MaCurdy meets McFadden

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

February 2025

ABSTRACT

This paper extends the neoclassical theory of individual consumption and labor supply to allow workers to choose the firm they work at, in addition to hours worked. Firms internalize the relatively inelastic response of hours (the intensive margin) and the relatively elastic response of firm choice (the extensive margin), understanding that poor workers are willing to change employers for modest wage differentials, while relatively rich workers are not. Parameters of the theory are identified by, and rationalize, a core set of empirical measurements: (i) labor supply elasticities to firms by worker wage and firm pay, (ii) wage decompositions into firm, worker and covariance terms, and (iii) marginal propensities to earn after shocks to non-labor income. After estimating the model, we find that more than half of markdowns are variable. Across firms, the vast majority of markdown differences are driven by differences in extensive margin elasticities. Larger, more productive firms have a relatively richer, less elastic labor force: workers' wages are high, but a smaller share of their marginal product. These variable markdowns compress firm wage premia, so firms matter greatly for wage setting and inequality despite negligible firm premia in wage decompositions. Lastly, the forces that drive variable markdowns have first-order implications for tax policy. High extensive margin elasticities amplify losses from tax progressivity, reconciling high elasticities of taxable income with low intensive margin elasticities of labor supply.

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 differ sharply in how responsive they are to wage differences across employers. Conditional on skill, workers at higher-wage firms are less responsive to firm-specific wage changes (Seegmiller, 2021; Bils et al., 2025a), and larger, higher-paying firms pay wider markdowns (Yeh et al., 2022; Chan et al., 2023). At the same time, labor earnings respond substantially to shocks to non-labor income, with especially large responses among higher-income workers (Golosov et al., 2024; Dinerstein et al., 2025; Vivalt et al., 2024). Existing theories treat these facts separately: models of labor market power take heterogeneity in labor supply to firms as reduced form, while models of labor supply and consumption abstract from firms. This paper links the two. We show that workers' marginal value of wealth shapes labor supply not only on the usual hours margin, but also across employers. Diminishing marginal utility means that wealthy workers place less weight on wage differences across firms relative to idiosyncratic non-wage amenities, making them less responsive to employer-specific pay. When labor markets deviate from perfect competition, firms internalize these differences in elasticities, so that workers' economic circumstances affect wage-setting, markdowns, and sorting.

We formalize this insight and show that it reshapes how we think about wage-setting, inequality, and tax policy. We develop a model that combines a standard consumption-labor supply framework (MaCurdy, 1981) with discrete choice across employers (McFadden, 1974; Card et al., 2018). On the supply side, workers choose hours and consumption along with a discrete choice over which firm to work for; they differ in skill, non-labor income, and idiosyncratic preferences for jobs. On the demand side, a finite number of firms compete strategically for workers, and production features complementarities between worker skill and firm productivity (Costinot and Vogel, 2010). The across-firm margin of labor supply is absent in competitive models, where it is effectively infinite, and is typically parametric in monopsony models. Once it is finite and varies with workers' marginal value of wealth, worker heterogeneity becomes a central determinant of firm wage-setting. Combined with complementarities in production, the model links worker wealth, firm pay, sorting, and wage inequality within a unified framework. Crucially, we add no free parameters beyond those standard in either the neoclassical labor supply model or the discrete-choice model of labor supply across firms.

The model admits a closed-form expression for markdowns as a function of an intensive margin through hours and an extensive margin through workers' choice of employer. The strength of the extensive-margin channel is governed by risk aversion: when income effects dominate substitution effects—the standard case used in the consumption-inequality literature and as implied by large negative marginal propensities to earn—wealthier workers are less elastic to wage differences across firms. This same parameter has been at the center of the macro–public-finance literature on optimal taxation and consumption insurance (Heathcote et al., 2014, 2017).

However, its implications for firm wage-setting have gone unrecognized because competitive models render the across-firm margin of labor supply irrelevant. Our contribution is to endogenize this heterogeneity from standard preferences: the marginal value of wealth, the object that the consumption-inequality literature has long disciplined, turns out to be a key determinant of labor market power.

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, 2021; Bils et al., 2025b; Volpe, 2024). 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, 2021; Bils et al., 2025b; Yeh et al., 2022). Through the lens of our framework, the first fact points to strong 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. In the absence of worker-firm complementarities, high earners "follow" their preference shocks toward low wage firms, a counterfactual implication. Thus matching the positive sorting in the data requires complementarities in production that raise the wages of high-skill workers relatively more at productive firms. Conditional on complementarities and preferences, any remaining difference in labor supply elasticities across firm size is attributed to oligopsony. Remaining parameters include the Frisch elasticity and dispersion of worker ability and firm productivity. These are estimated to match earnings declines following lotteries (Golosov et al., 2021), 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 model matches these moments under standard values of preference parameters and reproduces key untargeted patterns in the data. In particular, it reconciles the empirical finding that, within skill groups, workers at higher-wage firms are less wage-elastic and are paid less competitively. We view this joint rationalization as a contribution in its own right: McFadden's discrete choice model and MaCurdy's neoclassical labor supply model fit together without producing counterfactual implications. More broadly, the model implies that an important share of wage markdowns is endogenous to the composition of firms' workforces: firms employing workers with lower marginal values of wealth face less elastic labor supply and therefore set wages further below marginal product. This calibrated model yields three main sets of results.

First, we decompose the sources of firms' market power. Firms pay workers, on average, 36 percent below marginal product. Roughly half of this is "variable," tied to worker and firm characteristics and would respond to changes in policy or economic conditions, while the remainder reflects standard monopsony forces that are invariant to workforce composition. Among the variable components, workers' wealth effects are the dominant source, accounting for about 30

percent of total markdowns. Oligopsony, or size-based market power of the kind emphasized in Berger et al. (2022), accounts for about 10 percent. Asset-based sorting and tax progressivity each contribute about 5 percent. This decomposition stands in contrast to existing frameworks that attribute variation in labor market power primarily to firm size and market concentration.

Second, these forces have direct implications for the interpretation reduced form decompositions of wage inequality. The most productive firms pay the widest markdowns, compressing the distribution of wages relative to marginal products. As a result, firm fixed effects explain only a modest share of wage variance, consistent with (Bonhomme et al., 2019; Kline et al., 2020; Borovičková and Shimer, 2024; Lamadon et al., 2022). This finding is sometimes read as evidence that firms play a limited role in wage determination. Our model delivers the opposite conclusion. Counterfactually removing the variable components of markdowns (oligopsony, asset-based sorting, and wealth effects) raises the contribution of firm fixed effects to wage variance by about 5 percentage points, or roughly 40 percent. Firms are central to wage determination; they appear otherwise because their wage-setting behavior compresses the very variation that would reveal their importance.

Third, we apply the framework to progressive income taxation. Theoretically, we show that (i) higher tax progressivity makes labor less elastic with respect to the pre-tax wages chosen by firms, (ii) this results in wider markdowns, (iii) which generate aggregate effects that mimic a contraction in labor demand, rather than the usual contraction in labor supply. Quantitatively, increasing progressivity reduces output, but less than half of this decline reflects reduced hours. The dominant channel is a worsening allocation of workers across firms. In an exact decomposition, reduced sorting accounts for more than the entire decline in aggregate labor productivity, while heterogeneous responses in hours worked by low and high productivity workers accounts positively. This extensive-margin channel is entirely absent from standard models of optimal taxation. The framework also delivers an elasticity of taxable income of 0.24, within the empirical range of Saez et al. (2012), but through sorting distortions rather than reduced hours, providing a new account of why empirical estimates of earnings responses to taxes exceed model-based predictions. More broadly, the framework applies to any policy that significantly shifts the marginal value of consumption for large groups of workers, including universal basic income, unemployment insurance, and minimum wages, since these policies alter the across-firm margin of labor supply and therefore reshape firm wage-setting and sorting across firms.

Overview. We next review the relevant literature, and then lay out the baseline model in Section 2. Section 3 characterizes theoretical properties of the model in terms of labor supply elasticities to firms and sorting. Section 4 calibrates the model to statistics on heterogeneity in worker and firm fixed effects and their covariance as well as data on markdowns by firm and worker type. Section 5 applies the model to study the sources of markdowns in general, their

differences across firms and implications for wages and inequality. Section 6 applies the model to study the implications of tax progressivity on wages, sorting and aggregate productivity.

1.1. Literature

This paper sits at the intersection of two literatures: (i) the treatment of inequality in macroeconomics 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 influential 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. They build on Constantinides and Duffie (1996) by incorporating permanent and transitory idiosyncratic income risk into a Bewley model with endogenous hours, a la MaCurdy (1981). Their welfare and labor supply formulas critically depend on the tax-adjusted labor supply elasticity which is also at the heart of our monopsony mechanism (i.e., tax progressivity lowers labor supply elasticities and aggravates monopsony power). We contribute to the literature by incorporating non-competitive wage setting, firm heterogeneity and sorting.

An equally influential set of papers focuses on optimal progressivity in Mirrleesian matching frameworks, Scheuer and Werning (2017). Scheuer and Werning (2017) assume that worker and firm types are supermodular and show that this super modularity makes earnings elasticities greater among higher-type, higher-wage workers. Lower tax progressivity induces workers to reallocate to better matches, in which workers exert more effort, amplifying the gains.

Another strand of literature studies taxation 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 two sided heterogeneity ex-ante, heterogeneous marginal utilities of wealth, and endogenous labor supply as in MaCurdy (1981).

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,

2024). Despite rapid progress in modeling and measurement of monopsony power, 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 the neoclassical treatment (MaCurdy, 1981). We show that progressive taxation reduces hours supply through two channels, (i) the direct effect on choice of hours, holding wages fixed, and (ii) the indirect effect of lower wages due to lower labor supply elasticities and wider markdowns.

Our paper contributes to the ongoing debate over the nature of firm premia in U.S. (Bonhomme et al., 2019; Borovičková and Shimer, 2024; Kline, 2024). Using clustering methods, Bonhomme et al. (2019) argue that firm premia explain a trivial 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. Kline (2024) argue that the clustering methods dampen the rich firm heterogeneity present in the data. Our contribution is to show that a neoclassical model of labor supply with monopsony power is consistent with low firm wage premia and strong complementarities of production. Our model generates these patterns through two channels: (1) workers select into jobs based on unobserved asset positions and marginal utilities of wealth (this yields sorting), and (2) high-wage firms have more monopsony power (limiting firm wage premia). The highest paying firms have the least wage elastic workers, who are then paid large markdowns, which in turn compress the role of firm fixed effects in explaining the variance of wages. In a counterfactual economy with competitive firms, the role of firm fixed effects sharply increases.

Our model relates to work by Volpe (2024). Volpe (2024) extends Lamadon et al. (2022) to include 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 (2021) shows that, within an education group, higher paying firms face *less wage-elastic employment*. This is consistent with findings in Brazil (Bils et al., 2025a). 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).

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. We assume there is a finite number J_m of firms in each market, and a large number of markets $m \in \{1, \ldots, M\}$.[1] Firms have heterogeneous productivity levels z_{jm}. A measure $\overline{\Gamma}$ of workers is indexed by i, and they 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.

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 flow 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.

These preferences nest three specifications commonly adopted in the literature. First, absent firm specific shifters, these are simply MaCurdy (1981) preferences. Under $\sigma = 1$, preferences are balanced growth. Under $\sigma > 1$, wealth effects on labor supply dominate substitution effects. These are the same preferences deployed in papers that evaluate optimal tax progressivity in the incomplete markets literature in public/macroeconomics (e.g. Heathcote et al., 2017), which use log preferences in their benchmark exercises.

Second, under log utility, and absent an hours component, utility is as in Lamadon et al. (2022). If there are infinitely many firms in each market, or $\eta = \theta$ such that within- and across-market dispersion in tastes are the same, then the competitive structure also lines up with the monopsonistically competitive environment in that paper.

Third, under log utility and Gumbel distributed tastes, we can apply the aggregation results from Verboven (1996) as in Berger et al. (2022) to obtain nested-CES utility. Hence, again, the departure from log utility is key. We show below that this departure is necessary for the model to match reduced form estimates of differences in labor supply elasticities to firms of low and high wage workers.

[1] We assume that there is an arbitrarily large number of markets, making each firm have no influence on the aggregates.

Technology. An allocation can be expressed as 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 of labor, output of firm jm is given by:

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

The measure of workers of type e is $\Gamma(e)$. Of these a share $\rho_{jm}(e)$ work at firm jm, and they work $h_{jm}(e)$ hours. They produce $A(z_{jm}, e)$ units of output per hour worked (Costinot and Vogel, 2010), which depends on worker skill e.

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. To economize on notation, 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:

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

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 for hours and consumption:

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

2.3. Firm problem

We assume that firms set wages that discriminate by worker skill e. We will maintain this assumption when we introduce non-labor income, effectively assuming non-labor income is unobserved by the firm. In our results section we tighten this assumption to allowing only a firm wage w_{jm}.

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:

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

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 form of 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), or do not behave strategically with respect to their market (i.e. monopsonistic competition).

3. Characterization of equilibrium - 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. Worker 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:

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

Second, we log-differentiate to obtain the extensive margin elasticity:

$$\varepsilon^{\rho,i}_{jm} = \frac{\partial \log \rho^i_{jm}}{\partial \log w^i_{jm}} = \frac{\partial \log \rho^i_{jm}}{\partial \log v^i_{jm}} \times \frac{\partial \log v^i_m}{\partial \log w^i_{jm}} = \left[\eta(1-\rho^i_{j|m}) + \theta\rho^i_{j|m}\right] \times \left(\lambda^i_{jm} w^i_{jm} h^i_{jm}\right). \tag{3}$$

The extensive margin elasticity has two components. The first, says how sensitive an individuals' choice is to the value represented by a firm in the market, v^i_{jm}. This captures traditional *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 $\widetilde{v}^i_{jm}$, and hence the elasticity nears η. This could also occur 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 toward zero. When the firm is *large* in the 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 toward one and the first term approaching the across market parameter θ.

The second, reflects 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. Hence, *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. Hence, *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 $y^i_{jm} = w^i_{jm}h^i_{jm}$:

$$\varepsilon^{\rho,i}_{jm} = \left[\eta(1-\rho^i_{j|m}) + \theta\rho^i_{j|m}\right] \times \left(y^i_{jm}\right)^{-(\sigma-1)}. \tag{4}$$

If $\sigma > 1$, then, 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 given percentage wage increase at the high wage firm is therefore less likely to attract the worker than the same percentage increase at the low wage firm. This is encoded as a lower elasticity of labor supply to the high wage firm on the extensive margin. 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 relatively less value on working at higher wage firms, just like they place relatively less value on working longer hours or working at all.

Equation (4) also delivers a transparent approach to calibrating the model. The parameter σ governs how elasticities of labor supply across firms vary across workers.

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:

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

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

Empirical estimates of Frisch and Marshallian labor supply elasticities 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 wages were set with only this elasticity in mind, then an *upper bound* on markdowns would be one-half: $\varphi/(1+\varphi)$. Estimates of Marshallian elasticities tend to be below the Frisch, further compressing this upper bound.

3.2. Sorting

The combination of heterogeneity in extensive labor supply elasticities and complementarity in production yields 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):

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

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

A useful first 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:

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

We can now apply this to sorting of workers across firms, for which we want to compute Υ_ρ, the super-modularity associated with the extensive margin of labor supply. Using the worker 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:

$$\frac{\Upsilon_\rho}{\eta} = \underbrace{\Upsilon_A u'\Big(c_2^L\Big)c_2^L}_{\text{Technological complementarity}} + \underbrace{-\Big(\sigma-1\Big)u'\Big(y_1^L\Big)\Big(y_1^L - y_2^L\Big)\Delta\log y_1}_{\text{Wealth effects}} \quad , \quad \Delta\log y_1 = \frac{y_1^H - y_2^H}{y_2^H}. \tag{5}$$

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, then there will also be increasing differences in firm employment, by worker type. Hence this term is *positive*. The second piece reflects the heterogeneity in labor supply elasticities, holding fixed the complementarities in production. High wage workers have higher consumption, and if $\sigma > 1$, this reduces their marginal value of working at high wage firms. Their labor supply is less responsive to wages at high wage firms, and hence, in relative terms they will sort toward low wage firms. Hence this term is *negative*. Empirically, if there is positive sorting, then it is clear that the first force must dominate. Hence, conditional on labor supply elasticity differences across worker types, data on sorting is determinative of production complementarities.

3.3. Identification

These results will give a particular interpretation to data on sorting and labor supply elasticity differences across workers and firms. While the moments from the data will be coarser than

what follows, 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 strong wealth effects, and thus to $\sigma > 1$. This is because, in equation (4), the extensive margin elasticity declines with income only when wealth effects are sufficiently strong. However, if high earners are less elastic, our sorting expression in equation (5) implies that they tend to sort toward low wage firms. This requires complementarities in production ($\Upsilon_A > 0$) that raise the wages of high earners relatively more at productive firms. This mitigates high earners' tendency to ignore wages and follow idiosyncratic tastes, which is stronger for them than for low earners if $\sigma > 1$. Hence, once the model matches the first fact, the second necessarily implies complementarities in production. The larger the elasticity differences across workers, the larger complementarities will have to be to generate positive sorting. Any remaining elasticity difference between low and high productivity firms is driven by the *size-based* piece of the extensive margin elasticity (i.e., the value share terms ρ^i_{jm} in equation (4)). This reveals how much less substitutable markets are than firms from the worker's perspective, and therefore helps identify the gap between θ and η.

3.4. Comparison to Volpe (2024)

As an application of the above, and identification arguments, we contrast sorting via wealth effects and complementarities to sorting via amenities in Volpe (2024). Labor is only supplied on the extensive margin across firms and utility is:

$$U^i_j = \beta^i \log w_j(e^i) + \frac{1}{\eta} \log \phi_j(e^i) + \zeta^i_j.$$

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 above, labor demand and firms' elasticity are:

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

The individual elasticity is $\varepsilon^i_j = \eta\beta^i$, 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.

$$\Upsilon_\rho(e) = \log\left(\frac{\rho_1(e,\beta^H)/\rho_1(e,\beta^L)}{\rho_2(e,\beta^H)/\rho_2(e,\beta^L)}\right) = \eta\Big(\beta^H - \beta^L\Big)\Big(\log w_1(e) - \log w_2(e)\Big). \tag{6}$$

This is always *positive*. High-β types sort to high wage firms, implying a higher ε_j at these firms and narrower markdowns.

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

3.5. Summary

In this section we have shown how households' labor supply elasticities on the extensive margin across firms are shaped by various forces. These include wealth effects on labor supply, and the density of firms that are desirable from the workers' perspective which determines the market power those firms have over the worker. Together, these imply that counterfactual policies that shift the distribution of income and wealth will 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

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

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 $\tau_1 = 0$, there is a constant proportional tax at rate τ_0. If $\tau_1 > 0$, then taxes are *progressive*, with marginal tax rates increasing in income. This follows Benabou (2002) and Heathcote et al. (2020a), who find the 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 (i) isolate all working age individuals (20 to 55 years old) who earn at least \$1,000 (where all nominal variables are deflated to 2014), (ii) compute net worth, (iii) equivalize by the number of adults in the household, (iv) compute 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), (v) split households into 20 quantiles of income, (vii) compute the share of households in each quantile of income that have a^i in each of four bins: $(-\infty, 0)$, $[0, \$5,000)$, $[\$5,000, \$20,000)$ and $[\$20,000, \infty)$.[2] This gives the conditional density of $\Gamma(a^i|e^i)$, where we set $a^i \in \{\$0,\ \$2,500,\ \$12,500,\ \$20,000\}$ in the model. As examples, 28 (2) percent of lowest income households are in the lowest (*highest*) bin, while 2 (*58*) percent of the highest income households are 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 $w_{jm}(e)$ 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:

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

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 also change:

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

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 formulas deliver three important observations. First, higher tax progressivity reduces both elasticities. The labor supply elasticity is with respect to the *pre-tax wage*, which is what matters to firms' decision making. However, on the margin, a \$1 wage increase from the firm raises the after-tax income—which is what matters to workers' decision making—by progres-

[2] These four buckets represent 12%, 62%, 19% and 7% of households, respectively.

sively less when τ_1 is higher. Hence, *ceteris paribus*, higher tax progressivity dissuades workers from leaving one firm to join another and from supplying more hours when that firm raises its pre-tax wage. Second, additional non-labor income shifts elasticities in different directions. The extensive margin elasticity falls as the marginal value of higher wages at one firm versus another are lower, requiring larger wage changes to shift labor supply. The intensive margin elasticity rises as consumption increases less when income increases, dampening wealth effects that otherwise lower the Marshallian labor supply elasticity. Third, firms incorporate these effects into their prices of labor. If most high e households have high a, then $w_{jm}(e)$ will be set low, reflecting less elastic supply. This wage will be lower than if the firm could otherwise tailor wages to the low a, more elastic, households within this group. High a households benefit from firms lack of ability to set wages by a, while low a households lose.

4.2. Approach

We proceed to estimate the model by Simulated Method of Moments (SMM). All parameters of the model are estimated via 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. The key 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), and they compute an average MPE of -0.52 (Column (1) of Table III). We implement this experiment in our model by increasing a^i by \$7,500 and computing the change in labor earnings, holding equilibrium wages fixed.

The 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., 2025b), Norway (Volpe, 2024), and the US (Seegmiller, 2021). In all cases firms' *reduced form labor supply elasticities*—computed as the change in log employment divided by the change in log average wage—are higher for low income workers 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, 2024) 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 (2021) uses variation in firm-specific stock returns as instruments for the marginal revenue product of labor. Seegmiller (2021) splits workers into quintiles of skill based on AKM (Abowd et al., 1999) fixed effects, with the bottom two labeled low-skilled, and the top labeled high-skilled. Seegmiller (2021) estimates an average labor supply elasticity of 5.34 (which implies an average markdown of 0.80); an average gap in labor supply elasticities between high- and low-skilled workers of 5.69; and an average gap in labor supply elasticities between low and high value-added firms of 1.69. To compute the model counterparts 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.[3]

The number of firms per market J is chosen to match the payroll-weighted, wage-bill Herfindahl.[4] 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 productivity (e.g., Marimon and Zilibotti (1999) and Lise and Postel-Vinay (2015)). We specify the following functional form for $A(z_j, e^i)$, which 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 dependent on worker or firm type, but not both:

[3] This procedure follows Berger et al. (2022) which points out the issues associated with mapping reduced form labor supply elasticities to structural objects like ε_{jm} in concentrated markets.

[4] 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 $HHI_m = \sum_j s_{jm}^2$.

$$\log w_{jm}(e^i) = \log \mu + \log z_j + \log e^i.$$

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.[5] 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 (i.e., $var(\log c) = 0.27$ in CEX data tabulated in Heathcote et al. (2020b)). 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.[6] Note that the largest share of the variance is due to worker fixed effects (0.61), while the remaining 0.10 is error.[7] These shares are similar to those estimated by Bonhomme et al. (2019) in Sweden (77% worker, 5% firm, 17% covariance).

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%.[8] We set τ_0 to match the ratio of labor income taxes to total payroll of 29 percent.[9]

[5] 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.

[6] 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).

[7] 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 former 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.

[8] 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 ($\$2,251/\$50,436 = 4.46\%$).

[9] Labor income taxes to GDP are 19 percent in the U.S., to get their share of payroll we divide by a labor share of 0.65.

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	0.29
Productivity shifter	$\overline{Z}$	43.65	Average earnings	\$45,207	\$45,103
Labor disutility shifter	$\overline{\varphi}$	12,072	Average hours	1,777	1,782
Worker prod. dispersion	σ_e	0.72	Variance of log consumption	0.27	0.27
Firm prod. dispersion	σ_z	0.52	Share of log wage variance - Firm effects	0.13	0.13
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	0.66
Within-market dispersion	η	2.59×10^7	Average markdown*	0.80	0.78
Across-market dispersion	θ	$0.01 \times \eta$	Elasticity gap: Low - High wage firms*	1.69	1.72
Relative risk aversion	σ	2.49	Elasticity gap: Low - High wage workers*	5.69	5.87
Frisch elasticity	φ	0.08	MPE after \$7,500 windfall gain	-0.52	-0.53
Number of firms	J	16	HHI - Payroll	0.11	0.12

Table 1: Baseline calibration

Note: (i) $*$ denotes from (Seegmiller, 2021), (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.

4.3. Results - Estimation

We have four main results from the calibration itself which we connect to our discussion in Section 3.3. Table 1 provides parameter estimates and moments in model and data.

First, moderate wealth effects rationalize the gap between reduced form estimates of labor supply elasticities to firms of low versus high wage workers ($\sigma = 2.49$). We view the fact that this falls within ranges of standard estimates as a significant outcome for the model.

Second, there are notable complementarities ($\psi = 0.09$). With $\sigma > 1$, high wage workers have a lower marginal value of the high wages of high z firms, 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.[10]

Third, it is not simply the case that high wage firms are productive because they are a collections of high wage people ($\sigma_z = 0.52$). Despite a small share of variation in wages being accounted for by firm fixed effects, the heterogeneity in firm productivity is nonetheless large and in the range of estimates from administrative firm level data (Decker et al., 2013). We therefore share the conclusion of Borovičková and Shimer (2024) that firm-wage-variance decompositions can

[10]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 1.7 times relatively more 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 increases to 9.3.

	A. Data - Seegmiller (2023)		B. Model	
	Low z	High z	Low z	High z
Low skill workers	8.44	5.72	8.91	6.27
High skill workers	1.72	1.07	2.13	1.32

Table 2: Model vs. Data - Markdowns across workers skills and firm value-added quartiles

	All	Earnings quartiles		
		Q1 (low)	Q2-Q3	Q4 (high)
Data (Golosov et al., 2024)	-0.52	-0.31	-0.55	-0.67
Model	-0.53	-0.26	-0.52	-0.81

Table 3: Model vs. Data - Marginal propensity to earn

not be used to make prima-facie conclusions about sources of heterogeneity in worker and firm productivity.

Fourth, significant size-based market power is required ($\theta = 0.01 \times \eta$). Since $\sigma > 1$, workers are less elastic at high wage firms, as earning a high wage reduces the marginal value of wealth. Despite this, a value of $\theta < \eta$ is still required to further reduce the labor supply elasticity to large, productive firms in the market, consistent with the data.

More broadly, this provides our first main quantitative result of the paper. 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.

Note on scaling. 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 becomes: $\varepsilon^{\rho,i}_{jm} = [\eta(1-\rho^i_{j|m}) + \theta\rho^i_{j|m}] \times (y^i_{jm})^{-(\sigma-1)}$. Average per worker earnings are \$45,207. With $\sigma = 2.49$ the second term is extremely small, $(\widetilde{y}^i_{jm})^{-(\sigma-1)} \approx 1.6e^{-7}$. 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 to offset the downward pressure working through the utility scaling.

Additional moments. Table 2 provides the reduced form labor supply elasticity estimates for each cell. The model compares well to the data. Note that the model is sufficiently flexible to generate opposite patterns. For example, if large firms had more elastic labor supply—which is the empirical finding of Volpe (2024) in Norway—we would infer $\sigma < 1$. Table 3 compares MPE estimates across income groups. As in the data, the MPE is larger for high income households.

We view the model's success at replicating these non-targeted moments as additional validation of the model's wealth effect mechanisms.

Wages by types and sorting. Despite the presence of complementarities, wages in the model do not 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'. For a middle worker type, their wages might peak at a middle firm type. This is not the case. Figure 1A plots average log earnings of each worker type at each firm. For clarity we group firms and workers by quartiles of z and e. Our figure replicates the key Figure 2 in Bonhomme et al. (2019). For all workers, earnings are higher at higher productivity firms. Recall that we chose the structure $A(z,e) = ze \times \text{Penalty}_{\psi}(z,e)$, such that if the Penalty term is 1, we maintain log additivity in the marginal product of labor. Figure 1A shows that this dominates in shaping wages, which is emphasized by the dashed lines which plot $A(z,e)$ directly. The Penalty term exerts sufficient complementarities to undo the negative sorting via wealth effects and deliver positive sorting in equilibrium.

Figure 1B shows how sorting plays out. The Penalty function delivers the key 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 (Borovickova and Shimer, 2024) or sorting on amenities (Lamadon et al, 2025; Volpe 2025). These features co-exist in our model with markdown differences across firm and worker types that are consistent with empirical estimates.

Figure 1C adds to the suite of plots considered by Bonhomme et al. (2019) by turning to workers' perspectives. A key part of our theory is a definition of the firm that can be taken to the data. A cursory look at Panel B might leave the impression that for low skill workers, a larger share of them are employed at low productivity firms. This is not the case. High productivity firms are larger. For higher productivity firms to be larger, they must employ more of all types of workers. This has implications for wage setting, as high productivity firms have a high ρ^i_{jm} in the extensive margin elasticity for all worker types, lowering their elasticity of supply via size-based forces.

4.4. Comparative statics of sorting

To further illustrate key properties of the model, we consider two simple counterfactuals and their implications for sorting.[11]

No complementarities. Figure 2A describes sorting under $\psi = 0$. Absent production complementarities, wealth effects dominate sorting. With $\sigma > 1$, high productivity workers—who

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

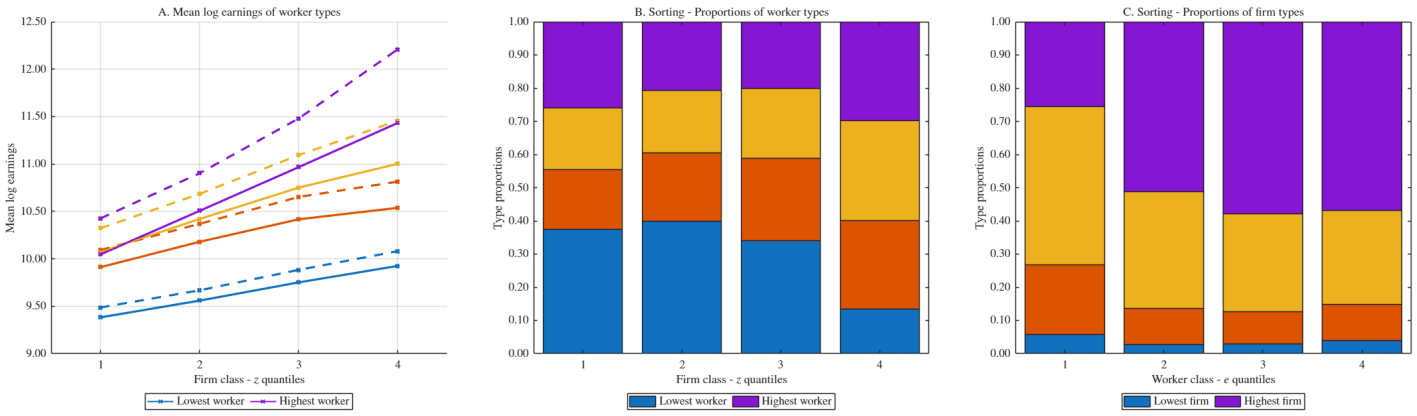


Figure 1: Wages and Sorting

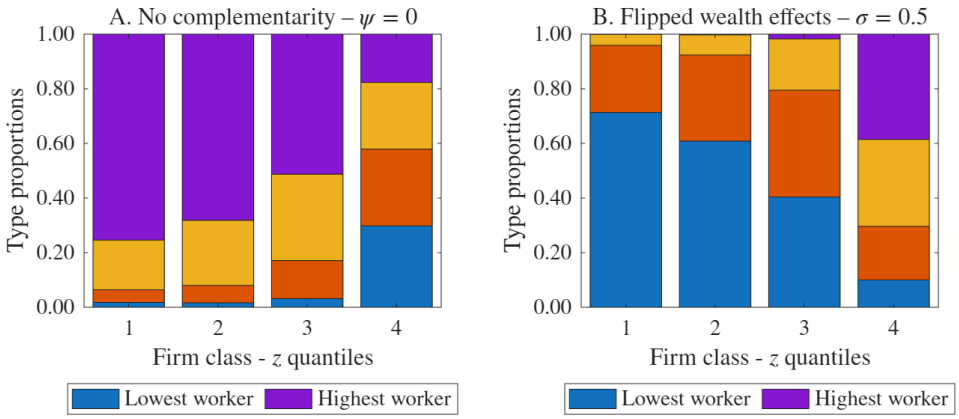


Figure 2: Sorting - Counterfactual patterns of sorting under $\psi = 0$ and $\sigma = 0.5$

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 2B 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.

Taking stock. We have shown that 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. Given these quantitative successes of the model, we turn to two applications. First, we use the model to quantify the sources of depar-

tures from competitive wages in general and across firms. In particular, we quantify the role of size based notions of market power and heterogeneity in worker elasticities and sorting. Second, motivated by the presence of tax progressivity in expressions for labor supply elasticities $\varepsilon_{jm}^{i,\rho}$ and $\varepsilon_{jm}^{i,h}$, we use the model to understand the implication of changes in tax progressivity to the structure of wages, sorting and production.

5. Application I - Sources of markdowns

5.1. Employment, Wages and Markdowns

First, we provide some background in the form of summary statistics from the cross-section of firms and workers. Figure 3 compares firm employment, hours and markdowns by worker types. We maintain our previous grouping of firms and workers into four quartiles of z and e. As expected, high productivity firms have higher employment, and with a higher marginal product of labor pay higher average hourly wages. However, Panel C shows that these wages are a larger markdown on worker marginal revenue productivity than at small firms. Next we describe how different model features contribute to these declining markdowns. For now, we focus on Panel C, which shows that markdowns are declining for all worker types, and more steeply declining for the highest productivity workers.

Our theory revealed two reasons why all workers' wages are a smaller slice of their marginal product at high productivity firms. First, within a worker-type, working at a high wage firm increases the worker's level of consumption, making them less elastic with respect to wage differences across firms. Second, since high productivity firms are larger in the labor market for all types of labor (Figure 1C), their dominance leads to wider markdowns for 'size-based' reasons.

These forces also shape the more pronounced markdowns for high e workers at high e firms. First, at high productivity firms, the complementarity between worker and firm types leads to relatively higher wages for the high wage type, leading them to be relatively less elastic and hence marked down even more than the low wage types. This implies that a firm's ability to discriminate on wages across worker types leads lower skill workers to have higher pay, and higher skill workers to have lower pay. We quantify these effects below. Second, this complementarity also leads high productivity firms to be especially large in the market for high skill labor (Figure 1C). This position of dominance in the market leads to even larger size-based reasons for wide markdowns on high skill labor at high productivity firms. We next answer which of these forces is bigger, and for which workers.

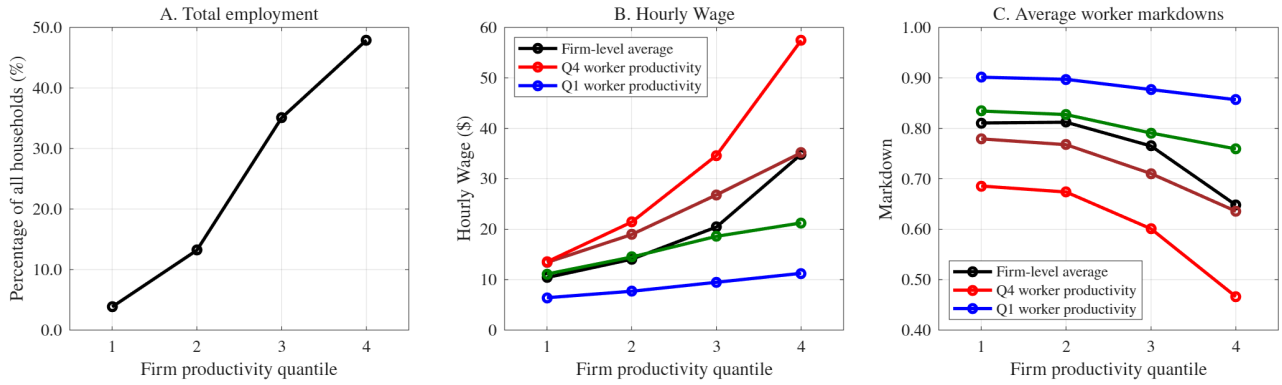


Figure 3: Employment, Wages and Markdowns

5.2. Understanding differences in labor supply elasticities across firms

The sorting patterns in the previous figures are driven by heterogeneity in labor supply elasticities and production complementarities, which sort workers with different elasticities into different firms in equilibrium. Figure 4 provides a decomposition of labor supply elasticities across firms and workers in order to unpack the different forces in our expression for the extensive margin elasticity $\varepsilon^{\rho,i}_{jm}$ in equation (3). We use the following decomposition, dropping the additional m subscript for clarity, and using i to capture heterogeneity in a^i and e^i. We start with a weighted average of elasticities across types within the firm:

$$\varepsilon_j := \sum_i \omega_j^i \underbrace{\left[\eta \rho_j^i + \theta\left(1-\rho_j^i\right)\right]}_{\text{A. Oligopsony}} \underbrace{u'\left(c_j^i\right)\widetilde{y}_j^i}_{\text{B. Consumption}} \quad , \quad \omega_j^i = \frac{\rho_j^i h_j^i \Gamma(a^i, e^i)}{\sum_k \rho_j^k h_j^k \Gamma(a^k, e^k)}$$

We then consider the following approximate decomposition:

$$\widetilde{\varepsilon}_j := \varepsilon_j^O \varepsilon_j^C \quad , \quad \widetilde{\varepsilon}_j^O := \sum_i \omega_j^i \underbrace{\left[\eta \rho_j^i + \theta(1-\rho_j^i)\right]}_{\text{A. Oligopsony}} \quad , \quad \widetilde{\varepsilon}_j^C := \sum_i \omega_j^i \underbrace{u'\left(c_j^i\right)\widetilde{y}_j^i}_{\text{B. Consumption}} .$$

We rank firms by their employment, with $j = 1$ being the smallest, and plot the decomposition in log deviations from the smallest firm.

Figure 4A shows that both the oligopsony and consumption components contribute to declining elasticities of labor supply by firm size. Large firms have wide markdowns because production complementarities lead them to employ more productive, less elastic labor (blue line). Large firms also have wide markdowns because for each labor type they tend to be large within the labor market. Overall, heterogeneity in the marginal value of wealth within-workers, across-firms, and how workers with different marginal values of wealth sort across firms, shapes the largest part of labor supply elasticity differences across firms. Panel B uses the same approach to decompose the consumption piece into the $V_a = u'(c)$ term (green) and $\widetilde{y}$

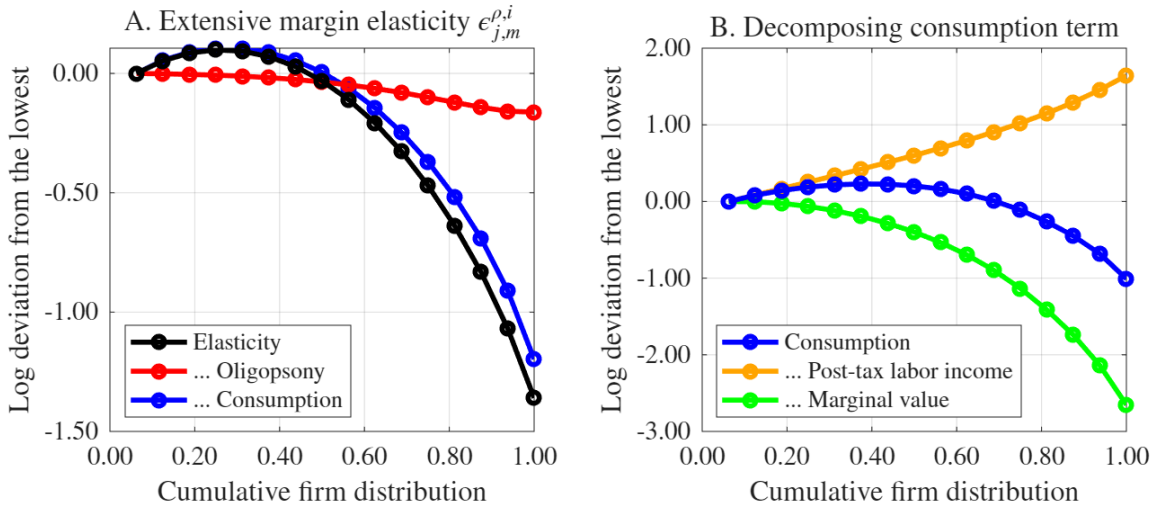


Figure 4: Decomposition of labor supply elasticities across firms

term (orange). The falling marginal value of wealth dominates.

5.3. Sources of markdowns

We now move on to what generates markdowns more broadly in the economy. We systematically isolate the sources of markdowns in our model. First, we define the payroll weighted average markdown μ as a weighted average of worker-specific markdowns $\mu_j(a,e)$:

$$E = \int_{A\times E} \sum_{j=1}^{J} y_j(a,e)\rho_j(a,e)\Gamma(a,e)\, da\, de$$

$$\mu := \frac{1}{E}\int_{A\times E} \sum_{j=1}^{J} y_j(a,e)\rho_j(a,e)\mu_j(a,e)\Gamma(a,e)\, da\, de, \quad \mu_j(a,e) := \frac{\varepsilon^{\rho}_j(a,e)+\varepsilon^{h}_j(a,e)}{1+\varepsilon^{\rho}_j(a,e)+\varepsilon^{h}_j(a,e)} \tag{8}$$

From the key equations (7) for $\varepsilon^{\rho}(a,e)$ and $\varepsilon^{h}(a,e)$, we isolate five sources of markdowns: (1) tax progressivity, (2) oligopsony, (3) asset-based sorting, (4) wealth effects, and (5) standard monopsony. We cumulatively remove these five sources in order to isolate their contribution to markdowns. Each maps into components of the extensive and intensive elasticities:

$$\varepsilon^{\rho}_{jm}(a,e) = \Big[\underbrace{\eta}_{\text{Monopsony}} + \underbrace{\big(\theta-\eta\big)\,\rho_{j|m}(a,e)}_{\text{Oligopsony}}\Big] \underbrace{u'\Big(c_{jm}(a,e)\Big)\widetilde{y}_{jm}(a,e)}_{\text{Asset sorting + Wealth effects}} \underbrace{\Big(1-\tau_1\Big)}_{\text{Tax progressivity}} \tag{9}$$

$$\varepsilon^h_{jm}(a,e) = \frac{\underbrace{\left(1-\sigma MPC_{jm}(a,e)\right)}_{\text{Asset sorting + Wealth effects}}\ \underbrace{\left(1-\tau_1\right)}_{\text{Tax progressivity}}}{\left(1+\frac{1}{\varphi}\right) - \underbrace{\left(1-\sigma MPC_{jm}(a,e)\right)}_{\text{Asset sorting + Wealth effects}}\ \underbrace{\left(1-\tau_1\right)}_{\text{Tax progressivity}}}. \tag{10}$$

Table 4 reports our results from this exercise. We start from an aggregate markdown of $(1-\mu)$ of 0.36, and remove each source, cumulatively. In each case we do so *ceteris paribus*. We (a) directly change terms in (9) and (10), (b) use these counterfactual expressions for the elasticities to obtain the associated markdowns at the worker-firm level, (c) aggregate them using our definition (8), weighting with the baseline allocation of employment.

First, we remove tax progressivity by setting $\tau_1 = 0$ in (9) and (10), *ceteris paribus*. The aggregate markdown falls to 0.35 (column 1), which accounts for 3.7 percent of the aggregate markdown (column 2).

Second, we isolate the role of oligopsony by computing markdowns under atomistic firm behavior. This removes the strategic component of markdowns by setting θ to the same value as η in (9), *ceteris paribus*. Removing oligopsony leads the aggregate markdown to fall from 0.35 to 0.32 (column 1), which accounts for 8.0 percent of the aggregate markdown (column 2) Column 3 reports that the cumulative removal of taxes and oligopsony accounts for 11 percent of markdowns. This is broadly consistent with the decomposition across firms in Figure 4.

To isolate the component of markdowns due to the sorting of workers based on assets and wealth effects requires some additional notation. Let $W_j(a,e) := u'(c_j(a,e))\widetilde{y}_j(a,e)$ summarize the wealth component of $\varepsilon^\rho_j(a,e)$. Let $\overline{W}(e)$ denote the employment weighted mean of $W_j(a,e)$, conditional on e, and $\overline{W}$ the overall employment weighted mean. We similarly construct means of $\overline{MPC}(e)$ and $\overline{MPC}$.

Third, we isolate the role of sorting on the basis of assets. We replace $W_j(a,e)$ and $MPC_j(a,e)$ in (9) and (10), respectively, with $\overline{W}(e)$ and $\overline{MPC}(e)$. Row 4 shows that this contributes 5.6 percent to the markdown, further reducing the aggregate markdown by 2 percentage points.

Fourth, we isolate the role of wealth effects. We now replace $\overline{W}(e)$ and $\overline{MPC}(e)$ with $\overline{W}$ and $\overline{MPC}$. Row 5 shows that pure wealth effects close markdowns from 0.30 to 0.18, accounting for 32 percent of markdowns. Cumulatively, the first four markdown factors account for 50 percent of markdowns.

The fifth and final component of markdowns is standard monopsony. We treat this component of markdowns as the residual. We achieve this by effectively sending η to infinity, which makes the extensive margin elasticity infinite. This has the additional effect of completely removing the role of the intensive margin in wage setting. By construction, this accounts for the remaining

Source removed (cumulatively)	Markdown $(1-\mu)$ (1)	Share (%) (2)	Cumulative Share (%) (3)
Baseline	0.36	—	—
No tax progressivity	0.35	3.7	3.7
No oligopsony	0.32	8.0	11.7
No asset-based sorting	0.30	5.6	17.2
No wealth effects	0.18	32.1	49.3
No monopsony	0.00	50.7	100

Table 4: 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.

50 percent of markdowns (Row 5).[12]

Summary: Half of markdowns are variable. Table 4 shows the wealth effect channel explains 30 percent of markdowns, while roughly 10 percent is attributable to oligopsony. Asset-based sorting and taxes each account for 5 percent or less of markdowns, while standard monopsony explains the remaining 50 percent of the aggregate markdown. These results imply that half of markdowns are "variable", driven by sources of both worker and firm heterogeneity.

5.4. Implications for wages

The model has important implications for the interpretation of empirical decompositions of wage inequality into firm, worker and covariance components. Such decompositions have become a key reference in the study of labor markets. Our calibrated model matched one such decomposition from Kline et al. (2020), with only 13 percent of the variance attributable to firm fixed effects. Such small estimates of the variance share attributed to firm fixed effects has been seen as evidence that firms are not central to wage setting.

Our model delivers the result that such conclusions require additional nuance regarding how firm wage setting *compresses* wages. A simple view of this is obtained from Figure 1A, which showed marginal products—which are equal to wages when firms do not set wages and labor markets are competitive—fanning out across firms and worker types while wages themselves remain compressed. We find that removing wage setting power from firms *increases* the role of firm fixed effects by 6 percentage points, a 40 percent increase over their baseline role. That is firms' price setting behavior itself leads to compression of wages and a smaller share of the wage variance associated with firm fixed effects.

[12]Setting $\eta = \infty$ would remove all markdowns stemming from the prior four markdown channels, and so η's contribution alone can arbitrarily range from 0 to 100%, depending on the ordering of the decomposition.

We follow an identical decomposition strategy as Section 5.3. Consecutively, we remove each source of markdowns from elasticities (9) and (10): (1) tax progressivity, (2) oligopsony, (3) asset-based sorting, (4) wealth effects, and (5) standard monopsony. From these we obtain new elasticities $\varepsilon_{jm}(e)$, which delivers new wages.

$$\varepsilon_{jm}(e) = \int \underbrace{\frac{\rho_{jm}(a,e)h_{jm}(a,e)}{\int \rho_{jm}(a',e)h_{jm}(a',e)d\Gamma(a'|e)}}_{\text{Fixed across decompositions}} \underbrace{\left[\varepsilon^{\rho}_{jm}(a,e) + \varepsilon^{h}_{jm}(a,e)\right]}_{\text{Varies across decompositions}} d\Gamma(a|e) \quad , \quad w_{jm}(e) = \underbrace{\frac{\varepsilon_{jm}(e)}{\varepsilon_{jm}(e)+1}}_{\text{Varies across decomp.}} A_{jm}(e).$$

When computing earnings, and aggregating within and across firms we use the baseline allocation of workers and hours.

Table 5 reports the results. Row 1 is our baseline economy. In Row 2, we remove tax progressivity. Overall wage inequality increases (Column 1), but there is little effect on the share of wage variance explained by firms. This is expected since tax progressivity alters elasticities roughly in a common across the firm distribution.

Removing oligopsony raises the fraction of the wage variance explained by firms by 2 percentage points, or equivalently 9 percent above the baseline (Row 2). Recall that more productive firms pay higher wages at wider markdowns. Removing this source of wider markdowns increases wages at these firms. This spreads out the distribution of wages. Unlike removing tax progressivity, wages fan out because of wages at high wage firms, so the variance of firm fixed effects increases.

Removing asset-based sorting sharply increases the variance of log wages, and the share due to firm fixed effects. Within a skill group e, richer households sort toward more productive firms. Since richer households are less elastic, this leads to wider markdowns at more productive firms.[13] Narrowing these disproportionately wide markdowns at high z firms increases the role of firm fixed effects by 6 percentage points.

Cumulatively taxes, oligopsony, and asset-based sorting revives firms' contribution to wages significantly. After jointly removing these sources, firm fixed effects explain 19 percent of wage variance, as opposed to 13 percent in our baseline model. This represents a 42 percent increase, suggesting an important *wage compression* role of firms.

In Row 5, we remove the remaining skill heterogeneity in wealth effects which sharply increases wage inequality, but entirely through the variance of worker fixed effects. Figure 3C showed that high skill workers tend to have wider markdowns at all firms. So, when this component is removed, wages rise disproportionately for high skill workers. This shows up as an increase in the variance of worker fixed effects.

[13] High e and high a are correlated (by construction, as informed by the SCF), and high e sort toward high z firms, where markdowns are wider.

Source removed (cumulatively)	Wage variance $Var(\log w_j^i)$ (1)	Due to firm F.E.'s $Var(\psi_j)$ (2)	Share of variance $Var(\psi_j)/Var(\log w_j^i)$ (3)	Relative to baseline (=1) (4)
Baseline	0.38	0.05	0.13	1.00
No tax progressivity	0.39	0.05	0.13	1.00
No oligopsony	0.42	0.06	0.15	1.09
No asset-based sorting	0.47	0.09	0.19	1.42
No wealth effects	0.68	0.09	0.13	1.00
No monopsony	0.68	0.09	0.13	1.00

Table 5: Role of firm fixed effects in earnings decompositions

Note: This table reports the variance of firm fixed effects and its share over log wage variance. Each row represents the marginal effect of removing the specified factor while maintaining the removal of all preceding factors. Column (1) reports the raw variance of log earning. Column (2) reports the raw variance of log firm fixed effects. Column (3) reports the share of log earning variance attributed to firms. Column (4) reports the share as an index relative to the baseline, normalized to 1.

Lastly, Row 6 sets η to infinity. Since we have already removed the other sources of market power, setting η to infinity represents a fixed percentage increase in wages for all worker and firm types. The average wage increases, with no effect on the variance.

Summary: Firms matter for wage setting, even if their fixed effects have small variance. Table 5 establishes that wage setting by firms is precisely why firm effects play a muted role in wage decompositions. Removing oligopsony and variable markdowns from asset-based sorting raises the contribution of firm fixed effects by 5 percentage points. This exercise suggests that firms' wage setting plays an important role in *mitigating* wage inequality (by as much as 40 percent) and remain integral in understanding wage inequality.

5.5. Implications for inequality

From the above results, it should be clear that firm wage-setting has implications for inequality in terms of wages and consumption, as well as measures of concentration of labor markets. To quantify these effects, we consider three modes of firm wage-setting: (1) price discrimination is extended to assets a, (2) price discrimination is limited to a firm-level wage, and (3) price discrimination is removed, i.e. competitive wages. Table 6 reports the results. In Panel A we recalibrate $(\overline{Z}, \overline{\varphi})$ to match average earnings and hours, while we do not in Panel B.

Our first result in Column (2) is that by allowing firms to (a, e) discriminate, the variance of log earnings increases by 1pp, while the variance of log consumption falls by 1pp. Since this is a static model, the only way this configuration can occur (higher $var(w)$, lower $var(c)$) is if wage gains occur among low asset households (who now eat more) and wage losses occur among high asset households (who now eat less). This is precisely what firms want to accomplish: they want to pay high asset households less, as they are especially inelastic.

If firms are restricted to equal pay for all employees, inequality *rises* (Column 3). Unable to discriminate on worker type, across-worker variation in markdowns in Figure 3C are flattened. Setting wages with the average-e in mind, leads firms to effectively narrow markdowns and increase wages for high-e workers, while widening markdowns and lowering wages for low-e workers. In partial equilibrium, this leads to more wage and consumption inequality. Wages were marked down most steeply for high-e workers at high-z firms. Hence, narrowing these markdowns improves high skill wages disproportionately at high wage firms. In response, high skill workers reallocate to high wage firms, and sorting (measured by the covariance of worker and firm fixed effects) increases dramatically, rising by 50 percent. In general equilibrium, this amplifies the increase in wage and consumption inequality.

If markets were competitive, then earnings inequality would increase further, nearly doubling from its baseline level (Column 4). As we have shown, imperfect competition disproportionately reduces the wages at the highest productivity firms and for the highest skill workers (recall Figure 1A and Figure 3C). In partial equilibrium, competitive wages undo this compression. In general equilibrium, this leads to a substantially more reallocation of high-e workers toward high-z firms, relative to low-e workers. The covariance of firm and worker fixed effects is twice the baseline level. This amplifies the effects of competitive wages on consumption and wage inequality. There is much more inequality in a competitive environment.

Table 6 also shows that our environment is one in which higher concentration can reflect gains in economic efficiency. Steps toward less wage discrimination increase market concentration. This is consistent with earlier findings in Berger et al. (2022). Since wages are most distorted at high wage firms with the highest market shares, then steps toward competitive wages increase employment at these firms, increasing concentration.

Summary: Firm conduct is a key determinant of wage inequality. Table 6 corroborates our conclusions in the prior section: firm conduct is key to understanding wage inequality. These findings have important policy implications. If, compared to low-wage workers, high-wage workers' wages reflect smaller pieces of their marginal product, then policies that counteract labor market power will inherently increase wage inequality while also increasing concentration. This contrasts with earlier views that monopsony was prevalent among the lowest earners in the labor market, in which case such policies would lower inequality. In contrast, our framework and results highlight that an efficiency-redistribution tradeoff remains central to policy evaluation of labor market power.

6. Application II - Tax progressivity

How do changes in tax progressivity affect wages, sorting, output and consumption? Our key equations (7) showed how higher progressivity ($\uparrow \tau_1$) reduced extensive and intensive margin

Wage discrimination:	**Baseline**	**More**	**Less**	**None**
Wages set at ...	(e,j)-level	(a,e,j)-level	(j)-level	$\mu_j(a,e)=1$
Moment	(1)	(2)	(3)	(4)
Panel A. With Recalibration				
Variance of log consumption	0.27	0.26	0.35	0.45
Variance of log wages	0.38	0.39	0.51	0.70
Variance of worker effects	0.25	0.26	0.35	0.47
Variance of firm effects	0.05	0.05	0.04	0.07
$2\times$ Covariance	0.06	0.06	0.09	0.12
HHI - Payroll	0.12	0.12	0.11	0.17
Panel B. Without Recalibration				
Variance of log consumption	0.27	0.26	0.35	0.45
Variance of log wages	0.38	0.39	0.50	0.63
Variance of worker effects	0.25	0.26	0.34	0.42
Variance of firm effects	0.05	0.05	0.04	0.09
$2\times$ Covariance	0.06	0.06	0.09	0.09
HHI - Payroll	0.12	0.12	0.11	0.13

Table 6: Market structure and its implications for inequality

Note: This table compares the effects of firms' wage setting structure on inequality and labor market concentration. To investigate this, we consider three specifications of firm conduct across the counterfactual. Column (1) is the Baseline. In column (2), firms are unrestricted and can wage discriminate by both skills e and assets a ((a,e,j)-level). In Column (3), firms are restricted to offer a single firm-level wage ((j)-level). In column (4), firms set a competitive wage equal to marginal product. Panel A reports results where the productivity shifter ($\bar{Z}$) and labor disutility shifter ($\bar{\varphi}$) are recalibrated in each counterfactual to maintain average earnings and hours constant at baseline levels. Panel B reports results without this recalibration, holding all structural parameters fixed at baseline estimates.

elasticities of labor supply to the firm. Qualitatively, this has two effects. First, keeping wages fixed, if individuals are less elastic with respect to wages then sorting toward high wage, high productivity firms will weaken. Second, as firms set wages, less elastic labor supply will lead to wider markdowns. Both effects would lead to efficiency losses in the economy: aggregate employment remains constant in terms of workers, but hours and output per hour would decline. In this section we investigate these mechanisms quantitatively.

Our exercise is simple. In our baseline calibration we vary $\tau_1 = 0.06$ between 0 and 0.20. The intercept in the tax function τ_0 changes to ensure that the overall fiscal position of the tax and transfer program remains constant.

6.1. Micro - Workers, Firms and Sorting

Workers. Higher progressivity has the expected effects in terms of narrowing inequality between workers. Figure 5A shows that the standard deviation of after tax earnings is reduced by nearly 8 percentage points as τ_1 increases from 0.06 to 0.20. Consumption inequality falls by slightly less due to the presence of transfers and non-labor income. These effects are expected. A novel aspect of our theory is linking tax progressivity to wage setting by firms. This shows

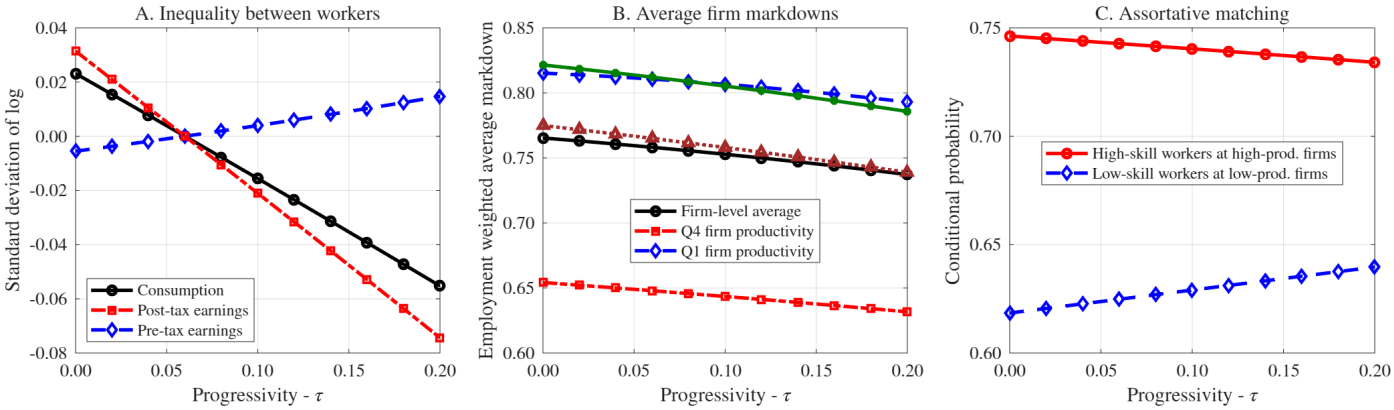


Figure 5: Tax progressivity - Workers and firms

up in the increase in dispersion in *pre-tax earnings*, a force against the compression of inequality in post-tax earnings and consumption. To understand this we now turn to firms.

Firms. Our framework links firm wage setting to tax policy via the endogenous interaction of tax policy and workers' elasticities of labor supply. Figure 5B shows that higher progressivity has a modest effect on widening markdowns as workers become less elastic with respect to pre-tax wages. On average, markdowns widen by about 2 percentage points. Effects are larger at high productivity firms due to non-linearities in the interaction between progressivity and labor supply elasticities. An increase in τ_1 delivers a larger reduction in markdowns when $\varepsilon_{jm}^{i,\rho}$ is smaller.[14]

Sorting. Figure 5C shows the response of sorting. As labor becomes less elastic, workers respond less to wages at high wage firms, and sort more toward lower productivity firms. For low skill workers (below median e), employment at low productivity firms (below median z) increases. For high skill workers, this is also true. Here we plot the flipside of this observation, which is that high skill employment at *high productivity firms* decreases.

6.2. Macro - Output, Consumption, Productivity

We now turn to aggregates (Figure 6). 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, which acts like a wedge in labor demand. As labor demand contracts, wages decrease, and equilibrium employment, output and consumption decline. Less than half of the 2.1 percent decline in output is due to hours, with the remainder being due to a decline in aggregate labor productivity (Panel A).

[14]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.

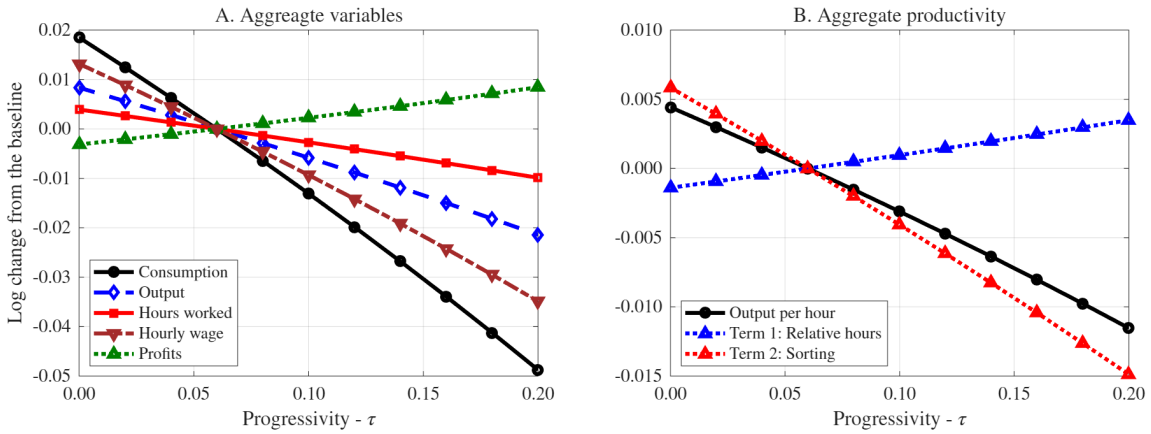


Figure 6: Tax progressivity - Macroeconomic aggregates

To understand the decline in aggregate productivity, we study an exact decomposition of the change in output-per-hour. Aggregate output and hours are:

$$Y = \int_{A\times E} \sum_{j=1}^{J} \rho(a,e,j) h(a,e,j) A(e,z_j) \Gamma(a,e) d(a,e) \quad , \quad H = \int_{A\times E} \sum_{j=1}^{J} \rho(a,e,j) h(a,e,j) \Gamma(a,e) d(a,e)$$

Let $Z = Y/H$ be aggregate labor productivity. To simplify notation, let k denote (a,e,j). We have following exact decomposition across tax policies, where $\widehat{X}$ is percentage change in X:

$$\widehat{Z} = \underbrace{\sum_k \omega(k)\left[\widehat{h}(k) - \widehat{H}\right]}_{\text{1. Relative hours (intensive margin)}} + \underbrace{\sum_k \omega(k)\widehat{\rho}(k)}_{\text{2. Sorting (extensive margin)}} + \underbrace{\sum_k \omega(k)\widehat{h}(k)\widehat{\rho}(k)}_{\text{3. Interaction}} \quad , \quad \underbrace{\omega(k) = \frac{y(k)}{Y}}_{\text{Output share}}$$

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 most productive 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 firms. The third term captures their interaction.

The sorting term is responsible for the entire decline in aggregate labor productivity (Figure 6B). Total labor productivity falls substantially: -1.14 percent. The intensive margin term actually contributes positively: $+0.35$ percent. The usual force of high skill workers reducing their hours more than low skill workers when facing progressive taxes is offset by large reductions in hours at low productivity firms. The extensive margin sorting term accounts for more than the entire decline: -1.49 percent. This elasticity is 'missing' from standard theories of optimal taxation.

We conclude that 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. Moreover, the model delivers

a novel understanding of the effect of an increase in taxes. Labor income taxes make it more expensive for a firm to get a dollar of wages to the worker, which when internalized, reduces labor demand, and lowers wages.

6.3. Implications for the *Elasticity of Taxable Income*

Our model may be useful for resolving a standard puzzle in public finance: empirical estimates of the effects of increases in marginal tax rates on earnings are large, while model based estimates are small. Two issues in standard models give rise to this result. First, such models are calibrated to micro-empirical estimates of labor supply elasticities on the hours and / or participation margin. These elasticities are small, hence taxes lead to small reductions in hours. Second, in these models, higher taxes reduce labor supply, causing equilibrium wages to increase. This partially offsets the decline in earnings due to lower hours. Our model addresses both issues. First, we maintain a low Frisch elasticity of labor supply (0.08), and match empirical measures of marginal propensities to earn following windfall gains (-0.53), however the key feature is a high elasticity of labor supply on the extensive margin across firms. Disproportionately taxing wages in good matches, pushes sorting away from these firms, reducing earnings. Second, as higher taxes reduce labor demand, equilibrium wages decrease, amplifying the decline in earnings.

When we use a standard approach to measuring the Elasticity of Taxable Income (ETI), we find estimates in the range described by Saez et al. (2012). The ETI is measured as

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

where $\overline{\tau}$ is the earnings-weighted average marginal tax rate. Increasing τ_1 from 0.06 to 0.20 increases the average marginal tax rate from 0.34 to 0.45 (recall that the budget remains balanced by τ_0). Total earnings decline by 4.3 percent. Hence the ETI is 0.24. This is within the range documented in Saez et al. (2012) of 0.12 to 0.40. The key novelty in our case, is that the change in labor earnings is delivered primarily through sorting and lower wages, rather than a reduction in total hours and disproportionate reduction in hours of productive workers.

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 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.

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, we showed that 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," *The 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," Tech. rep., 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*, JSTOR.

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

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

——— (2025b): "Robinson Meets Roy: Monopsony Power and Comparative Advantage," *Working Paper*.

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," Tech. rep., 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.

CHAN, M., E. MATTANA, S. SALGADO, AND M. XU (2023): "Wage setting and passthrough: The role of market power, production technology and adjustment costs," *Working Paper*.

CONSTANTINIDES, G. M. AND D. DUFFIE (1996): "Asset pricing with heterogeneous consumers," *Journal of Political economy*, 104, 219–240.

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*.

DINERSTEIN, M., S. EARNEST, D. K. KOUSTAS, AND C. YANNELIS (2025): "Student loan forgiveness," Tech. rep., National Bureau of Economic Research.

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 (2021): "How Americans Respond to Idiosyncratic and Exogenous Changes in Household Wealth and Unearned Income," NBER Working Paper 29000, National Bureau of Economic Research.

——— (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.

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

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., 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 (2015): "Multidimensional skills, sorting, and human capital accumulation," *(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.

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. (2021): "Valuing labor market power: The role of productivity advantages," Tech. rep., 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.

VIVALT, E., E. RHODES, A. W. BARTIK, D. E. BROOCKMAN, AND S. MILLER (2024): "The employment effects of a guaranteed income: Experimental evidence from two US states," Tech. rep., National Bureau of Economic Research.

VOLPE, O. (2024): "Job Preferences, Labor Market Power, and Inequality," Tech. rep., Working Paper.

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