_j(e)$ is independent of e, and can be written $\rho(z)$. Let $\mathbb{E}_\rho[\log z_j]$ denote the employment-weighted mean of log productivity:

$$\mathbb{E}_\rho[\log z_j] = \mathbb{E}_z \left[\frac{z^{\eta(1-\tau)}}{\mathbb{E}_z\left[z^{\eta(1-\tau)}\right]} \log z \right]$$

- Given that $\mathbb{E}[\log e] = 0$, then we have the planner's optimality condition:

$$\frac{\partial \log \lambda(\tau)}{\partial \tau} = \left[\log \mu(\tau) + \mathbb{E}_\rho[\log z_j] + \log h(\tau) \right] - (1-\tau) \frac{\partial \log \mu(\tau)}{\partial \tau}.$$

- **Interpretation** - This is a clean expression that characterizes the planner's optimality condition for progressivity.
  - On the right hand side is average log earnings. The adjustment reflects that at this value of τ, increasing in τ further widens markdowns. Therefore the last term is positive.
  - The LHS is a decreasing function of τ. As τ increases, the additional revenue from further increases falls. Therefore, increasing the size of the RHS will reduce optimal

τ.

- However note that this argument is incomplete. A decrease in $\mu(\tau)$ reduces worker consumption. However if the planner taxes the implied increase in profits to the firm and rebates these back to workers via $\lambda(\tau)$, then the net effect may be smaller. Below we show that this is exactly the case. In fact *all* terms on the RHS will be recycled back to workers and hence also appear in $\partial \log \lambda(\tau) / \tau$.

- **Nested case** - We will refer to the *standard model* as one in which $\sigma_z = 0$, such that there is no firm heterogeneity, and $\mu(\tau) = 1$, as wages are set competitively. In this case, if μ_z were such that mean log productivity was equal to zero, then, this would reduce to:

$$\frac{\partial \log \lambda(\tau)}{\partial \tau} = \log h(\tau).$$

B.1.D. Government budget

- We next compute the derivative on the left.
- The government taxes γ share of firm profits.
- Hence, government revenue is all worker earnings, $\mu(\tau) e z_j h(\tau)$, and γ share of firm profits $\gamma(1-\mu(\tau)) e z_j h(\tau)$.

$$\begin{aligned} Revenue(e, z_j) &= \mu(\tau) y_j(e) + \gamma(1-\mu(\tau)) y_j(e) \\ Revenue(e, z_j) &= [\gamma + (1-\gamma)\mu(\tau)] y_j(e) \end{aligned}$$

- If $\gamma = 1$, then the government absorbs all output $y_j(e)$. If $\gamma < 1$, then a fraction $(1-\gamma)\mu_j(\tau)$ is 'leaks' out of the income tax system to the owners of firms. The planner doesn't like this leakage, which worsens as τ increases and $\mu(\tau)$ widens.
- Define $\omega(\tau)$ as follows:

$$\begin{aligned} \omega(\tau) &= \frac{(1-\gamma)\mu(\tau)}{\gamma + (1-\gamma)\mu(\tau)} \\ \frac{\partial \log[\gamma + (1-\gamma)\mu(\tau)]}{\partial \tau} &= \omega(\tau) \frac{\partial \log \mu(\tau)}{\partial \tau} \end{aligned}$$

- The government then returns post-tax income to all workers:

$$\lambda(\tau)\int\int \rho_j\left(\mu(\tau) z_j e h(\tau)\right)^{1-\tau} djdF(e) = \left[\gamma+(1-\gamma)\mu(\tau)\right]\int\int \rho_j z_j e h(\tau)\, djdF(e)$$

$$\lambda(\tau) = \left[\gamma+(1-\gamma)\mu(\tau)\right]\frac{\mathbb{E}_e[e]\,\mathbb{E}_\rho\left[\sum_j \rho_j z_j\right]}{\mathbb{E}_e\left[e^{1-\tau}\right]\mathbb{E}_\rho\left[\sum_j \rho_j z_j^{1-\tau}\right]}\mu(\tau)^{-(1-\tau)} h(\tau)^{\tau}$$

- Then

$$\log\lambda(\tau) = \log\left[\gamma+(1-\gamma)\mu(\tau)\right] + \left(\log\mathbb{E}_e[e] - \log\mathbb{E}_e\left[e^{1-\tau}\right]\right) + \left(\log\mathbb{E}_\rho[z_j] - \log\mathbb{E}_\rho\left[z_j^{1-\tau}\right]\right) - (1-\tau)\log\mu(\tau) - \tau\log h(\tau)$$

- Starting with the *worker productivity* term:

$$\log e \sim N\left(0,\sigma_e^2\right)$$
$$\log e^{1-\tau} \sim N\left(0,(1-\tau)^2\sigma_e^2\right)$$
$$\log\mathbb{E}_e[e] - \log\mathbb{E}_e\left[e^{1-\tau}\right] = \left\langle\frac{1}{2}\sigma_e^2\right\rangle - \left\langle(1-\tau)^2\frac{1}{2}\sigma_e^2\right\rangle$$
$$\frac{d}{d\tau}\left(\log\mathbb{E}_e[e] - \log\mathbb{E}_e\left[e^{1-\tau}\right]\right) = (1-\tau)\sigma_e^2$$

  - This has the interpretation of the *net revenue* that the government obtains. It brings in e from the log normal distribution, but it only has to pay out according to the left-shifted distribution. At the margin, on net, increasing $1-\tau$ increases revenue by exactly σ_e^2. There is no interaction of the new parts of the model with this term.

- Moving to the *firm productivity* term.

  - Note that, since $\rho_j \propto z_j^{\eta(1-\tau)}$, then the distribution of $\log z$ *conditional on this sorting* is given by:

$$\log z|\text{Sorting} \sim N\left(\mu_z+\eta(1-\tau)\sigma_z^2,\sigma_z^2\right)$$
$$\log z^{1-\tau}|\text{Sorting} \sim N\left((1-\tau)\mu_z+\eta(1-\tau)^2\sigma_z^2,(1-\tau)^2\sigma_z^2\right)$$

  - As η increases, or $\tau\to 0$ (linear taxes), workers become more elastic, sorting to higher wages, which implies higher z. This shifts the distribution to the right. The fact that the variance remains the same does not generalize to other distributions. Under a Pareto distribution, the variance also shifts.

– Using these distributions:

$$\log \mathbb{E}_\rho [z_j] - \log \mathbb{E}_\rho \left[z_j^{1-\tau}\right] = \left\langle \mathbb{E}_\rho [\log z_j] + \frac{1}{2}\sigma_z^2 \right\rangle - \left\langle (1-\tau)\, \mathbb{E}_\rho [\log z_j] + \frac{1}{2}(1-\tau)^2 \sigma_z^2 \right\rangle$$

$$\log \mathbb{E}_\rho [z_j] - \log \mathbb{E}_\rho \left[z_j^{1-\tau}\right] = \left\langle \tau \mathbb{E}_\rho [\log z_j] + \frac{1}{2}\sigma_z^2 \right\rangle - \left\langle \frac{1}{2}(1-\tau)^2 \sigma_z^2 \right\rangle$$

$$\frac{d}{d\tau}\left(\log \mathbb{E}_\rho [z_j] - \log \mathbb{E}_\rho \left[z_j^{1-\tau}\right]\right) = \mathbb{E}_\rho [\log z_j] + \tau \frac{d\mathbb{E}_\rho [\log z_j]}{d\tau} + (1-\tau)\sigma_z^2$$

– The second term is novel and gives the downward shift in mean log productivity due to an increase in τ. At the margin, the planner loses τ of this shift on net.

- Combining terms:

$$\log \lambda(\tau) = \log[\gamma + (1-\gamma)\mu(\tau)] + \left(\log \mathbb{E}_e [e] - \log \mathbb{E}_e \left[e^{1-\tau}\right]\right) + \left(\log \mathbb{E}_\rho [z_j] - \log \mathbb{E}_\rho \left[z_j^{1-\tau}\right]\right) - (1-\tau)\log \mu(\tau) - \tau \log h(\tau)$$

$$\frac{\partial \log \lambda(\tau)}{\partial \tau} = (1-\tau)\left(\sigma_e^2 + \sigma_z^2\right) + [\log \mu(\tau) + \mathbb{E}_\rho [\log z_j] + \log h] + \omega(\tau)\frac{\partial \log \mu(\tau)}{\partial \tau} - (1-\tau)\frac{\partial \log \mu(\tau)}{\partial \tau} + \tau \frac{\partial \log h(\tau)}{\partial \tau} + \tau \frac{d\mathbb{E}_\rho [\log z_j]}{d\tau}$$

- Compare this to the first order condition

$$\frac{\partial \log \lambda(\tau)}{\partial \tau} = [\log \mu(\tau) + \mathbb{E}_\rho [\log z_j] + \log h(\tau)] - (1-\tau)\frac{\partial \log \mu(\tau)}{\partial \tau}.$$

- **Result** - *Under 100% profit taxes, the effect of progressivity on markdowns drops out of the optimal tax formula*.

– Under $\gamma = 1$, then there is no *leakage*, and $\omega(\tau) = 0$. Hence the additional markdown terms cancel.

– Fix an allocation on the extensive and intensive margin. The amount that the worker is paid is incidental to the planner, as lower pay simly means higher profits, which are then redistributed back to workers. Under log utility, all workers value that redistribution identically. To a first order, the planner does precisely this computation, i.e. fixing an allocation.

– Hence, the planner only cares about how the changes in wages due to changes in taxes impact *allocations*.

– We also see this in the terms that reflect mean earnings. These are taxed and recycled

back to households, and hence drop out of both sides of the optimal tax expression.

$$\mathbb{E}\left[\log y_j\left(e\right)\right] = \int \sum_j \rho_j \log \mu\left(\tau\right) z_j e h\left(\tau\right) dF\left(e\right)$$
$$\mathbb{E}\left[\log y_j\left(e\right)\right] = \log \mu\left(\tau\right) + \mathbb{E}_\rho\left[\log z_j\right] + \mathbb{E}_e\left[\log e\right] + \log h\left(\tau\right)$$

B.1.E. Optimal tax formula

- Combining, we find that *all* terms in the first order condition representing the effect of taxes on individual welfare *holding fixed* $\lambda\left(\tau\right)$ net out against terms that appear in the equilibrium change in $\lambda\left(\tau\right)$. We are left with the following:

$$\left(1-\tau\right)\left(\sigma_e^2+\sigma_z^2\right) = -\omega\left(\tau\right)\frac{\partial \log \mu\left(\tau\right)}{\partial \tau} - \tau\frac{\partial \log h\left(\tau\right)}{\partial \tau} - \tau\frac{d\mathbb{E}_\rho\left[\log z_j\right]}{d\tau}$$

- At the optimum, the government balances out marginal fiscal gain obtained by exploiting dispersion in the tail of income, against marginal fiscal loss associated with behavioral responses and leakage via widening markdowns.

- In terms of marginal gain: at each income level, the planner gets \$1 of income, but only has to pay out $(1-\tau)$. The left-hand side represents an integral where that change in τ occurs along the entire distribution. It can be shown that without any restrictions on the distribution of e and z, the right hand side remains the same. It is completely robust to distributional assumptions. When we move to more general distributions of productivity, the left-hand side changes.

- The right hand side contains the *behavioral effects* of higher progressivity. Higher τ reduces hours, and this reduces net revenue by τ. Higher τ reduces aggregate productivity, and this reduces net revenue by τ.

- From above we can compute these derivatives:

$$\left(1-\tau\right)\left(\sigma_e^2+\sigma_z^2\right) = -\omega\left(\tau\right)\frac{\partial \log \mu\left(\tau\right)}{\partial \tau} - \tau\frac{\partial \log h\left(\tau\right)}{\partial \tau} - \tau\frac{d\mathbb{E}_\rho\left[\log z_j\right]}{d\tau}$$
$$\underbrace{\left(1-\tau\right)\left(\sigma_e^2+\sigma_z^2\right)}_{\text{Marginal revenue benefit, fixing behavior}} = \underbrace{\omega\left(\tau\right)\left(\frac{1}{1-\tau}\right)\left(\frac{1}{\eta\left(1-\tau\right)+1}\right)}_{\text{Marginal leakage cost}} + \underbrace{\tau\left(\frac{\varphi}{\varphi+1}\right)\left(\frac{1}{1-\tau}\right)+\tau\eta\sigma_z^2}_{\text{Marginal behavioural costs}}$$

- This expression keeps terms in their respective places.

- Marginal revenue gain from taxes is *decreasing in* τ, while the marginal behavioral cost of taxes is *increasing in* τ. This delivers a unique τ^*.

- A higher value of σ_z^2 increases the marginal revenue gains, by delivering more dispersion in earnings, which can be taxed. It also increases the marginal behavioral costs, as reallocation of workers across firms leads to steeper revenue losses. In the text, our numerical approach neutralizes the first term, by comparing economies indexed by $\zeta = \sigma_z^2/\left(\sigma_z^2 + \sigma_e^2\right)$, while keeping $\sigma_z^2 + \sigma_e^2$ constant.

- **Optimal tax formula**

  - In the text, we re-arrange this formula to obtain the following:

$$\frac{\tau}{(1-\tau)^2} = \frac{\sigma_e^2 + \sigma_z^2}{\left(\frac{\varphi}{\varphi+1}\right) + (1-\tau)\,\eta\sigma_z^2 + \frac{\omega(\tau)}{\tau}\left(\frac{1}{\eta(1-\tau)+1}\right)}.$$

  - Larger distrortions in the denominator push τ downards. The three distortions enter the denominator. The left hand side is *increasing* in τ. Hence, as a simple example, a higher Frisch φ makes hours more elastic to taxes, increasing the hours distortion, reducing the denominator: $\lim_{\varphi\to\infty}\left(\frac{\varphi}{\varphi+1}\right) \nearrow 1$.

- **Special cases**

  - Under the case of *no leakages*, $\omega(\tau) = 0$, in which case we obtain:

$$\frac{\tau}{(1-\tau)^2} = \frac{\sigma_e^2 + \sigma_z^2}{\left(\frac{\varphi}{\varphi+1}\right) + (1-\tau)\,\eta\sigma_z^2}$$

  - In the *standard model*, $\sigma_z^2 = 0$, the only trade-off is against the behavioural cost due to hours changes. A higher Frisch φ leads to lower optimal τ.

$$\frac{\tau}{(1-\tau)^2} = \frac{\sigma_e^2}{\left(\frac{\varphi}{\varphi+1}\right)}$$

C. Additional discussion

C.1. Comparison to Volpe (2025)

As an application of the identification arguments described in the paper, we contrast *sorting via wealth effects and complementarities* to *sorting via amenities* in Volpe (2025).

In Volpe (2025) labor is only supplied on the extensive margin across firms and utility is:

$$U_j^i = \beta^i \log w_j(e^i) + \frac{1}{\eta} \log \phi_j(e^i) + \zeta_j^i.$$

Households differ in skill e^i, and within each skill, there is heterogeneity in β^i, the coefficient on log wages in utility. We simplify to $e^i \in \{e^L, e^H\}$ and $\beta^i \in \{\beta^L, \beta^H\}$. Firms have heterogeneous amenity values and these are valued differently by workers of different skills, $\phi_j(e^i)$. These values are independent of β^i.

Applying our derivations in the main text, labor demand and firms' elasticity are:

$$\rho_j\Big(e^i, \beta^i\Big) = \frac{\phi_j(e^i) w_j(e^i)^{\eta\beta^i}}{\sum_k \phi_k(e^i) w_k(e^i)^{\eta\beta^i}} \quad , \quad \varepsilon_j(e^i) = \eta \int \frac{\rho_j(e^i, \beta^i)}{\int \rho_j(e^k, \beta^k)\, dk}\, \beta^i \, di$$

The individual elasticity is $\varepsilon_j^i = \eta\beta^i$, and depends only on an individual's price sensitivity β^i and dispersion in preferences η.

Sorting can then be thought of in two ways. First, consider sorting by β, conditional on e. Fix e and consider two firms $z_1 > z_2$. The measure of super-modularity can be solved by hand, and depends only on differences in β^i across individuals and the differences in firm wages.

$$\Upsilon_\rho(e) = \log\left(\frac{\rho_1(e, \beta^H)/\rho_1(e, \beta^L)}{\rho_2(e, \beta^H)/\rho_2(e, \beta^L)}\right) = \eta\Big(\beta^H - \beta^L\Big)\Big(\log w_1(e) - \log w_2(e)\Big). \tag{B1}$$

This is always *positive*. High-β types sort to high wage firms, implying a higher ε_j at these firms and narrower markdowns. This is inconsistent with **Fact 3** for U.S. and Brazilian data, as described in the main text.

Second, consider sorting by e, conditional on β. Fixing β, it is straight-forward to show that $\Upsilon_\rho(\beta) = \Upsilon_\phi$. That is, sorting is driven entirely by workers' systematic differences in amenity values at high and low wage firms.

In this environment, **Fact 1** (high skill workers have lower elasticities) can only be rationalized by high skill types having lower β's. However, by itself, this yields *negative sorting*: high skill workers would sort toward low wage firms. Hence, matching positive sorting in the data (**Fact 2**) requires high e workers obtain systematically higher amenities from high wage firms: $\Upsilon_\phi > 0$.

C.2. Simulation of Seegmiller (2025) in structural model

When estimating the model, we simulate the experiment implied by the empirical approach in Seegmiller (2025). First, assign all workers in the economy into three groups based on labor income: $q_w \in \{1, \ldots, Q_w\}$. The first group corresponds to quintiles one and two of wages, the second corresponds to quintiles three and four, the third corresponds to the top quintile. Second, assign all firms in the economy into employment-weighted quartiles of value-added per worker, denoted $q_f \in \{1, \ldots, Q_f\}$. We implement the following algorithm, which consists of three steps:

- **Step 1 - Construct comparable data**
- For each $j \in \{1, \ldots, J\}$
    - **Exercise** j
    - Multiply firm j productivity in all matches $A(z, e)$ by a factor $\chi > 1$. All other firms' productivity within the market is constant. All other markets are unaffected. Since each market is infinitesimal, this implies that $\overline{v}^i$, the worker's value in other markets, is unchanged.
    - Solve again for the equilibrium. This leads to reallocation of labor into the market, and across firms within the market.
    - Using the baseline and counterfactual equilibrium, assemble measures of average earnings and employment at firm j for each group q_w of workers: n_{j,q_w}, $\overline{y}_{j,q_w}$.
    - Repeat this exercise for all j firms.
- **Step 2 - Compute comparable treatment effects**
- We have $Q_w \times Q_f$ cells of worker×firm types. For each cell, we seek to measure the effect of the shock on log employment and log average wage. Denote these estimands $\beta^n_{q_w,q_f}$ and $\beta^y_{q_w,q_f}$.
- For each firm group $q_f \in \{1, \ldots, Q_f\}$
    - Take all treated firms $j \in q_f$
    - For each worker group $q_w \in \{1, \ldots, Q_w\}$, we compute the unweighted mean of $\Delta \log n_{j,q_w}$ and $\Delta \log y_{j,q_w}$, across firms within the firm group. Denote these $\beta^n_{q_w,q_f}$ and $\beta^y_{q_w,q_f}$.
- We now have $Q_w \times Q_f$ log earnings and log employment coefficients.
- From these we compute the *reduced form labor supply elasticities* $\epsilon_{q_w,q_f} = \beta^n_{q_w,q_f}/\beta^y_{q_w,q_f}$. These are comparable to the coefficients in Seegmiller (2025) **Table IA.3**.
- **Step 3 - Aggregate to three statistics used in Table 1**
- From these we obtain three statistics: (i) average elasticity, (ii) elasticity gap between high and low firms, (iii) elasticity gap between high and low workers

(i) **Average elasticity** - Using employment in each cell, we compute the employment n_{q_w,q_f}-weighted average of ϵ_{q_w,q_f}.

(ii) **Worker gap** - Using employment in each cell, we compute the employment weighted average of ϵ_{q_w,q_f} for high skill workers using the top quintiles $(q_w = 3)$ and for low skill workers using the bottom two quintiles $(q_w = 1)$. We then compute the difference.

(iii) **Firm gap** - Using employment in each cell, we compute the employment weighted average of ϵ_{q_w,q_f} for low productivity firms in the bottom two quartiles $q_f \in \{1, 2\}$ and for high productivity firms using the top two quartiles $q_f \in \{3, 4\}$. We then compute the difference.

- This same procedure is used to compress the $Q_w \times Q_f = 12$ statistics in Seegmiller (2025) Table IA.3 into three comparable statistics.

We choose the value of $\chi = \exp\{0.04\}$ given the increase in average earnings of 4 percent o deliver an average increase in earnings of 4 percent Seegmiller (2025) (Figure 1).

C.3. Conditional downward job moves

Our validation statistic is the change, following a windfall, in the probability that a job-to-job move is downward (i.e., to a lower-paying firm) conditional on moving. This conditional downward move probability is computed as the downward share of all moves (down over down-plus-up). Golosov et al. (2024) report the sample average probabilities of any, downward and upward moves. They also compute the post-win difference-in-differences estimates of the effect of winnings on the probability of any, downward, and upward moves, at horizons one, two, three, four and five years after the lottery win. If we had these statistics, we could compute the increase in the conditional downward-move probability at a five year horizon, which is appropriate for our static model which effectively compares across steady-states. However, Golosov et al. (2024) only released and report the average of these effects over the 5 years post-win. Using these we compute that, on average over these periods, the downward-move probability rises from 0.49 to 0.54, an increase of 4.6 percentage points. Using the event-study profile of Golosov et al. (2024) (their Figure B.11(d)), the $t = +5$ conditional downward-move response is a factor of 1.51 larger than the pooled, horizon-averaged response (2.7 versus 1.8 percentage points). We therefore rescale the pooled estimates by this factor, which delivers the moments we compare to the model: 4.6×1.51 is 6.9 from Table 2.