# Dynamic Wage Setting: The Role of Monopsony Power and Employment Costs*

Mons Chan† Elena Mattana‡ Sergio Salgado§ Ming Xu¶

August 26, 2026

## Abstract

We develop and estimate a model of dynamic wage setting that jointly identifies latent worker ability, firm productivity, and wage markdowns, defined as the ratio of wages to the marginal revenue product of labor. Using matched employer-employee data and firm accounts, we document that markdowns widen with average worker ability and firm productivity. We decompose markdowns into a static monopsony wedge, determined by the firm-level labor supply elasticity, and a dynamic cost wedge reflecting non-wage employment costs and the continuation value of labor. Static monopsony power accounts for most of the average level of markdowns but explains very little of their cross-sectional variance. Nearly all dispersion in markdowns arises from the dynamic cost wedge. Standard wedge-accounting exercises that treat observed markdowns as measures of employer market power would misattribute most of the within-market markdown dispersion to the wrong source.

**JEL classification numbers:** J31, J42, J23, D24
**Keywords:** Wage Setting, Monopsony, Markdowns, Labor Supply Elasticity, Adjustment Costs, Misallocation.

*A previous version of this paper circulated as "Wage Setting and Passthrough: The Role of Market Power, Production Technology and Adjustment Costs." We thank David Berger, Manudeep Bhuller, João Galindo da Fonseca, Kyle Herkenhoff, Kory Kroft, Rasmus Lentz, Huw Lloyd-Ellis, Jakob Roland Munch, Qingsong Pan, Luke Rawling, Rune Vejlin, and seminar participants at Aarhus, Bank of Canada, Calgary, Copenhagen, Queen's, York, Shandong, Sun Yat-sen, Waseda, Waterloo, BSE Summer Forum, Lisbon Macro Workshop, CMSG, MEBDI, St. Louis Fed Macro-Labor Conference, Advances in Labor Macroeconomics Conference at Richmond Fed, Columbia Junior Micro/Macro Labor Conference, and the Nordic Summer Institute in Labour Economics for helpful comments and suggestions. The work presented here is supported by the Carlsberg Foundation, grant CF23-1544.

†Queen's University. E-mail: mons.chan@queensu.ca
‡Aarhus University. E-mail: emattana@econ.au.dk
§The Wharton School, University of Pennsylvania. E-mail: ssalgado@wharton.upenn.edu
¶Queen's University. E-mail: ming.xu@queensu.ca

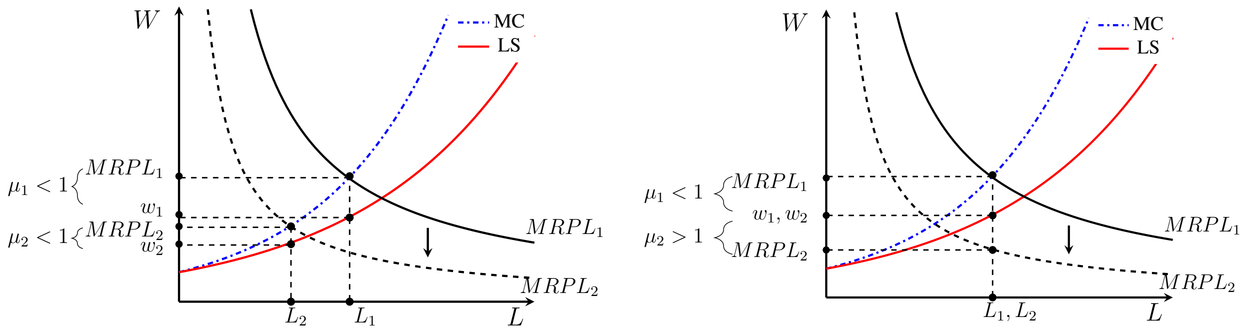
# 1 Introduction

Workers are often paid less than their marginal revenue product of labor (MRPL) and firm-level wages move when firms are hit by productivity or demand shocks. These facts are central to current debates on labor market power, rent sharing, wage inequality, and the transmission of firm shocks to workers' wages (Card et al., 2018; Berger et al., 2022; Lamadon et al., 2022; Yeh et al., 2022; Chan et al., 2024). Yet, the wage–MRPL gap does not have a unique structural interpretation. In the canonical static monopsony model, the markdown, $\mu_{jt} \equiv W_{jt}/MRPL_{jt}$, is pinned down by the labor supply elasticity faced by the firm, so a wider gap maps directly into greater employer market power. Labor, however, is a dynamic input. Hiring, training, reorganizing, and dismissing workers are costly, and the marginal worker may be valuable because of future production rather than current MRPL alone. The same observed gap may therefore reflect both limited worker responsiveness to wages and the firm's marginal cost and shadow value of employment. This distinction changes both the relevant policy response and the welfare interpretation of markdown dispersion. Our central question is therefore: how much of observed wage markdowns reflects monopsony power, and how much reflects non-wage employment costs and firms' forward-looking labor-demand decisions?

Figure 1 illustrates the difference between these two interpretations of the wage–MRPL gap. Panel (a) shows the static benchmark, where a negative productivity shock shifts the firm's MRPL schedule downward, the firm contracts employment along its labor supply curve, and wages and marginal products fall together. The markdown remains below one and is determined by the labor supply elasticity. If the firm instead faces costly adjustment and uncertainty, it may optimally respond to the same shock by retaining more labor than it would choose if employment were freely adjustable. Paying additional wages today may be cheaper than firing and then rehiring needed labor tomorrow (Hopenhayn and Rogerson, 1993). This may even lead firms to pay $W_{jt} > MRPL_{jt}$ despite facing upward-sloping labor supply, as depicted in Panel (b).

Expanding on this simple example, we develop a tractable model of dynamic wage determination in which the firm chooses employment taking into account its monopsony power in the labor market, current non-wage employment costs, and the expected future value of its workforce. In this framework, the marginal cost curve relevant for wage and employment decisions depends on both the labor supply elasticity and net non-wage employment costs. The model delivers a multiplicative decomposition of the expected wage markdown into two components: a monopsony wedge $\mu^{\varepsilon}_{jt}$, governed by the static elasticity of labor supply to the firm, and a cost wedge $\mu^{\phi}_{jt}$, which captures the net present value of labor. The expected markdown can therefore be written as $\overline{\mu}_{jt} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}$. When $\mu^{\phi} = 1$, the firm is at the static monopsony optimum for wages and

Figure 1: FIRM RESPONSES TO PRODUCTIVITY SHOCKS




**(a)** Unconstrained firm **(b)** Fully constrained firm

*Notes:* Figure 1 depicts a firm facing upward-sloping labor supply (LS) and a marginal cost of labor (MC) that exceeds the wage. A negative productivity shock shifts the MRPL schedule downward. In panel (a), an unconstrained firm adjusts employment freely along its labor supply curve; both wages and employment fall, and the markdown $\mu \equiv W/MRPL$ remains below one in both periods. In panel (b), a fully constrained firm does not adjust employment due to adjustment costs. Employment and wages remain at their pre-shock levels while MRPL falls below the wage, producing a markdown above one ($\mu_2 > 1$) despite the firm facing the same upward-sloping labor supply.

employment. A cost wedge greater than one implies the future value of labor exceeds current costs and the firm may pay wages exceeding the MRPL. On the other hand, a cost wedge below one implies firms set wages and employment below their static-monopsony levels.

Our paper makes three main contributions: First, we develop a structural framework that delivers firm-level distributions of productivity, worker ability, and markdowns; our structural approach controls for unobserved worker heterogeneity in the estimation of productivity and markdowns. Second, our framework separates the markdown into a monopsony and a dynamic employment cost component; we provide direct evidence, prior to imposing the structural model, that labor adjustment frictions are a first-order determinant of firm-level markdowns. Third, we quantify the monopsony and cost wedges and show that although static monopsony determines the level of markdowns, dynamic adjustment frictions drive markdown dispersion almost entirely.

To estimate the model, we first combine matched employer–employee data covering the population of Danish workers with accounting data for the universe of private-sector Danish firms. We use a model-implied AKM-style wage decomposition to recover worker ability and firm-year wages per efficiency unit, construct an ability-adjusted labor input, and then estimate a flexible revenue production function following Gandhi et al. (2020). This delivers the joint distributions of firm productivity, MRPL, and wage markdowns without imposing the full dynamic wage-setting model or taking a stance on whether wages are posted, bargained, or determined through another firm-level mechanism.

The median firm pays workers 79% of their MRPL, and about 18% of firms pay workers more than their current marginal product. Wage–MRPL gaps are particularly wide among the largest and most productive firms, firms employing higher-ability workers, and firms with larger local employment shares. Correcting for latent worker ability is essential for these comparisons. If we instead treat workers as homogeneous and measure labor using total hours, the standard deviation of markdowns rises from 0.27 to 0.65, the fraction above one rises from 18% to 30%, and the standard deviation of persistent productivity nearly doubles. Differences in labor quality would be loaded mechanically onto firm productivity and MRPL, creating spurious markdown dispersion and substantially overstating the heterogeneity that must be explained by labor market power or employment frictions.

Expected markdowns rise sharply following negative productivity shocks but respond little to positive shocks. Labor also adjusts much less to these shocks than intermediate inputs, consistent with firms retaining workers when employment is costly to adjust. Episodes in which wages exceed MRPL are transitory: affected firms are less profitable, more leveraged, and more likely either to exit or to return toward the economy-wide markdown. We establish these results before imposing the structural wage-setting model, providing direct evidence that observed markdowns reflect firms' dynamic employment decisions as well as worker-side labor supply.

To identify the monopsony wedge $\mu^{\varepsilon}_{jt}$, we develop a tractable effective-labor extension of the nested-logit labor-supply frameworks used in Lamadon et al. (2022) and Chan et al. (2024). Our formulation allows heterogeneous workers to sort systematically across firms, while ensuring that the estimated elasticity applies to the same ability-adjusted labor input used in production, the wage bill, and the firm's first-order condition. The mean labor supply elasticity is 2.87, with both a mean and median monopsony wedge of 0.74. This is remarkably close to the median expected markdown of 0.77 recovered independently from the production side. Static monopsony therefore provides a useful benchmark for the level of markdowns. It does not, however, explain their dispersion: the monopsony wedge accounts for only about 4–5% of the variance of log expected markdowns. Alternative labor-supply specifications and less structured exercises yield the same conclusion.

We recover the cost wedge as $\mu^{\phi}_{jt} = \overline{\mu}_{jt}/\mu^{\varepsilon}_{jt}$. The cost wedge reflects current non-wage employment costs and the continuation value of labor. The median firm values workers 5% more than their current MRPL, while firms at the 90th percentile value the marginal unit of labor 47% more than its current MRPL. Importantly, the cost wedge is not simply an unexplained residual. Observed firm states and persistent firm heterogeneity account for more than 80% of its variation, with variables related to the expected future value of labor, workforce structure, and firm amenities playing the largest roles. Consistent with these results, the relationships between markdowns and firm size or productivity are driven almost entirely by the cost wedge, whereas the relationship with labor-market shares has a substantially clearer monopsony component. Exit rates are also V-shaped in the cost wedge, with their minimum near $\mu^{\phi}_{jt} = 1$, further supporting its interpretation as an economically meaningful measure of departures from the static monopsony benchmark.

The decomposition also changes the interpretation of both aggregate wage gaps and cross-firm markdown dispersion. Static monopsony accounts for approximately 72% of the aggregate wage–MRPL gap, confirming that employer market power is central to the overall division of surplus between firms and workers. By contrast, the cost wedge accounts for roughly 96% of the within-market revenue-loss index generated by markdown dispersion. Our results suggest that the accounting losses due to dispersion in markdowns should not automatically be interpreted as a welfare distortion. This dispersion may reflect costly adjustment, labor hoarding, or the value of retaining workers under uncertainty. Observed markdowns are also poor policy screens. Among firms in the decile with the widest observed wage–MRPL gaps, less than one-quarter are also in the decile with the strongest monopsony component. Observed wage–MRPL gaps are therefore informative accounting objects, but they are not sufficient statistics for labor market power, misallocation, or the appropriate policy response.

**Relation to the literature.** Our paper first contributes to the literature on employer wage-setting power and wage markdowns. Direct estimates of labor supply to the firm show that employers face finite elasticities, although the magnitudes vary across settings and identification strategies (Sokolova and Sorensen, 2021). Matched-worker and quantitative models connect this wage-setting power to rent sharing, wage inequality, sorting, and the allocation of labor across firms (Card et al., 2018; Berger et al., 2022; Lamadon et al., 2022; Chan et al., 2024). Production-side approaches complement this evidence by recovering wage–MRPL gaps from firm accounts (Brooks et al., 2021; Yeh et al., 2022); see Azar and Marinescu (2024), Kline (2025), and Berger et al. (2026) for recent surveys. Relative to this literature, we use production-side markdowns and labor-supply elasticities jointly, rather than as alternative measures of the same object, to identify how much of the observed markdown is attributable to static monopsony.

A growing body of recent work asks how much of the measured wage–MRPL gap reflects static monopsony power rather than dynamic labor market frictions. A long literature establishes that labor is costly to adjust and that the structure of these costs shapes firms' responses to shocks (Hamermesh, 1989; Hopenhayn and Rogerson, 1993). Agostinelli et al. (2025) show that dynamic wage–tenure contracts make instantaneous markdowns misleading, because they ignore the present value of future wage growth committed through renegotiation. Seegmiller (2025) develops a related framework and decomposes the markdown into a dynamic component driven by hiring costs and a static markup wedge, finding that much of the gap is dynamic. Other recent work emphasizes wage-setting and institutional constraints (Bhuller et al., 2025) as well as future wage paths and backloaded compensation (Han, 2024; Rubens and Silveira, 2026).

On the empirical side, Faia et al. (2026) find that measured markdowns respond by only 0–25% of the amount predicted by the monopsony model following the introduction of Germany's minimum wage, consistent with a large non-monopsony component in the measured markdown. We examine two of their proposed explanations – that markdowns are mismeasured and that labor may face adjustment costs. Dhyne et al. (2025) show that fixed overhead labor costs, not monopsony, dominate wage responses to demand shocks in Belgium, while Petrin and Sivadasan (2013) argue that labor adjustment costs play a significant role in driving gaps between the MRPL and wages. Similarly, Asker et al. (2014) show that dynamic capital adjustment can generate dispersion in static MRPK measures without implying capital misallocation; we establish an analogous result for labor. We contribute to this literature by providing a unified decomposition of the measured markdown into a monopsony wedge and a cost wedge, and by quantifying their relative roles in explaining markdowns.

Finally, our approach to estimating productivity while controlling for unobserved labor quality builds on Hellerstein and Neumark (2007), Fox and Smeets (2011), Bagger et al. (2014), and Bagger and Lentz (2019). We differ in using a flexible production function that does not restrict the form of factor-augmenting technology (see also Doraszelski and Jaumandreu, 2018; Demirer, 2026), and in constructing ability-adjusted labor inputs that enter both production estimation and the structural wage-setting model.

The rest of the paper proceeds as follows. Section 2 develops the model. Section 3 describes the data. Section 4 presents the identification and estimation strategy. Section 5 reports the distributions of wages, productivity, and markdowns, and presents evidence on firm dynamics and adjustment costs. Section 6 quantifies the relative roles of monopsony power and dynamic employment costs in wage setting. Section 7 examines the implications of our decomposition for markdown accounting and labor misallocation. Section 8 concludes.

# 2 Model of Firm Dynamics and Wage Setting

In this section, we develop a dynamic model of wage setting in which the observed markdown reflects two forces: upward-sloping labor supply and dynamic employment costs. We divide the exposition of our framework into three subsections. Subsection 2.1 contains the theory and assumptions needed to estimate worker ability, firm productivity, ability-adjusted wages, MRPL, and markdowns. Subsection 2.2 discusses labor supply and adjustment costs. Subsection 2.3 presents the firm's dynamic optimization problem and derives the markdown.

## 2.1 Production and Worker Ability

We consider an environment with a finite number of firms indexed by $j$ and a continuum of workers indexed by $i$. Firms maximize expected discounted profits in each period $t$ conditional on their payoff-relevant information set, $\mathcal{I}_{jt}$, and operate a continuous and differentiable production function

$$Q_{jt} = F(K_{jt}, L_{jt}, M_{jt})e^{\nu^Q_{jt}},$$

with Hicks-neutral physical productivity term $\nu^Q_{jt}$, which combines capital ($K_{jt}$), labor ($L_{jt}$), and intermediate inputs ($M_{jt}$) to produce output ($Q_{jt}$). Firm $j$ faces inverse demand

$$P_{jt} = P(Q_{jt}, \chi_{jt}),$$

where $\chi_{jt}$ is a firm-specific demand shifter. Firm revenue is therefore $R_{jt} = P(Q_{jt}, \chi_{jt})Q_{jt}$. We assume that the demand system and physical production function admit a multiplicatively separable revenue representation,

$$R_{jt} = R(K_{jt}, L_{jt}, M_{jt})e^{\nu_{jt}}, \tag{1}$$

where revenue is a function of observed production inputs and a scalar revenue productivity term $\nu_{jt}$ that nests unobserved demand shifters and production efficiency ($\chi_{jt}$ and $\nu^Q_{jt}$).[1]

The term $e^{\nu_{jt}}$ captures the firm's Hicks-neutral revenue TFP (or TFPR), which in logs follows

$$\nu_{jt} = \omega_{jt} + \epsilon_{jt} = h(\omega_{jt-1}) + \eta_{jt} + \epsilon_{jt}, \tag{2}$$

where $\omega_{jt}$ is the persistent component of productivity and $\epsilon_{jt}$ is an ex-post productivity shock. The productivity terms $\nu_{jt}$, $\omega_{jt}$, $\eta_{jt}$, and $\epsilon_{jt}$ may therefore nest the demand state $\chi_{jt}$ along with physical productivity. The persistent productivity, $\omega_{jt}$, evolves as a first-order Markov process with $h(\omega_{jt-1}) \equiv E[\omega_{jt} \mid \omega_{jt-1}]$ and i.i.d. innovation $\eta_{jt}$. The ex-post productivity innovation, $\epsilon_{jt}$, is i.i.d. across firms and time, with $\mathbb{E}[\epsilon_{jt}] = 0$ and $\epsilon_{jt} \perp\!\!\!\perp \mathcal{I}_{jt}$.

We assume that firms are price takers in intermediate input and capital markets, but not in labor markets.[2] Firms differ in a set of exogenous characteristics (e.g., industry), denoted $Z_{jt}$, and endogenous workforce characteristics (e.g., average education of their workers), denoted $\widetilde{Z}_{jt}$.

We make standard assumptions about input timing (Gandhi et al., 2020). First, the persistent

[1] A sufficient condition is isoelastic demand as in Klette and Griliches (1996) and De Loecker (2011). If $P(Q,\chi) = S(\chi)Q^{\rho-1}$, with $\rho \in (0,1]$, then firm revenue satisfies $P(Q,\chi)Q = S(\chi)Q^{\rho}$. Substituting physical output $Q = F(K,L,M)e^{\nu^Q}$ gives $P(Q,\chi)Q = [F(K,L,M)]^{\rho}\exp\{\rho\nu^Q + \log S(\chi)\}$, so demand shifters are absorbed into the scalar productivity term and input revenue elasticities are stable functions of inputs.

[2] This assumption on capital markets is much stronger than needed to identify the revenue production function and markdowns. A sufficient condition is that capital is predetermined.

innovation $\eta_{jt} \in \mathcal{I}_{jt}$, and thus $\omega_{jt} \in \mathcal{I}_{jt}$, are observed prior to choosing period-$t$ inputs. Second, capital is predetermined by investment decisions in period $t-1$ (i.e., $K_{jt} \in \mathcal{I}_{jt}$). Third, labor and intermediates are chosen in period $t$ conditional on $\mathcal{I}_{jt}$. We assume intermediate inputs are fully flexible with no adjustment costs and are optimally chosen, while labor is dynamic and the level may depend on lagged labor and labor force characteristics. This implies $\{L_{jt-1}, \widetilde{Z}_{jt-1}\} \subset \mathcal{I}_{jt}$ while $M_{jt-1} \notin \mathcal{I}_{jt}$. Finally, firms observe the ex-post productivity shock $\epsilon_{jt}$ after making input decisions, so period-$t$ input choices are not conditional on $\epsilon_{jt}$, that is $\epsilon_{jt} \notin \mathcal{I}_{jt}$.[3] Workers' realized wages, however, may depend on the ex-post shock.

Firm $j$'s labor input is given by an ability-weighted sum of total hours worked

$$L_{jt} = \sum_{i \in j} A_{it} H_{it}, \tag{3}$$

where $A_{it}$ represents worker $i$'s time-varying ability (capturing innate unobserved ability and observed characteristics such as education or experience) and $H_{it}$ represents the quantity (in hours) of labor that individual $i$ provides to firm $j$ in period $t$.[4] This specification implies that workers are perfect substitutes in production conditional on their ability. The marginal revenue product of an hour supplied by worker $i$ is $A_{it} MRPL_{jt}$, so the marginal revenue product per ability unit is common within the firm. We also impose that firm $j$ pays a common price $W_{jt}$ per hour of ability-adjusted labor.[5] Worker $i$'s hourly wage is then $W_{ijt} = A_{it} W_{jt}$. We allow the firm-level wage to depend on the ex-post shock $\epsilon_{jt}$ by taking the following form: $W_{jt} = W^c_{jt} f^c(\epsilon_{jt})$, where $W^c_{jt}$ is a contract wage rate offered by the firm when making hiring decisions and $f^c$ is a function that represents the degree to which firms modify the ex-post wage relative to the expected wage once $\epsilon_{jt}$ is realized (e.g., performance bonuses). The expected wage then takes the form

$$\overline{W}_{jt} = E_{\epsilon_{jt}}[W_{jt} | W^c_{jt}] = W^c_{jt} \mathcal{E}^c,$$

where $\mathcal{E}^c \equiv E_{\epsilon_{jt}}[f^c(\epsilon_{jt})]$ is the expected ex-post wage adjustment. Though the value of $\epsilon_{jt}$ is unknown at the time of hiring, workers and firms know the distribution of $\epsilon_{jt}$ and the shape of $f^c(\epsilon_{jt})$, and base their employment and hiring decisions on $\overline{W}_{jt}$.

[3] Ex-post shock $\epsilon_{jt}$ therefore captures variation in revenue that is uncorrelated with inputs.

[4] We can rewrite the labor input as $L_{jt} = H_{jt} \bar{A}_{jt}$ where $H_{jt}$ is the total hours of labor employed, and $\bar{A}_{jt}$ is the hours-weighted time-varying mean worker ability at the firm. This is similar to the literature that estimates firm-level labor-augmenting productivity (e.g., Demirer (2026) and Doraszelski and Jaumandreu (2018)). The key difference is that those papers assume labor markets are perfectly competitive with no adjustment frictions, while we achieve identification in the presence of both labor market power and dynamic adjustment costs.

[5] This means that firms are indifferent in production between one hour of labor from a highly skilled worker and two hours of labor from workers with half the skill of the first. Workforce composition may nevertheless affect recruiting, screening, training, retention, management, and other non-wage labor costs.

**Discussion.** The assumptions we made thus far, along with a few more technical assumptions that we discuss in Section 4, are sufficient to provide identification of the production function, firm productivity, MRPL, firm-level (expected) wages, workers' ability, and wage markdowns. Our assumption that the law of one price holds within the firm (conditional on ability) in particular is what allows us to identify and control for unobserved worker heterogeneity when estimating productivity. While strong, this assumption is consistent with $W_{jt}$ being determined through wage posting, bargaining, or other firm-level wage-setting mechanisms. The restriction on $P(Q_{jt}, \chi_{jt})$ is what allows us to recover the revenue production function and MRPL without observing firm-specific output prices or product markups. It nests output price-taking, isoelastic or CES residual demand with constant markups, and constant-conduct environments in which the relevant marginal revenue wedge is constant conditional on the variables included in $R(\cdot)$. It also allows unobserved demand-level shifters to vary across firms and time, provided they enter the scalar productivity terms and do not separately change input elasticities. We do not, however, need to make any assumptions about the optimality of labor or capital input choices, the nature of adjustment or fixed costs, or how wages are determined to estimate the productivity, workers' ability, and markdown distribution.

## 2.2 Labor Supply and Employment Costs

Labor markets are imperfect, and firms may have monopsony power. Firm $j$ faces an upward-sloping supply of total effective labor, $L_{jt} = L\left(\overline{W}_{jt}, Z_{jt}, \widetilde{Z}_{jt-1}, u_{jt}, I_{gt}\right)$, which depends on the expected wage, $\overline{W}_{jt}$, and predetermined firm and workforce characteristics $\{Z_{jt}, \widetilde{Z}_{jt-1}\}$. We also allow labor supply to depend on firm-specific job amenities, $u_{jt}$, and a competition index $I_{gt}$ that varies across time $t$ and local labor markets $g$. This general labor supply function allows firms to make their labor input decisions conditional on other firms' actions, for example, as in oligopsonistic competition.[6]

At the chosen scale $L_{jt}$, the firm also selects the current workforce state $\widetilde{Z}_{jt}$ through recruitment policies, screening, training, retention, and job design, all of which may be costly. We summarize the current non-wage cost of employing effective labor and implementing this workforce state by

$$\Phi_{jt} = \Phi\left(L_{jt}, L_{j,t-1}, K_{j,t+1}, K_{jt}, Z_{jt}, \widetilde{Z}_{jt}, \widetilde{Z}_{j,t-1}, \Phi_j\right).$$

The cost function includes labor and capital adjustment costs, costs of changing workforce composition, and other hiring, dismissal, management, organizational, and congestion costs. It may

[6]We remain agnostic about the sources and nature of labor market imperfection for now. This general labor supply function nests many common theoretical frameworks from the literature (e.g., Card et al., 2018, Berger et al., 2022, and Chan et al., 2024).

therefore depend on the number and type of employees rather than only on ability-adjusted hours. We assume the current workforce state is not directly production relevant, conditional on $L_{jt}$, but it affects current non-wage costs and the firm's continuation value in period $t+1$. Firm-level heterogeneity in fixed costs or marginal cost efficiencies is captured by $\Phi_j$.

Firms make their employment decisions based on their information set in $t$ and their expectations over the ex-post productivity shock. We specify the firm's information set at the beginning of period $t$ as

$$\mathcal{I}_{jt} = \left\{ K_{jt}, L_{jt-1}, \widetilde{Z}_{jt-1}, \omega_{jt-1}, \eta_{jt}, Z_{jt}, P_t, u_{jt}, I_{gt}, \Phi_j \right\},$$

where $P_t$ is a vector of aggregate prices including the intermediate input price $P_t^M$ and capital investment price $P_t^K$. We assume firms take $P_t^M$ and $P_t^K$ as given.

## 2.3 The Problem of the Firm

Conditional on choosing to operate in period $t$, firms maximize their present discounted stream of current and future profits by choosing their capital investment $K_{jt}^I$, effective labor input $L_{jt}$, intermediates $M_{jt}$, and endogenous workforce characteristics $\widetilde{Z}_{jt}$. Firms thus face the following dynamic profit maximization problem:

$$\begin{aligned} V_{jt}(\mathcal{I}_{jt}) = \max_{L_{jt}, K_{jt}^I, M_{jt}, \widetilde{Z}_{jt}} & E_{\epsilon_{jt}}\left[R(K_{jt}, L_{jt}, M_{jt})e^{\nu_{jt}}\right] - \overline{W}_{jt}L_{jt} - \Phi_{jt} \\ & - P_t^K K_{jt}^I - P_t^M M_{jt} + \beta E_{\epsilon_{jt}, \eta_{jt+1}, P_{t+1}, I_{gt+1}}\left[V_{jt+1}(\mathcal{I}_{jt+1})\right] \end{aligned} \quad (4)$$

$$\begin{aligned} \text{s.t. } K_{jt+1} &= (1-\delta)K_{jt} + K_{jt}^I \\ L_{jt} &= L\left(\overline{W}_{jt}, Z_{jt}, \widetilde{Z}_{jt-1}, u_{jt}, I_{gt}\right) \\ \Phi_{jt} &= \Phi(L_{jt}, L_{jt-1}, K_{jt+1}, K_{jt}, Z_{jt}, \widetilde{Z}_{jt}, \widetilde{Z}_{jt-1}, \Phi_j), \end{aligned}$$

where the continuation value of the firm depends on the distributions of $\epsilon_{jt}$, $\eta_{jt+1}$, and the firm's expectations over future prices and the competition index. Solving the first-order condition (FOC) with respect to intermediates gives

$$\frac{\partial R(\cdot)}{\partial M_{jt}} e^{\omega_{jt}} \mathcal{E} = P_t^M, \quad (5)$$

where $\mathcal{E} \equiv E\left[e^{\epsilon_{jt}}\right]$ and the left-hand side is the expected marginal revenue product of intermediate inputs, or $\overline{MRPM}_{jt}$. This is a key equation used to identify the revenue production function. Deriving this equation relies only on the assumptions made in Section 2.1 and on firms taking the intermediate input price as given.

The FOC for labor equates the total expected marginal cost and marginal benefit of labor,

$$\underbrace{\overline{W}_{jt}\left(1+\frac{1}{\varepsilon^{L}_{\overline{W}_{jt}}}\right)+\frac{\partial \Phi_{jt}}{\partial L_{jt}}}_{\mathcal{MC}^{L}_{jt}}=\underbrace{\overline{MRPL}_{jt}+\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}}}_{\mathcal{MB}^{L}_{jt}},$$

where $\varepsilon^{L}_{\overline{W}_{jt}}$ is the labor supply elasticity with respect to the expected wage, $\overline{V}_{jt+1}$ is the expected discounted continuation value, and $\overline{MRPL}_{jt} \equiv \frac{\partial R}{\partial L_{jt}} e^{\omega_{jt}} \mathcal{E}$ is the expected MRPL.[7] The first term on the left is the monopsony marginal cost term, where expanding employment requires raising wages along an upward-sloping labor supply curve. But this is only one component of the firm's marginal cost schedule. Current non-wage employment costs contained in $\Phi_{jt}$ including adjustment, training, management, or congestion costs, also enter marginal cost. Holding all other state and choice variables fixed, the slope of the total marginal cost schedule is

$$\frac{\partial \mathcal{MC}^{L}_{jt}}{\partial L_{jt}}=\frac{\partial}{\partial L_{jt}}\left[\overline{W}_{jt}\left(1+\frac{1}{\varepsilon^{L}_{\overline{W}_{jt}}}\right)\right]+\frac{\partial^{2} \Phi_{jt}}{\partial L^{2}_{jt}}.$$

Thus, the labor supply curve and the associated monopsony wedge determine only the first component of the slope. Convex employment costs in $\Phi_{jt}$ can independently steepen the firm's marginal cost curve and widen markdowns. Accordingly, the $MC$ curve in Figure 1 should be interpreted as the firm's total current marginal cost of labor, not solely as a representation of monopsony power. On the benefit side, the marginal worker is worth both current MRPL and the shadow value of carrying that labor into the future. This continuation value is what can induce a firm to retain workers after an adverse productivity shock lowers current MRPL.

Solving the first-order condition for the expected wage and using $W_{jt} = \overline{W}_{jt} f^{c}(\epsilon_{jt})/\mathcal{E}^{c}$ gives the realized optimal wage,

$$W_{jt}=\frac{\varepsilon^{L}_{\overline{W}_{jt}}}{1+\varepsilon^{L}_{\overline{W}_{jt}}}\left(\overline{MRPL}_{jt}+\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}}-\frac{\partial \Phi_{jt}}{\partial L_{jt}}\right)\frac{f^{c}(\epsilon_{jt})}{\mathcal{E}^{c}}. \tag{6}$$

Note that while the expected wage and employment pair always lies on the labor supply curve, the realized wage may deviate from it after $\epsilon_{jt}$ is realized.

[7] See Appendix A.1 for derivations of the first-order conditions and key model equations.

Finally, writing $W_{jt} = \mu_{jt} MRPL_{jt}$ gives us our expression for the realized markdown,

$$\mu_{jt} = \frac{\varepsilon^{L}_{\overline{W}_{jt}}}{1+\varepsilon^{L}_{\overline{W}_{jt}}} \left[1 + \left(\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}}\right) \overline{MRPL}_{jt}^{-1}\right] \frac{f^{c}(\epsilon_{jt})}{\mathcal{E}^{c}} \frac{\mathcal{E}}{e^{\epsilon_{jt}}}. \tag{7}$$

Equation (7) separates three forces. The first factor is the standard static monopsony wedge generated by the finite labor supply elasticity. The term in brackets captures the net marginal value of labor: its expected future shadow value less its current non-wage marginal cost, scaled by expected MRPL. The remaining terms map the expected markdown into the realized markdown through the ex-post productivity and wage adjustments. Output-market conditions enter through MRPL and the continuation value rather than as a separate labor-market wedge.

This additional structure allows us to characterize which mechanisms drive variation in wages and markdowns. It also means that our subsequent decomposition assumes that the expected wage–employment pair is determined along the firm's labor supply curve, ruling out bargaining as a separate ex-ante wage-setting mechanism. The realized wage is not subject to this restriction because of the ex-post wage adjustment function, which may capture, among other mechanisms, bargaining over unexpected firm surplus.

Firms facing adjustment costs can therefore pay workers differently than a static monopsony model implies and, in some cases, can pay more than their current MRPL. Evaluated before the ex-post shock is realized, the expected wage exceeds expected MRPL if and only if

$$\frac{\overline{W}_{jt}}{\overline{MRPL}_{jt}} > 1 \quad \Longleftrightarrow \quad \frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}} > \frac{\overline{MRPL}_{jt}}{\varepsilon^{L}_{\overline{W}_{jt}}}.$$

The condition says that the shadow value of retaining the marginal unit of labor must exceed its current non-wage marginal cost by enough to offset the usual monopsony wedge. A firm may therefore optimally pay above current MRPL not because it lacks monopsony power, but because the worker's value to the firm extends beyond current production. Because this net marginal value enters the markdown relative to MRPL, firms with high MRPL may face large non-wage marginal costs in levels while remaining less constrained in proportional terms than firms with lower MRPL.

# 3 Data

We use linked firm- and worker-level administrative registers provided by Statistics Denmark containing detailed panel information on firm and worker characteristics.

**Worker-Level Data**. Our worker-level data come from the Integrated Database for Labor Market

Research, an annual register that covers the full Danish population from 1991. For each individual, we observe hourly and annual wages, hours and days worked across up to three job spells per year, as well as demographic and employment characteristics including age, gender, education, occupation, and position in the firm.

**Firm-Level Data**. We complement the worker data with the Firm Statistics Register (Statistics Denmark), which provides balance-sheet and production information for all private-sector firms in Denmark from 1996.[8] From this register we extract revenue, value added, the book value of the capital stock, intermediate input expenditures, and employment measured in full-time equivalents, along with firm age, location, and industry classification.

**Sample Selection and Characteristics**. We impose minimal sample selection criteria. To identify the distribution of worker ability and each worker's contribution to the firm's labor input, we use worker-level data between 1991 and 2010 covering employment, wages, and demographics. We consider workers aged 15 and older, keeping information on up to three job spells per year. After applying our firm selection criteria (below), this leaves us with 8.8 million worker-year observations. Appendix Table G.1 displays descriptive statistics of our sample of workers in terms of annual earnings, annual wages, hourly wages, and age distributions.

On the firm side, we drop observations with imputed variables, observations without price deflators (pre-1999), and observations without all the variables needed for estimating productivity (firms with missing capital, revenue, materials, labor, or firm-level labor characteristics for year $t$ or $t-1$). This leaves us with about 550,000 firm-level observations. After merging with the worker data, the full sample contains about 374,000 firm-year observations.[9] Appendix Table G.2 displays cross-sectional moments for various firm characteristics including employment, revenue, value added, value added per worker, and firm age. We express nominal values in 2010 prices using the Danish CPI.

**Danish Labor Market Features**. Denmark's "flexicurity" model combines relatively low costs of hiring and dismissing workers with generous income support for the unemployed (Andersen and Svarer, 2007), resulting in high worker mobility and short unemployment spells even during downturns (Andersen, 2021). During our sample period, wages for most workers were negotiated at the firm level rather than through centralized collective bargaining, and wage flexibility was comparatively high by international standards (Dahl et al., 2013).[10] This institutional environment

[8] The firm register begins in 1996 with coverage of the manufacturing sector. Other sectors were added from 1997–2000, with universal coverage achieved in 2000.

[9] To avoid the disclosure of any sensitive information, we have rounded the number of observations throughout the paper to the nearest thousand.

[10] Prior to the early 1990s, wage determination in Denmark was substantially more centralized through state-supported bargaining arrangements, but these institutions were largely dismantled during the subsequent labor market reforms.

with decentralized wage setting, flexible employment relationships, and high turnover makes Denmark a useful setting for studying firm-level wage determination and the role of labor adjustment frictions.

# 4 Estimating Wage Markdowns

In this section, we estimate the distribution of firm-level markdowns, defined as the ratio of ability-adjusted wages to the MRPL. We show that the structure of the model allows us to separately identify latent worker ability, firm-level wages, and firm-level productivity without solving the full structural model.

## 4.1 Estimating Wages

To estimate the firm-level wages, we first assume that workers are perfectly substitutable in production conditional on their ability. Firms offer the same wages per unit of ability, and workers' hourly wages take the following form: $W_{ijt} = A_{it} \times W_{jt}$. We aim to separately identify the distributions of the worker-level ability $(A_{it})$ and the firm-level wages $(W_{jt})$, neither of which is directly observed in the data.

We further assume that worker ability is given by $A_{it} = A_i \times \Lambda_t(X_{it})$ where $A_i$ denotes time-invariant innate ability and $\Lambda_t(X_{it})$ captures the contribution of observable worker characteristics, such as education or experience, with the contribution of each characteristic allowed to vary over time. We then exploit the log additivity of the wage form to estimate the worker and firm components of wages jointly:

$$w_{ijt} = a_i + \lambda_t(x_{it}) + w_{jt}, \tag{8}$$

where $a_i$ is a worker fixed effect. If we further assume that the time-varying worker characteristic function takes the form $\lambda_t(x_{it}) = \sum_k \lambda_{kt} x_{ikt}$, where $k$ indexes worker characteristics, the individual log-wage function in Equation (8) becomes $w_{ijt} = a_i + \sum_k \lambda_{kt} x_{ikt} + w_{jt}$. This wage function resembles the two-way fixed-effect model introduced by Abowd et al. (1999) (AKM hereafter). We use a similar strategy in our estimation and identification. Specifically, we estimate the individual wage equation as

$$w_{ijt} = \hat{a}_i + \sum_k \hat{\lambda}_{kt} x_{ikt} + \hat{w}_{jt} + \hat{\xi}_{ijt}, \tag{9}$$

where $w_{jt}$ is a time-varying firm effect and $\xi_{ijt}$ is a residual that we treat as measurement error.

Our specification differs from the standard two-way fixed effects regression model commonly used in the literature in two aspects. First, we use match-specific hourly wages rather than annual earnings, which allows us to include both full- and part-time spells without confounding the wage rate with hours worked. Second, we recover the firm-level wage as the time-varying firm effect in the wage equation.[11] We estimate the model in Equation (9) using the largest connected set of firm-time observations, which includes 94% of the firms and 99% of all of the workers in our original sample.[12] In our estimation, we consider a rich set of worker-level observables including dummies for occupation, education, and position within the firm (general employee, manager, etc), as well as continuous measures of labor market experience, most of which are typically absent in other administrative data sets. Following Card et al. (2013), we allow for the effect of education, occupation, and worker position within the firm to change over time to capture both the effects of time-varying observable individual characteristics and aggregate trends such as skill-biased technical change and outsourcing. We assume these are worker-level observables that are portable across firms.

Our estimates are in line with previous studies: We find that around 48% of the variance of the log hourly wages is accounted for by workers' observed and unobservable characteristics, whereas 16% is accounted for by firms' time-varying characteristics. The estimation of Equation (9) gives us two key distributions. First, a measure of each individual's ability $A_{it} = \exp(\hat{a}_i + \sum_k \hat{\lambda}_{kt} x_{ikt})$ which we use to construct the firm-level labor input $L_{jt}$ when we estimate firms' MRPL. Second, the firm's ability-adjusted hourly wage $W_{jt} = \exp(\hat{w}_{jt})$, which we use in our markdown estimation.

## 4.2 Estimating Revenue Productivity

We estimate the revenue production function using the flexible-intermediate-input approach of Gandhi, Navarro, and Rivers (2020) (GNR). The firm's FOC for intermediate inputs, Equation (5),

[11] As we further discuss in Appendix D, identification of Equation (9) requires that labor mobility is orthogonal to the residual $\xi_{ijt}$. Workers may nevertheless switch firms in response to shifts in the firm-time wage, observed firm characteristics, amenities, and recruiting policies. The time-varying firm effect therefore allows firm shocks to affect wages and worker allocation – a mechanism ruled out by the standard AKM model with time-invariant fixed effects. A few other papers also allow for time-varying coefficients in AKM regressions. The first example, to our knowledge, is Bagger et al. (2014) who allow for occupation-firm fixed and firm-time effects in their wage model. Bagger and Lentz (2019) also incorporate time-varying firm-level observables into their estimation. More recent examples include Chan et al. (2026), Engbom et al. (2023) and Lachowska et al. (2023) who also estimate a wage equation with time-varying firm effects. Our approach to controlling for labor quality in production estimation builds on Hellerstein and Neumark (2007) and Fox and Smeets (2011). For the identification of factor-augmenting technology in production functions, see Doraszelski and Jaumandreu (2018) and Demirer (2026).

[12] In Appendix D, we provide a graphical representation of how we construct our connected set using all job observations for each worker in the sample. Importantly, if we restrict our sample to only include the firm that provides the largest labor earnings for each worker (defined by total annual earnings), the largest connected set decreases in size, covering only 89% of firms.

implies

$$s_{jt} = \ln \mathcal{E} + \ln D(k_{jt}, l_{jt}, m_{jt}) - \epsilon_{jt} \equiv \ln D^{\mathcal{E}}(k_{jt}, l_{jt}, m_{jt}) - \epsilon_{jt}, \tag{10}$$

where $s_{jt} \equiv \ln(P_t^M M_{jt}/R_{jt})$ is the log share of intermediate-input expenditures in revenues, $D(k_{jt}, l_{jt}, m_{jt}) \equiv \partial r(k_{jt}, l_{jt}, m_{jt})/\partial m_{jt}$ is the revenue elasticity of intermediate inputs, and $D^{\mathcal{E}} \equiv \mathcal{E}D$. Because period-$t$ inputs are chosen before the ex-post shock $\epsilon_{jt}$ is observed, $\mathbb{E}[\epsilon_{jt} \mid k_{jt}, l_{jt}, m_{jt}] = 0$. The conditional mean of Equation (10) therefore identifies $D^{\mathcal{E}}$. Its residual identifies $\epsilon_{jt}$ and hence $\mathcal{E} \equiv \mathbb{E}[e^{\epsilon_{jt}}]$, which recovers $D$. Integrating $D$ with respect to intermediate inputs and using the Markov structure for persistent productivity in Equation (2) then identifies the remaining components of the revenue production function and the productivity terms $\omega_{jt}$ and $\eta_{jt}$. Appendix E.1 provides the full estimating equations and implementation details.

Our implementation differs from GNR and much of the literature in two ways that are central to our analysis. First, we construct labor as the ability-weighted sum of total hours: $L_{jt} = \sum_{i \in j} A_{it} H_{ijt}$. This avoids imposing homogeneous worker ability or perfect competition in the labor market, and it accounts for endogenous changes and cross-sectional differences in workforce ability, removing them from our estimates of firm productivity. Second, labor is a dynamic input chosen after the persistent productivity innovation $\eta_{jt}$ is observed, and is therefore endogenous in the second-stage production-function equation. We identify its contribution using predetermined capital and lagged labor and workforce states. In particular, lagged workforce composition $\widetilde{Z}_{j,t-1}$ is excluded from current production but shifts current employment through the firm's dynamic employment problem. This provides additional identifying variation relative to GNR, where labor is predetermined, and avoids treating labor as a fully flexible input.

We otherwise measure inputs and output following standard practice in the production-function literature (Syverson, 2011). The resulting estimates provide the non-parametric revenue production function, the input revenue elasticities, and the firm-level distributions of $\omega_{jt}$, $\eta_{jt}$, and $\epsilon_{jt}$. We then construct the MRPL and the firm-level markdown $\mu_{jt} = W_{jt}/MRPL_{jt}$, as well as worker-level marginal revenue products $MRPL_{ijt} = A_{it} MRPL_{jt}$.

**Discussion**

Identification of the wage equation and production function requires much weaker assumptions than those implied by the structure of the full model in Section 2.2. For the wage equation, the key assumptions are that workers are perfectly substitutable in production conditional on ability, and that firms pay the same wage per unit of ability within a firm-year.[13] These restrictions imply that wage differences within firms reflect differences in worker ability. Combined with worker

[13] Both of these restrictions could be easily relaxed to allow for multiple types of imperfectly substitutable workers with different wage equations.

mobility and the connected set, this gives us the variation needed to separately identify the worker and firm components of wages. These assumptions rule out worker-firm match effects that would be observationally confounded with ability or firm wages, but they do not require us to specify how the firm-level wage $W_{jt}$ is determined. For example, $W_{jt}$ may reflect labor market power, bargaining, rent-sharing, or firm-specific pay policies.

Similarly, identification of the production function does not require us to solve the firm's full dynamic problem. The key ingredients are the share equation for intermediates and the timing and price-taking assumptions on productivity and input markets, all standard in the production function literature. Intuitively, these assumptions allow us to separate true revenue productivity from endogenous input choices, and in our setting, from changes in the ability composition of the workforce. They do not require us to specify the exact form of the firm's cost or demand function, wage-setting policies, or even that labor and capital are chosen optimally in every period.

Since our approach identifies the revenue production function rather than the physical production function, we cannot recover physical productivity or returns to scale. However, we are able to identify the MRPL, our main object of interest, while allowing for both input and output market power. Our assumptions in Section 2.1 allow for markup variation that is either fixed at the estimation level or entirely absorbed by variation in inputs. Bias may arise if product-market heterogeneity affects revenue elasticities conditional on inputs, but our framework could be expanded to admit more flexible variable-markup systems such as those in Edmond et al. (2015) by including controls for variable demand elasticities (such as output market shares) when estimating revenue productivity.

# 5 Estimation Results

## 5.1 Wages, Productivity, and Markdowns

Our model estimates give us the distributions of firm-level wages ($W_{jt}$), worker-level ability ($A_{it}$), the input-revenue elasticities, and firm-level revenue productivity. We report key moments of these distributions in Panel A of Table 1.[14] The first two columns report the distributions of the estimated log wage and log MRPL. Firms at the 90th percentile of the wage distribution pay about 88% higher wages, conditional on ability, than those at the 10th percentile. Despite being estimated using very different data, the distributions of the MRPL and firm-level wage are similar, as we show in Figure 2a.[15] This similarity gives us a tight distribution of markdowns, which we calculate

[14] We report the distributions of revenue elasticities and returns to scale in Appendix Table G.3, and average markdowns and elasticities by economic sector in Appendix Table G.4.

[15] Firm-level wages are estimated using the worker-level matched employer-employee data, while the MRPL is estimated using the firm accounting data.

Table 1: MODEL ESTIMATES: ABILITY-ADJUSTED VS. NON-ABILITY-ADJUSTED

| | Wages and markdowns | | | Productivity | | |
|---|---|---|---|---|---|---|
| Statistic | $w_{jt}$ | $mrpl_{jt}$ | $\mu_{jt}$ | $\nu_{jt}$ | $\omega_{jt}$ | $\eta_{jt}$ |
| *Panel A. Ability-adjusted (baseline)* | | | | | | |
| Mean | 6.59 | 6.83 | 0.83 | — | — | 0.01 |
| Std. Dev. | 0.28 | 0.27 | 0.27 | 0.19 | 0.09 | 0.05 |
| p10 | 6.26 | 6.52 | 0.55 | −0.22 | −0.10 | −0.05 |
| p50 | 6.61 | 6.84 | 0.79 | 0.01 | 0.02 | 0.01 |
| p90 | 6.89 | 7.13 | 1.11 | 0.21 | 0.10 | 0.05 |
| *Panel B. Non-ability-adjusted* | | | | | | |
| Mean | 5.26 | 5.49 | 0.93 | — | — | 0.01 |
| Std. Dev. | 0.30 | 0.47 | 0.65 | 0.24 | 0.17 | 0.07 |
| p10 | 4.90 | 4.93 | 0.41 | −0.28 | −0.17 | −0.07 |
| p50 | 5.25 | 5.50 | 0.78 | −0.01 | 0.02 | 0.01 |
| p90 | 5.62 | 6.06 | 1.50 | 0.28 | 0.18 | 0.08 |

*Notes:* Table 1 compares cross-sectional moments of the model estimates under the baseline ability-adjusted specification (Panel A) and the non-ability-adjusted specification (Panel B). In Panel B, the firm-level wage $w_{jt}$ is the mean hourly wage and the production function is estimated using total hours of employment as the labor input of the firm. Wages ($w_{jt}$) and MRPL ($mrpl_{jt}$) are reported in log real DKK. The means of $\nu_{jt}$ and $\omega_{jt}$ are omitted as they are normalized. The sample includes 374,000 firm-year observations, rounded to the nearest thousand. Percentiles are calculated as the mean of adjacent milliles.

as $\mu_{jt} = W_{jt}/MRPL_{jt}$. We report the distribution of $\mu_{jt}$ in Column 3 and plot it as the solid line in Figure 2b. The mean markdown is 0.83, with the median firm paying workers 79% of their marginal product. While most firms have markdowns below one, around 18% pay their workers more than their marginal product.[16] Columns 4 through 6 report moments of the estimated productivity terms. Our approach provides relatively narrow productivity distributions, with a p90/p10 TFPR ($\nu_{jt}$) ratio for the entire economy of 1.54. This is comparatively low relative to other studies that find ratios around 1.9 even within narrowly defined manufacturing industries.[17]

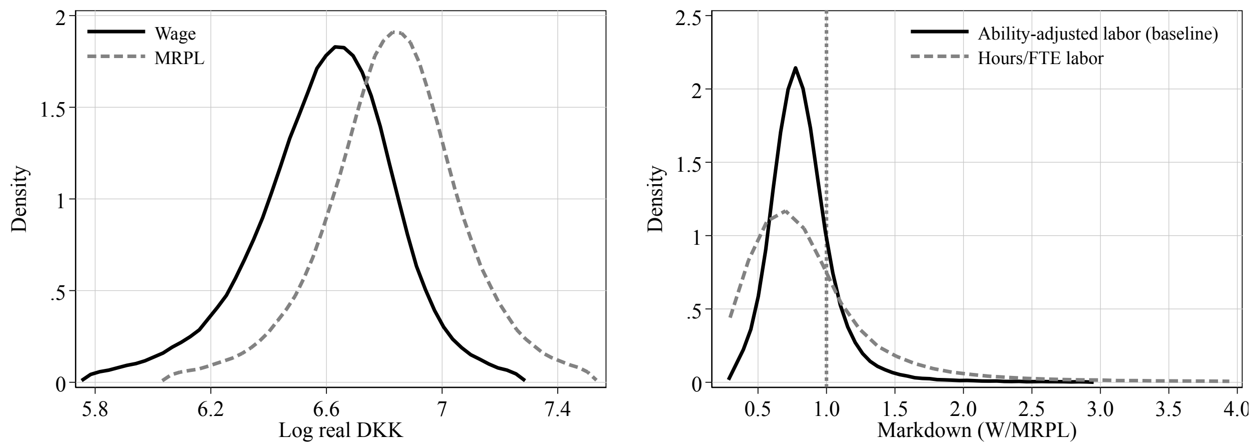
A key contribution of our approach is that we account for variation in latent worker-level ability ($A_{it}$) when estimating the revenue production function. This has first-order effects on measured productivity and markdown dispersion. Panel B of Table 1 presents the same model objects estimated under the assumption that workers are homogeneous and so $A_{it} = 1$ for all $i$. We find that controlling for latent ability significantly reduces the dispersion of our productivity and markdown estimates. Without the ability adjustment, the standard deviation of persistent productivity ($\omega_{jt}$)

[16] Other papers that estimate markdowns have found a similar (or larger) proportion of markdowns greater than one, typically attributed to measurement or mis-specification error. For example, Yeh et al. (2022) obtain a median inverse markdown (i.e., MRPL/W) of 1.364 (which is a markdown of 0.73) in the U.S. manufacturing sector, with an interquartile range of 0.6 and standard deviation of 0.7, implying a significant share of firms with markdowns greater than one. Similarly, Brooks et al. (2021) obtain a median inverse markdown of around 0.5 using Indian data, implying that the median firm pays workers twice their marginal product.

[17] For example, GNR find a mean p90/p10 ratio of 1.86 for 3-digit industries in the Colombian manufacturing sector, while Syverson (2011) reports an average ratio of 1.92 for 4-digit industries in the U.S. manufacturing sector.

Figure 2: WAGES, MARGINAL PRODUCTS, AND MARKDOWNS

(a) Ability-adjusted wage and MRPL (b) Markdown and ability correction




*Notes:* Figure 2 shows density plots for (log) firm wages, log MRPL, and the markdown. Figure 2a plots kernel densities of log firm-level wages per efficiency unit of labor ($w_{jt}$) and log MRPL. Figure 2b plots densities of the firm-level markdown, $\mu_{jt} = W_{jt}/MRPL_{jt}$, estimated using the ability-adjusted labor input and unadjusted hours. All density plots exclude the top and bottom percentiles, except for the markdown based on unadjusted hours, whose longer right tail is cropped for legibility.

nearly doubles from 0.09 to 0.17 and the persistent shock ($\eta_{jt}$) rises from 0.05 to 0.07. Controlling for ability also has significant implications for the estimated markdown distribution. Without the ability adjustment, the resulting markdown distribution is 140% more dispersed (a standard deviation of 0.65 vs 0.27), as we show in Figure 2b.[18] As with TFP, cross-sectional heterogeneity in labor quality drives a significant portion of the observed dispersion in markdowns. The uncorrected measure would also imply that an even larger share of firms (30%) pay workers more than their marginal revenue product.

While our identification strategy for the markdown is agnostic about how wages are set, the additional structure we place on the model in Section 2.2 provides two reasons why firms might set wages above the MRPL. First, if the expected future marginal benefit of labor is sufficiently greater than the current marginal cost of labor, firms may set wages higher than the MRPL to keep workers in the firm to reap those expected future continuation values. Second, ex-post productivity shocks may push worker wages above the firm's MRPL. To separate these effects, we construct the

[18]Our non-ability-corrected inverse markdown has a standard deviation of 0.77, similar to Yeh et al. (2022) though their sample is restricted to manufacturing, while our baseline inverse markdown has a standard deviation of 0.4.

expected markdown prior to observing $\epsilon_{jt}$,[19]

$$\overline{\mu}_{jt} = \frac{\varepsilon^L_{W_{jt}}}{1+\varepsilon^L_{W_{jt}}} \left( 1 + \frac{\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}}}{\overline{MRPL}_{jt}} \right), \tag{11}$$

which will be greater than 1 if $\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}} > \overline{MRPL}_{jt}/\varepsilon^L_{W_{jt}}$. This is the relevant markdown for the firm's employment decisions. We find that 16% of firms have $\overline{\mu}_{jt} > 1$. This suggests that ex-post wage adjustment accounts for very few markdowns above one; most arise from firm dynamics and adjustment costs. Intuitively, firms that are hit with a large negative persistent productivity shock would prefer to reduce their labor input and scale production down, lower wages, and increase the MRPL. However, if the shadow value of keeping the workers outweighs the costs, these firms may instead find it optimal to keep their labor input above the level which would be optimal if labor were a flexible input. We show evidence for this intuition in Section 5.3.

## 5.2 Markdown Heterogeneity

We plot the relationship between markdowns, wages, MRPL, and firm size in Figure 3. We find that markdowns have an inverse-U-shaped relationship with firm size. For small firms, markdowns increase from 0.78 for single-employee firms to 0.85 for firms with around 10 workers, while among the larger employers, markdowns decrease to 71% of MRPL on average among firms with 150 workers or more.[20] The employment- and revenue-weighted mean markdown is 0.7, implying that most workers receive slightly more than two-thirds of their marginal product. Figure 3b shows the source of this nonmonotonic pattern in markdowns. Firm-level wages increase monotonically with firm size, while the MRPL first decreases with size for small firms then increases for larger firms. The gap between MRPL and firm wages generates the inverse-U-shaped relationship between markdowns and size.

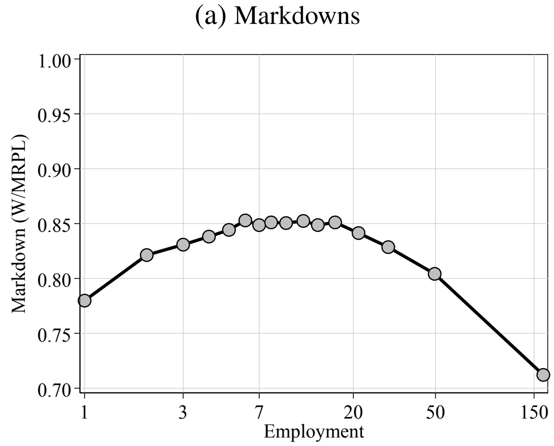
We also examine the relationship between markdowns and firm productivity, worker ability, and employment share in Figure 4. More productive firms tend to have substantially wider wage–MRPL gaps. Moving from the least to the most productive firms, the average markdown decreases from 0.96 to 0.74, a wage–MRPL gap of 22 percentage points (Figure 4a).[21] A similar, though

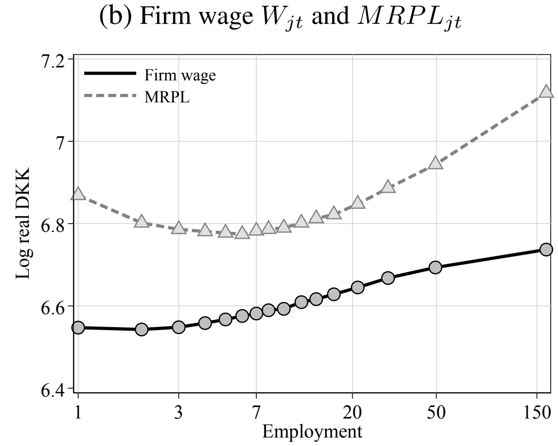
[19]Since all terms involving $\epsilon_{jt}$ enter the wage function multiplicatively, we can directly estimate $f^c(\epsilon_{jt})$ and $\mathcal{E}^c \equiv E[f^c(\epsilon_{jt})]$ and set $\overline{\mu}_{jt} = \mu_{jt} \frac{\mathcal{E}^c}{\mathcal{E}} \frac{e^{\epsilon_{jt}}}{f^c(\epsilon_{jt})}$. See Appendix E.2 for details. We also report the distributions of the expected wage, $\overline{W}_{jt}$, $\overline{MRPL}_{jt}$, and $\overline{\mu}_{jt}$ in Appendix Table E.2.

[20]Throughout our paper, an increase in markdown means workers are getting paid more relative to MRPL while a decrease in markdown indicates workers are getting paid less.

[21]Appendix Figure G.1 further shows that labor shares decline sharply with firm productivity (Panel a) and that markdowns widen as the labor share falls (Panel b), consistent with more productive firms being less labor intensive, and thus being less constrained by employment-related adjustment frictions.

Figure 3: MARKDOWNS, WAGES, AND MRPL BY FIRM SIZE




*Notes:* Figure 3 shows unconditional binscatter plots relative to bins of the log firm size (full-time equivalent employment) distribution. Figure 3a plots the within-bin mean of the firm-level markdown $\mu_{jt}$. Figure 3b plots the within-bin means of both log firm wages and log MRPL.

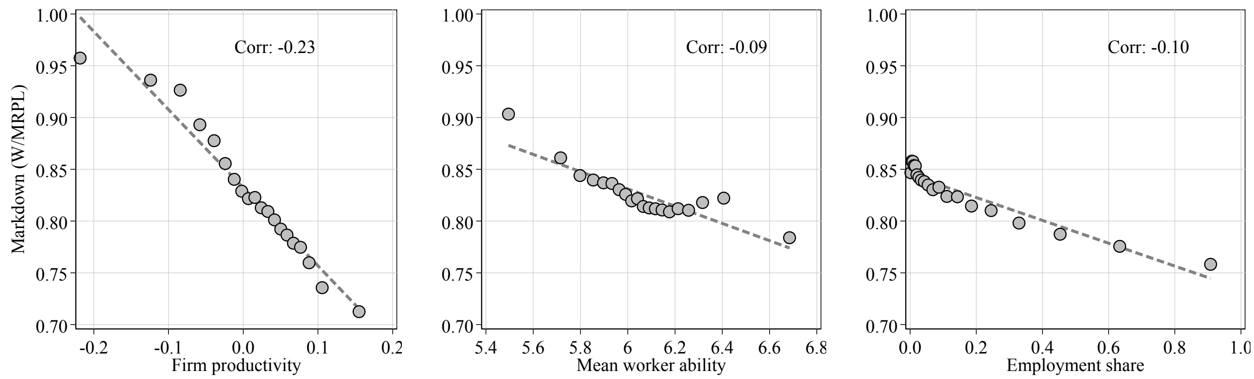
more moderate, pattern holds for mean worker ability (Figure 4b). Firms employing higher-ability workers pay a smaller share of MRPL, with markdowns declining from 0.90 to approximately 0.81 across the average worker ability distribution. Importantly, the relationships in Figures 4a and 4b are not driven solely by sorting of high-productivity firms employing higher-ability workers. Appendix Figure G.2 shows that markdowns vary with both productivity and worker ability even after conditioning on the other. Firms in the bottom decile of both distributions have an average markdown of 1.14, while those in the top decile of both have an average markdown of 0.72. Conditional on the productivity decile, markdowns generally decline as mean worker ability increases; the same pattern holds in the other direction, conditioning on ability.

Consistent with oligopsony models, firms with larger employment shares within their local labor market also have wider wage–MRPL gaps (Figure 4c).[22] The average markdown decreases from 0.85 for the smallest employers to 0.76 for those with the largest market shares. Taken together, these patterns show that the widest measured wage–MRPL gaps occur at large, productive firms with high-ability workforces and large local employment shares. This could reflect more productive and ability-intensive firms extracting larger rents from their workers, or facing greater variable employment costs that shift their marginal cost curve upward. Alternatively, dynamic concerns may constrain these firms below their optimal employment level. We examine these explanations in the following sections, and explicitly decompose the patterns in Figure 4 along each dimension in Section 6.

[22] Our definition of market share here is the firm-level mean employment share across all markets in which it hires in period $t$, with markets defined by 6-digit industry x municipality. This differs from the efficiency-unit based market share measure we use to estimate labor supply later in the paper.

Figure 4: MARKDOWNS BY FIRM CHARACTERISTICS

(a) Productivity ($\omega_{jt}$) (b) Mean Ability ($\overline{A}_{jt}$) (c) Employment Share




*Notes:* Figure 4 shows binscatter plots of firm-level markdowns by equal-sized groups of firm productivity ($\omega_{jt}$), mean worker ability ($\overline{A}_{jt}$), and employment share, based on a sample of approximately 374,000 firm-year observations. Productivity and average worker ability are in logs; markdowns and employment shares are in levels.

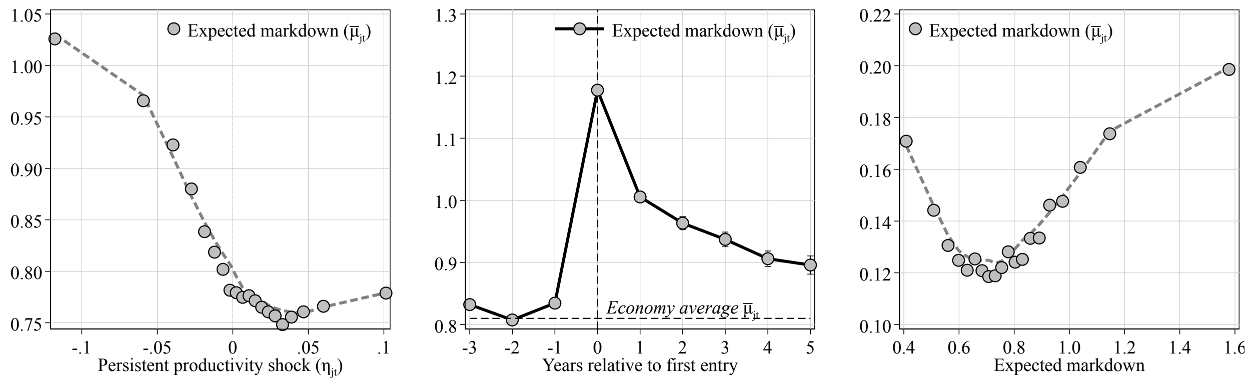
## 5.3 Adjustment Costs and Markdowns

The preceding results display features of the markdown distribution that static monopsony models do not account for: markdowns above one and a non-monotonic size–markdown relationship. In this section, we present evidence that these patterns reflect firms' dynamic response to productivity shocks in the presence of labor adjustment frictions.

**Productivity and Markdowns.** As shown in Section 2, forward-looking firms facing adjustment costs and uncertainty may set markdowns above or below the static monopsony level, particularly after large negative productivity shocks. Intuitively, in our model, firms that pay workers more than their marginal revenue product (i.e., $\mu_{jt} > 1$) are likely to have recently been hit by a large negative productivity shock. We test this by grouping firms by their persistent productivity shocks, $\eta_{jt}$, and plotting the average (expected) markdown within each bin in Figure 5a. We find that for firms experiencing positive productivity shocks, their markdown is around 0.78 and does not vary much with the magnitude of the shock. For firms experiencing negative shocks, the markdown is strongly negatively correlated with the magnitude of the shock. Firms experiencing the largest negative shocks have mean markdowns above one, suggesting that they are the most constrained, consistent with our model's predictions. This asymmetric relationship between productivity shocks and expected markdowns is consistent with models with convex adjustment costs: firms that need to shrink their workforce cannot do so quickly, and the resulting labor hoarding pushes wages above the now-lower marginal product. In Appendix B, we formalize this intuition and show that, given adjustment costs, there exists a threshold $\eta^*$ below which the firm retains more workers than the frictionless optimum and sets markdowns above one (Proposition 1 and Remark 2).

Figure 5: MARKDOWNS, PRODUCTIVITY SHOCKS AND EXIT RATES

(a) Markdown by $\eta_{jt}$ (b) Markdown persistence (c) Exit rate by markdown



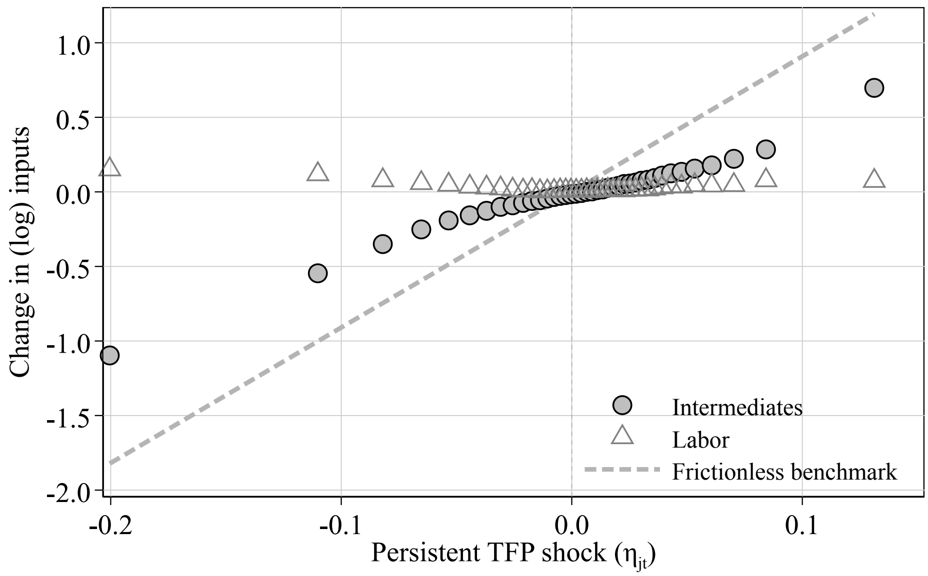

*Notes:* Figure 5a shows the average (expected) markdown $\overline{\mu}_{jt} = \overline{W}_{jt}/\overline{\text{MRPL}}_{jt}$ for each bin of the persistent firm-level productivity shock $\eta_{jt}$. Figure 5b traces the mean expected markdown from three years before to five years after a firm first enters the $\overline{\mu}_{jt} > 1$ region; the sample is restricted to clean entrants (firms observed with $\overline{\mu}_{jt} \leq 1$ in both $t-2$ and $t-1$), and the dashed line marks the economy-wide mean markdown (0.83). Figure 5c shows the next-period exit rate for each bin of $\overline{\mu}_{jt}$.

Importantly, if the asymmetry in Figure 5a reflects labor adjustment frictions, labor should respond less to productivity shocks than intermediate inputs, which can be adjusted more flexibly. Figure 6 shows exactly this pattern by plotting changes in log inputs against the persistent productivity shock. Intermediate inputs respond strongly and approximately symmetrically to positive and negative productivity innovations, with estimated slopes of about $5.1$ on both sides of zero. Labor is much less responsive: the corresponding slopes are $0.29$ for positive shocks and $-0.42$ for negative shocks. For scale, the dashed line in Figure 6 shows the response in a simple joint-frictionless Cobb–Douglas benchmark. With predetermined capital and both labor and intermediate inputs freely adjustable, the benchmark slope is $(1-\alpha_L-\alpha_M)^{-1} = 9.1$ when evaluated at our mean estimated revenue elasticities.[23] Relative to this benchmark, intermediate inputs move substantially in the frictionless direction, whereas labor barely adjusts, particularly following adverse shocks. This pattern is consistent with labor hoarding and costly labor adjustment as in Hopenhayn and Rogerson (1993), and suggests that the asymmetry in markdowns is not simply generated by a general inability of firms to adjust production, thus providing direct support for the labor-specific dynamic mechanism underlying the markdown response in Figure 5a.

**Markdown Persistence.** If a markdown above unity ($\overline{\mu}_{jt} > 1$) reflects constraints and firms' dynamic response to shocks, it should be a temporary state that cannot be sustained for long. We

[23] The $9.1$ slope is a joint-frictionless benchmark rather than the response that either input must attain independently. If one input is constrained, the optimal response of the other will generally differ from $9.1$ even when that input is itself fully flexible. For example, conditional on a labor response $\beta_L$, frictionless materials satisfy $\beta_M = (1+\alpha_L\beta_L)/(1-\alpha_M)$.

Figure 6: INPUT ADJUSTMENT TO PRODUCTIVITY SHOCKS




*Notes:* Figure 6 plots binned means of changes in log ability-adjusted labor and intermediate inputs against the persistent productivity innovation $\eta_{jt}$. Estimated slopes for intermediate inputs are approximately $5.1$ on both sides of zero; labor slopes are $0.29$ for positive shocks and $-0.42$ for negative shocks. The dashed line is the joint-frictionless Cobb–Douglas benchmark, $(1 - \alpha_L - \alpha_M)^{-1} = 9.1$, evaluated at the sample mean revenue elasticities $\alpha_L = 0.35$ and $\alpha_M = 0.54$.

test this conjecture by performing a simple event study, focusing on those firms that have just entered a constrained state ($\overline{\mu}_{jt} > 1$). Figure 5b shows the results. Before the event (years $-1$, $-2$, and $-3$), these firms have a mean expected markdown hovering around 0.8, close to the economy-wide mean. At the event ($t = 0$), the mean expected markdown spikes to 1.18, reflecting potential constraints and labor hoarding, then falls back to a markdown below one on average after one year and declines monotonically in subsequent periods. This pattern is consistent with the dynamic mechanism we have described: markdowns increase sharply in response to productivity shocks. If markdowns above unity reflected mostly measurement error, we would not expect to see this systematic pattern.

**Markdowns and Firm Exit.** If markdowns reflect firms' dynamic decisions and constraints, and are economically meaningful measures of firm distress, we should expect firms' exit rates to differ by their markdown levels. To examine this, we plot the relationship between markdowns and exit rates (which we define as the probability that a firm does not appear in our data in period $t + 1$) in Figure 5c. Exit rates display a V-shaped pattern in the level of markdowns. The exit rate is lowest among firms with $\overline{\mu}_{jt} \approx 0.73$ (firms paying workers 73% of their MRPL) and rises sharply as markdowns deviate from this level. This suggests a level of markdowns that minimizes the probability of exit, $\mu^*$, with deviations above or below it increasing the firm exit rate. If a firm is

Table 2: HIGH-MARKDOWN FIRMS: FINANCIAL CHARACTERISTICS AND GROWTH

| *Panel A. Financial characteristics by markdown group* | | |
|---|---|---|
| | $\mu_{jt} \leq 1$ | $\mu_{jt} > 1$ |
| Profit rate (profits/revenues) | 10% | −1% |
| Leverage (debt/assets) | 84% | 96% |

| *Panel B. Future outcomes of high-markdown firms* | | | | |
|---|---|---|---|---|
| | Revenue growth | | Profit growth | |
| | $g^{\text{sym}}(Rev)_{t+1}$ | $g^{\text{sym}}(Rev)_{t+1}$ | $g^{\text{sym}}(\Pi)_{t+1}$ | $g^{\text{sym}}(\Pi)_{t+1}$ |
| | (1) | (2) | (3) | (4) |
| $\mathbf{1}\{\mu_{jt} > 1\}$ | 0.031 | 0.056 | 0.338 | 0.583 |
| | (0.001) | (0.002) | (0.005) | (0.008) |
| Fixed effects | Year, industry | Year, firm | Year, industry | Year, firm |
| Observations | 287,000 | 273,000 | 287,000 | 273,000 |

*Notes:* Panel A reports mean financial characteristics by markdown group. Panel B reports regressions of one-period-ahead revenue and profit growth on an indicator for $\mu_{jt} > 1$. Symmetric growth is defined as $g^{\text{sym}}(x)_{t+1} = 2(x_{t+1} - x_t)/(|x_{t+1}| + |x_t|)$. Standard errors are reported in parentheses, and observation counts are rounded to the nearest thousand. Regressions using $\Delta \log Rev_{t+1}$ yield the same qualitative results as the symmetric revenue-growth specifications.

least distorted (relative to the monopsony baseline) and thus least likely to exit at $\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}} = 0$, then $\mu^*_{jt}$ should equal the monopsony wedge, $\frac{\varepsilon^L_{W_{jt}}}{1+\varepsilon^L_{W_{jt}}}$. As we will show in Section 6.2, our mean estimated monopsony wedge is very close to the level suggested here.[24] In particular, firms with $\mu_{jt} > 1$ experience the highest exit rates, reflecting a transitory, unsustainable state driven by adjustment frictions. Firms with smaller markdowns ($\mu_{jt} < \mu^*$) also experience larger exit rates, potentially because of very high variable employment costs or because the shadow value of labor is below its marginal cost, and firms employ less labor than they would at the optimum absent adjustment frictions.[25]

**Profits and leverage.** Our results indicate that periods with $\mu_{jt} > 1$ are transitory episodes in which firms hoard labor after negative shocks. Three testable implications follow: these firms should be less profitable, more leveraged, and, conditional on survival, exhibit higher subsequent profit growth as the constraint unwinds. We test these conjectures in Table 2.[26] In Panel A, we compare firms' profit rate and leverage (defined as debt divided by over assets) for those who are likely constrained ($\mu_{jt} > 1$) to those who are less constrained ($\mu_{jt} < 1$). We find that firms with $\mu_{jt} > 1$ have an average profit rate of $-1\%$, compared to $10\%$ for firms with $\mu_{jt} \leq 1$,

[24] In Section 6.2, we estimate a mean monopsony wedge close to 0.73, consistent with the least distorted firms having $\mu^{\phi}_{jt} = 1$ and therefore $\mu^*_{jt} = \mu^{\varepsilon}_{jt}$.

[25] This finding is consistent with Petrin and Sivadasan (2013) who also show that firm exit rates are correlated with the gap between the MRPL and wages.

[26] Profits are defined as revenues less the cost of capital (we use 8% of the capital stock value), intermediate expenditures, and labor costs. Because we use realized profits, these results also use realized rather than expected markdowns.

and are substantially more leveraged (96% vs. 84%), consistent with these firms being financially constrained after negative shocks.[27] In Panel B, we further examine these firms' revenue growth and profit growth. We run simple regressions of revenue and profit changes from period $t$ to $t+1$ on an indicator for $\mu_{jt} > 1$, controlling for firm characteristics and fixed effects, and find that firms with $\mu_{jt} > 1$ have significantly higher revenue and profit growth rates across all specifications.

To sum up, the markdown distribution exhibits features that static monopsony models alone do not account for, and the evidence presented above is consistent with firms facing labor adjustment costs. Critically, these empirical patterns do not depend on assumptions about how wages are set. In the next section, we impose stronger assumptions and model structure to decompose and quantify the contributions to the markdown due to monopsony power, employment/adjustment costs, and dynamics.

# 6 Quantifying Monopsony Power and Employment Costs

In this section, we decompose the markdown into two components: a *monopsony wedge*, $\mu^{\varepsilon}_{jt}$, driven by finite labor-supply elasticity, and a *cost wedge*, $\mu^{\phi}_{jt}$, driven by dynamic frictions and employment costs. We define

$$\mu^{\varepsilon}_{jt} = \frac{\varepsilon^{L}_{\overline{W}_{jt}}}{1+\varepsilon^{L}_{\overline{W}_{jt}}}, \qquad \text{and} \qquad \mu^{\phi}_{jt} = 1 + \left(\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}}\right) \overline{MRPL}^{-1}_{jt},$$

such that $\overline{\mu}_{jt} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}$. This decomposition requires additional structure, which we introduce next.

## 6.1 Estimating Labor Supply

To separately identify the monopsony wedge, $\mu^{\varepsilon}_{jt}$, from the cost wedge, $\mu^{\phi}_{jt}$, we impose additional structure on the supply of ability-adjusted labor. Recall that the firm's labor input is measured in efficiency units. Let $q_{it} \equiv A_{it}H_{it}$ denote worker $i$'s effective-labor endowment in period $t$, where $A_{it}$ is worker ability and $H_{it}$ is the worker's exogenous hours endowment. If worker $i$ is employed by firm $j$, the worker contributes $q_{it}$ units to $L_{jt}$. This is the same labor unit that enters production and the firm's wage bill in Section 2.

Worker $i$'s utility from working at firm $j$ is given by

$$U_{ijt} = f^{u}(\overline{W}_{jt}) + X_{jt}\beta + u_{jt} + \varepsilon^{u}_{ijt},$$

[27] We show in Appendix Figure G.3 that there is a strong positive relationship between leverage and markdowns, even when controlling for year, industry, employment, and revenues.

where $\overline{W}_{jt}$ is the expected wage per unit of ability-adjusted labor, $X_{jt} = \{Z_{jt}, \widetilde{Z}_{j,t-1}\}$ contains observed firm characteristics and the inherited workforce state, $u_{jt}$ captures firm-specific amenities common to all workers, and $\varepsilon^u_{ijt}$ is an idiosyncratic component which captures worker preferences and worker-specific firm hiring policies. The outside option is

$$U_{i0t} = f^u(W_{0t}) + \varepsilon^u_{i0t},$$

where $W_{0t}$ denotes the monetary value of non-employment. Workers choose the alternative that gives the highest utility. The vector $\boldsymbol{\varepsilon}^u_{it} = (\varepsilon^u_{i0t}, \{\varepsilon^u_{ijt}\}_{j=1}^{J_t})$ summarizes heterogeneity in nonpecuniary preferences, commuting costs, recruiting contacts, screening, job access, and other worker–firm-specific determinants of employment. Its joint distribution with worker characteristics is an equilibrium object shaped by workers' preferences and firms' hiring and retention policies.

The relevant population for the firm's labor-supply problem is the population of efficiency units. Let $Q_t \equiv \sum_i q_{it}$ denote total potential effective labor. The firm's effective-labor share is $s_{jt} \equiv L_{jt}/Q_t$, the outside share is $s_{0t} \equiv L_{0t}/Q_t$, and the within-market share is $s_{jgt} \equiv s_{jt}/\sum_{k \in g} s_{kt}$. All shares are therefore defined over efficiency units rather than worker counts.[28]

Following Lamadon et al. (2022) and Chan et al. (2024), conditional on the aggregate state and inherited firm states, we assume that the marginal distribution of $\boldsymbol{\varepsilon}^u_{it}$ under the effective-labor-weighted measure is nested generalized extreme value. Let $\lambda_{cz} \equiv 1/\sigma_{cz} \in (0,1)$ denote the nested-logit dissimilarity parameter, with firms in the same local labor market belonging to the same nest. We allow this parameter to vary across commuting zones. Worker characteristics may be correlated with the complete preference and access vector. This correlation determines the allocation of worker types across firms, while the nested-GEV marginal determines the total mass of effective labor allocated to each choice region. Appendix C states the joint-distribution restriction formally and derives the resulting aggregate shares.

The conventional individual-level nested logit is obtained when worker characteristics are independent of the preference and access vector under the effective-labor-weighted measure.[29] In the general case, the conditional distribution of worker characteristics varies across regions of preference and access space, generating systematic sorting in ability and other workforce characteristics.

[28] We can equivalently define an effective-labor-weighted probability $\Pr^q_t(B) \equiv Q_t^{-1} \sum_i q_{it} \Pr(B \mid i, t)$, so that $s_{jt} = \Pr^q_t(i \text{ chooses } j)$.

[29] Let $\nu^q_t$ denote the effective-labor-weighted population distribution of worker characteristics and let $K^q_t(dx \mid \boldsymbol{\varepsilon})$ denote their conditional distribution given the preference and access vector. The conventional model imposes $K^q_t(dx \mid \boldsymbol{\varepsilon}) = \nu^q_t(dx)$ for all $\boldsymbol{\varepsilon}$. Every worker type then has the same nested-logit choice probability, implying $P_{ijt} = P_{jt}$, $L_{jt} = Q_t P_{jt}$, and the same expected workforce composition at every firm. Appendix C provides the formal result.

Let

$$\delta_{jt} \equiv f^u(\overline{W}_{jt}) + X_{jt}\beta + u_{jt}, \qquad \delta_{0t} \equiv f^u(W_{0t}).$$

The nested-logit inversion for effective-labor shares is

$$\log s_{jt} - \log s_{0t} = f^u(\overline{W}_{jt}) - f^u(W_{0t}) + X_{jt}\beta + \left(1 - \frac{1}{\sigma_{cz}}\right) \log s_{jgt} + u_{jt}. \tag{12}$$

Equation (12) describes the allocation of total effective labor. The common wage index $f^u(\overline{W}_{jt})$ governs the response of aggregate effective labor to the firm's wage, while the joint distribution of worker characteristics and preference and access values governs workforce composition.

This scale–composition distinction maps directly into the firm's dynamic problem. The inherited workforce state $\widetilde{Z}_{j,t-1}$ may shift current labor supply and therefore enters $X_{jt}$. During period $t$, firms use recruiting, screening, training, retention, and job design to implement the workforce characteristics summarized by $\widetilde{Z}_{jt}$. These policies determine which worker types supply a given quantity of effective labor. Their non-wage implementation costs enter $\Phi_{jt}$, and the resulting workforce state enters the continuation value. Thus, Equation (12) governs the scale of effective labor, while $\widetilde{Z}_{jt}$ and $\Phi_{jt}$ govern the composition and cost of the workforce at that scale.

We define markets $g$ using municipality $\times$ four-digit-industry markets and construct all shares using ability-adjusted labor. We specify $f^u(\cdot)$ as a quadratic function of the log expected wage and include in $X_{jt}$ the same exogenous characteristics $Z_{jt}$ and lagged labor-force characteristics $\widetilde{Z}_{j,t-1}$ used elsewhere in the model.[30] We absorb variation in the outside option, including $f^u(W_{0t})$, with time effects. Since $u_{jt}$ is unobserved, we follow Chan et al. (2024) and assume that $u_{jt} = u_j + \widetilde{u}_{jt}$, where $\widetilde{u}_{jt}$ is i.i.d.. We then estimate Equation (12) in four-year differences, which removes the permanent firm component $u_j$. Because wages and within-market shares may be correlated with $\widetilde{u}_{jt}$, we use firm-level productivity innovations and characteristics of rival firms as instruments. Identification requires these excluded variables to be orthogonal to the four-year change in the unobserved supply component.[31] We also estimate a specification in which $f^u(\cdot)$ is linear in the log expected wage and obtain similar results.

[30]The aggregation result applies to any common differentiable wage index, including the quadratic specification used here. If pecuniary utility is instead derived from a common utility function over worker $i$'s actual wage $A_{it}\overline{W}_{jt}$, the log-linear specification has a direct separability interpretation because $\log(A_{it}\overline{W}_{jt}) = \log A_{it} + \log \overline{W}_{jt}$. A quadratic function of log actual earnings would generate an ability–wage interaction. We therefore interpret the quadratic baseline as a flexible common index of the efficiency-unit wage. Appendix C provides the formal discussion.

[31]The instrument set is $\eta_{jt}$, $\eta_{jt}^2$, $\Delta\omega_{jt}$, $\Delta\omega_{jt}^2$, one-period changes in the inside share interacted with commuting-zone indicators, and the four-year change in the sum of each observed firm characteristic among rivals in market $g$, $\sum_{k\neq j,\, k\in g} X_{kt}^h$, for every characteristic $h$ in $Z$ and $\widetilde{Z}_{t-1}$. These are the labor-market analogues of the instruments in Berry et al. (1995).

Because $Q_t$ is common across firms and fixed with respect to a unilateral change in firm $j$'s wage,

$$\frac{\partial \log L_{jt}}{\partial \log \overline{W}_{jt}} = \frac{\partial \log s_{jt}}{\partial \log \overline{W}_{jt}}.$$

The nested-logit share formulas imply

$$\frac{\partial \log s_{jt}}{\partial \delta_{jt}} = \sigma_{cz} + (1-\sigma_{cz})s_{jgt} - s_{jt}.$$

The firm-level labor-supply elasticity is therefore

$$\varepsilon^L_{Wjt} \equiv \frac{\partial \log L_{jt}}{\partial \log \overline{W}_{jt}} = \frac{\partial f^u(\overline{W}_{jt})}{\partial \log \overline{W}_{jt}} \left[\sigma_{cz} + (1-\sigma_{cz})s_{jgt} - s_{jt}\right], \tag{13}$$

and the monopsony wedge is

$$\mu^{\varepsilon}_{jt} = \frac{\varepsilon^L_{Wjt}}{1+\varepsilon^L_{Wjt}}.$$

The elasticity in Equation (13) is the elasticity of total effective labor conditional on the inherited workforce and recruiting state. This is the elasticity that enters the firm's wage-setting first-order condition in Equation (6). This specification is more flexible than the standard log-linear nested-logit or nested-CES specifications used in, for example, Berger et al. (2022) or Lamadon et al. (2022) because the elasticity can vary with both the firm's market shares and its wage.

We report moments of the distribution of firm-level labor supply elasticities, $\varepsilon^L_{\overline{W}_{jt}}$, and the implied monopsony wedge, $\mu^{\varepsilon}_{jt}$, in columns 2 and 3 of panel A of Table 3.[32] The mean labor supply elasticity is 2.87, with a mean monopsony wedge ($\mu^{\varepsilon}_{jt}$) of 0.74.[33] Column 1 reports moments of the expected markdown distribution from our production-side estimates, with a mean expected markdown of 0.81. If markdowns only represented static monopsony power, a mean markdown of 0.81 would imply a labor-supply elasticity of 4.26, which is approximately the 99th percentile of the estimated elasticity distribution.[34] This implies that the average firm is paying workers substantially

[32] We report the estimated model parameters and the resulting supply elasticity and monopsony wedge distributions for both the linear and quadratic specifications in Appendix Tables G.5 and G.6.

[33] The mean estimated labor supply elasticity is in line with the literature: Caldwell et al. (2026) find a mean estimated elasticity of 2.88 across 91 studies. We also estimate our labor supply elasticity simply by running a series of log-linear and polynomial OLS and IV regressions of effective employment on expected log wages, and find results in line with the nested logit model. The details of this exercise are presented in Appendix E.3.

[34] To assess whether the limited contribution of the monopsony wedge to markdown dispersion is an artifact of our labor-supply specification, Appendix E.4 reports a reverse-mapping exercise that does not use the nested-logit functional form. Under pure static monopsony, the markdown satisfies ($\mu = \varepsilon/(1+\varepsilon)$). The p10, p50, and p75 expected markdowns of 0.54, 0.77, and 0.92 would therefore require labor-supply elasticities of 1.17, 3.35, and 11.50, respectively, while the p90 markdown of 1.10 cannot be rationalized by any positive finite static labor-supply elasticity. We also perform the same inversion around large negative productivity shocks and show that attributing the associated

Table 3: ESTIMATED MARKDOWNS, LABOR SUPPLY ELASTICITIES, AND COST WEDGES

| *Panel A. Moments* | | | | |
|---|---|---|---|---|
| | Expected Markdown | Baseline LS specification | | |
| Statistic | $\overline{\mu}_{jt}$ | $\varepsilon^{L}_{W_{jt}}$ | $\mu^{\varepsilon}_{jt}$ | $\mu^{\phi}_{jt}$ |
| | (1) | (2) | (3) | (4) |
| Mean | 0.81 | 2.87 | 0.74 | 1.10 |
| Std. Dev. | 0.27 | 0.47 | 0.03 | 0.36 |
| p10 | 0.54 | 2.27 | 0.69 | 0.74 |
| p50 | 0.77 | 2.84 | 0.74 | 1.05 |
| p90 | 1.10 | 3.45 | 0.78 | 1.47 |
| *Panel B. Variance decomposition of log expected markdown* | | | | |
| Statistic | $\log(\overline{\mu}_{jt})$ | | $\log(\mu^{\varepsilon}_{jt})$ | $\log(\mu^{\phi}_{jt})$ |
| Share of variance (%) | 100.0 | – | 4.2 | 95.8 |

*Notes:* Panel A reports moments of the expected markdown, labor supply elasticity, monopsony wedge, and cost wedge under the baseline nested-logit labor supply specification. The expected markdown satisfies $\overline{\mu}_{jt} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}$. Panel B decomposes the variance of $\log(\overline{\mu}_{jt})$ into the labor-supply component and the dynamic cost component, with the covariance term ($cov(\log\mu^{\varepsilon}_{jt}, \log\mu^{\phi}_{jt}) = 0.002$) split equally between the two. The sample includes 374,000 firm-year observations, rounded to the nearest thousand. Percentiles are calculated as the mean of adjacent milliles. Appendix Table E.8 reports additional distributional moments.

more of their marginal product than it would in a static environment without adjustment frictions or other dynamic forces, in which $\overline{\mu}_{jt} = \mu^{\varepsilon}_{jt}$.

## 6.2 Monopsony Wedges and Cost Wedges

Given the estimated monopsony wedge, $\mu^{\varepsilon}_{jt}$, we can recover the cost wedge as the ratio of the expected markdown to the monopsony wedge:

$$\mu^{\phi}_{jt} \equiv \frac{\overline{\mu}_{jt}}{\mu^{\varepsilon}_{jt}} = 1 + \left(\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}}\right)\overline{MRPL}^{-1}_{jt}.$$

We report the cross-sectional moments of $\mu^{\phi}_{jt}$ in column 4 of Table 3. Deviations of the cost wedge from one capture the degree to which a firm is constrained above or below the static optimum quantity of labor. We find that the mean cost wedge is 1.10, which implies that, in equilibrium, the expected future marginal value of labor tends to outweigh the current marginal non-wage cost of labor, or $\frac{\partial \overline{V}_{jt+1}}{\partial L_{jt}} > \frac{\partial \Phi_{jt}}{\partial L_{jt}}$. However, this distribution is right-skewed, with a median cost wedge of only 1.05, reflecting that the median monopsony wedge of 0.74 is quite close to the median ex-

markdown movements entirely to monopsony would require very large one-period changes in firm-level labor-supply elasticities.

pected markdown of 0.77. There is significant dispersion in the net marginal value of labor. Firms at the 90th percentile value the marginal worker 47% more than their current marginal revenue product, while those at the 10th percentile value workers 26% below their MRPL.

The monopsony wedge accounts for the bulk of the average markdown level: as we show formally in Table 9, static monopsony accounts for roughly 72% of the aggregate wage–MRPL gap while dynamics account for the remaining 28%. In contrast, with a standard deviation of only 0.03, the monopsony wedge is unable to explain much of the *variation* in markdowns. Noting that $\log(\overline{\mu}_{jt}) = \log(\mu^{\varepsilon}_{jt}) + \log(\mu^{\phi}_{jt})$, we can decompose the variance of the log expected markdown into the components due to the monopsony and cost wedges. As we show in Panel B of Table 3, the monopsony wedge estimated from the nested-logit labor supply model only explains 4.2% of the overall variance in log expected markdowns, whereas the rest of the variance is accounted for by the cost wedge. We show in Appendix E.6 that using a calibrated version of the nested CES labor supply model used in Berger et al. (2022) provides similar results, explaining only 5.1% of the variance of the log markdown. We plot the distributions from both models in Appendix Figure E.1.

As a complementary reduced-form exercise, we ask how much markdown variation is predictable from a broad set of observables capturing firms' static labor-market position versus states associated with their dynamic labor-demand problem. Table 4 shows that static market position is substantially more informative about cross-sectional markdown levels than about movements in markdowns: across linear, polynomial, and spline specifications, static variables explain 6.8–8.5% of within-market dispersion in levels but only 1.1–1.3% of within-market variation in changes. Dynamic firm states explain substantially more in both cases, accounting for 12.3–16.0% of level dispersion and 11.7–13.6% of changes. Moreover, even if we attribute all within-market variation predicted by the broad static block to monopsony, it accounts for only 4.2–5.2% of total variance in markdown levels and less than 1% of variance in changes. These reduced-form patterns closely mirror the structural decomposition: static market power is associated with persistent differences in markdown levels, while dynamic firm states account for substantially more of their cross-firm dispersion and movement. Appendix E.7 provides the full exercise and additional diagnostics.

**Monopsony and cost wedges by firm characteristics.** Figure 7 decomposes the markdown profiles documented in Section 5.2 into the monopsony wedge $\mu^{\varepsilon}_{jt}$ and the cost wedge $\mu^{\phi}_{jt}$. Two findings stand out. First, the hump-shaped relationship between markdowns and firm size (Figure 3a) is driven entirely by the cost wedge (Figure 7a). The cost wedge rises from approximately 1.03 for the smallest firms to 1.15 for mid-sized firms, then declines back to 0.98 for the largest employers. The monopsony wedge, in contrast, is essentially flat across the size distribution at ap-

Table 4: Static Market Position and Dynamic Firm States

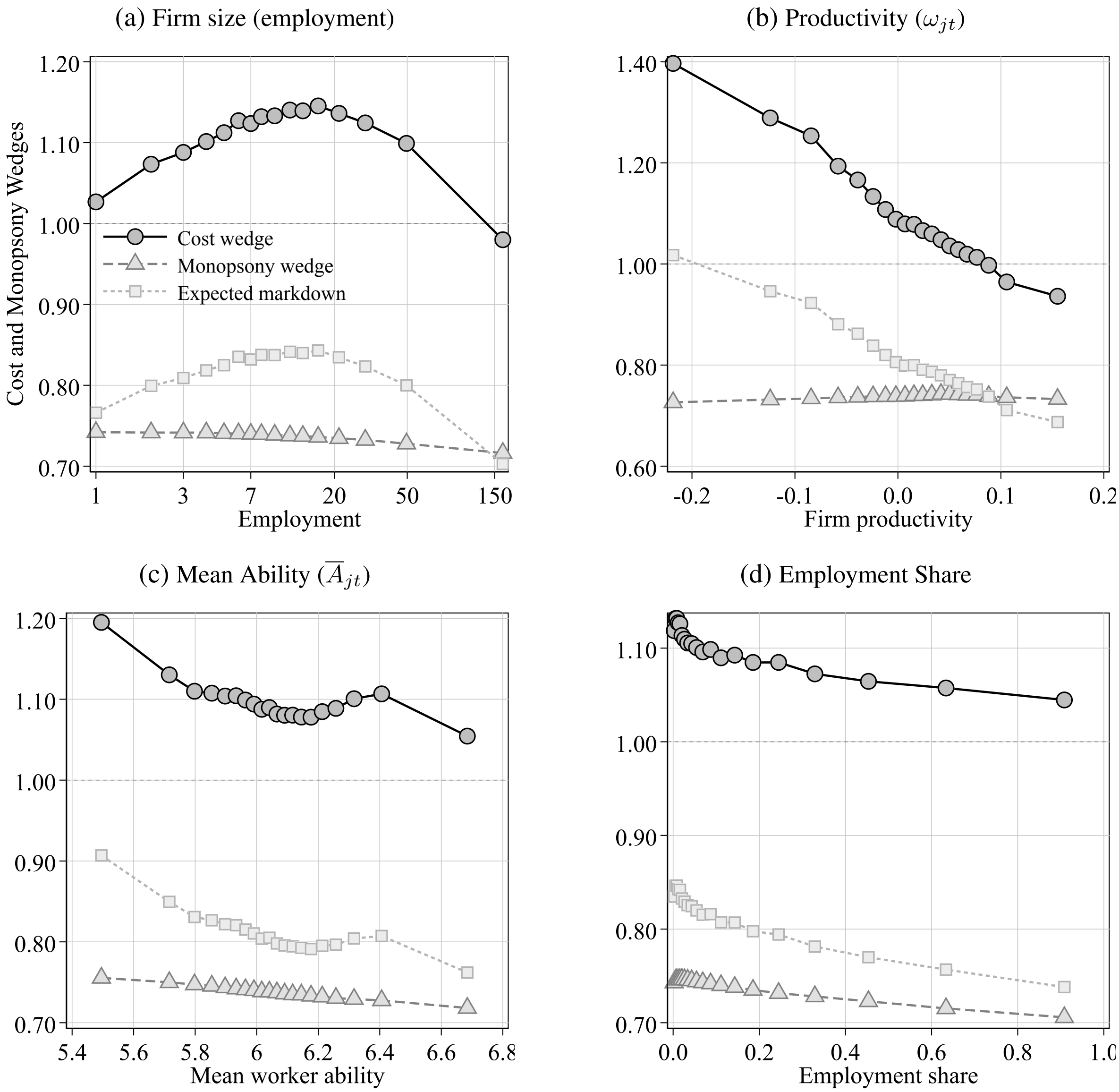
| | Levels | Changes |
|---|---|---|
| *Variance decomposition* | | |
| Between-market-year share of total variance | 38.33 | 31.01 |
| Within-market-year share of total variance | 61.67 | 68.99 |
| *Within-market $R^2$* | | |
| Static market-position block | 6.77–8.46 | 1.10–1.25 |
| Dynamic firm-state block | 12.28–15.99 | 11.68–13.60 |
| Both blocks | 22.59–26.01 | 13.84–16.91 |
| *Monopsony ceiling: share of total variance* | | |
| Within-market component, $s_W R_S^2$ | 4.17–5.22 | 0.76–0.86 |
| Total, $s_B + s_W R_S^2$ | 42.51–43.55 | 31.77–31.88 |

*Notes:* Entries are percentages. The static block contains firm shares of ability-adjusted labor and employment, corresponding within-market shares, wage-bill shares, firm size, market-structure measures, predetermined workforce characteristics, and interactions between firm market position and market structure. The dynamic block contains persistent productivity innovations, lagged persistent productivity, lagged employment, employment growth, capital, and the investment rate. Ranges report the minimum and maximum across linear, fourth-order polynomial, and five-knot restricted cubic spline specifications. The $R^2$ rows report within-market-year fit. Let $s_B$ and $s_W$ denote the between- and within-market-year variance shares, and let $R_S^2$ denote the within $R^2$ from the static-only specification. The within-market ceiling is $s_W R_S^2$; the total ceiling additionally assigns all between-market-year variation to monopsony and equals $s_B + s_W R_S^2$. These are accounting ceilings rather than identified structural variance shares. Appendix Table E.9 reports exact estimates and additional diagnostics.

proximately 0.74. Second, the negative gradient between markdowns and productivity (Figure 7b) is also driven by the cost wedge, which declines steeply from 1.40 for the least productive firms to 0.94 for the most productive. The monopsony wedge is again nearly constant.

This result is important because in most models of imperfect labor markets, the negative correlation between markdowns and productivity arises from less productive firms having smaller labor market employment shares and therefore facing more elastic labor supply. Our decomposition shows that this is not the only mechanism at work: monopsony power does not vary much with productivity. Instead, less productive firms have higher markdowns because (i) they face lower marginal employment costs (lower $\partial\Phi/\partial L$), (ii) they have a greater shadow value of labor (higher $\partial\overline{V}/\partial L$) consistent with these firms being more constrained after adverse shocks, or (iii) they have lower marginal products of labor and so the net marginal benefit of labor is larger relative to MRPL than it is at more productive firms. The bottom panels of Figure 7 show that the clean separation between the two wedges breaks down along dimensions that simultaneously proxy for market structure and firm dynamics. Both wedges decline with mean worker ability (Figure 7c) and employment share (Figure 7d), though neither dominates. Disentangling these channels requires the structural decomposition we undertake in Section 6.3.

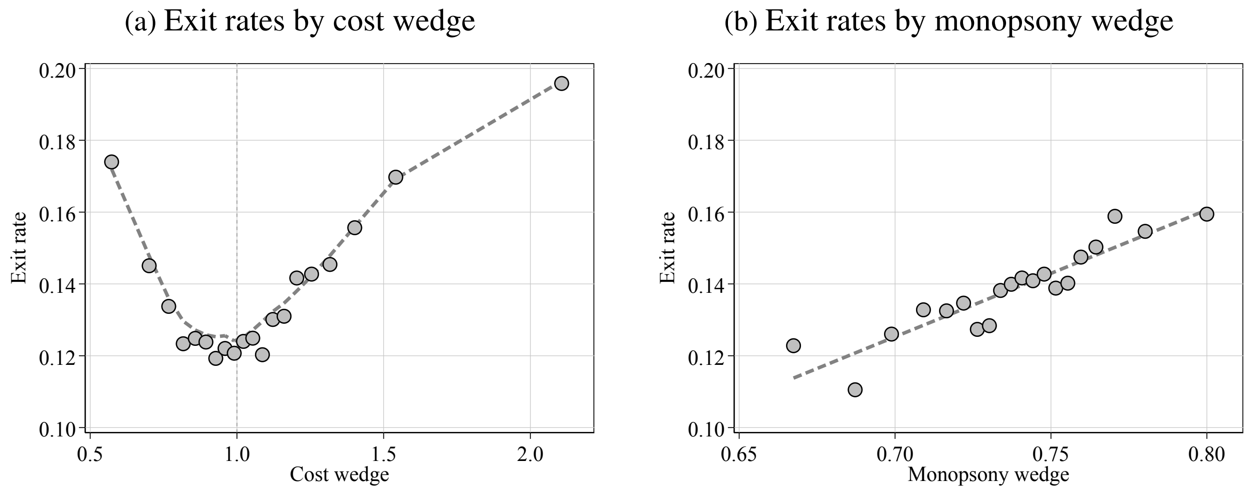
Figure 7: MARKDOWN, COST WEDGE, AND MONOPSONY WEDGE BY FIRM CHARACTERISTICS




*Notes:* Figure 7 shows the expected markdown ($\overline{\mu}_{jt}$), cost wedge ($\mu^{\phi}_{jt}$), and monopsony wedge ($\mu^{\varepsilon}_{jt}$) by bins of log firm size, productivity, mean worker ability, and employment share. The horizontal dashed line marks a value of one. The expected markdown satisfies $\overline{\mu}_{jt} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}$. The employment share is as defined in Section 5.2.

**Exit rates.** The cost wedge $\mu^{\phi}_{jt}$ captures the extent to which a firm is constrained away from its static monopsony optimum, where $\mu^{\phi}_{jt} = 1$ (see Section 2.3). This interpretation implies that firms further from $\mu^{\phi}_{jt} = 1$ should be more fragile and more likely to exit. Figure 8a strongly supports this prediction. Exit rates exhibit a clear V-shaped relationship with the cost wedge, with a minimum at $\mu^{\phi}_{jt} \approx 1$. Firms with high or low values of $\mu^{\phi}_{jt}$, corresponding to over- or under-employment, are significantly more likely to exit. This pattern holds when controlling for firm size, productivity, and capital (see Appendix Figure G.6), and is consistent with $\mu^{\phi}_{jt}$ capturing

Figure 8: EXIT RATES BY WEDGE COMPONENTS




*Notes:* Figures 8a and 8b plot the probability that a firm exits in $t+1$ (i.e., is not observed in the data in the subsequent period) against different wedge measures. The cost wedge is defined as $\mu^{\phi}_{jt}$ and the monopsony wedge as $\mu^{\varepsilon}_{jt}$. Exit rates are computed within bins of each variable.

economically meaningful constraints rather than, for instance, measurement error. In contrast, as shown in Figure 8b, exit rates increase monotonically with the monopsony wedge, as firms with $\mu^{\varepsilon}_{jt}$ closer to one tend to be smaller and less profitable, and are therefore more likely to exit. We show in Appendix Figure G.5 that the firms in the tails of the cost wedge distribution are also the most leveraged, which is consistent with increased risk of exit.

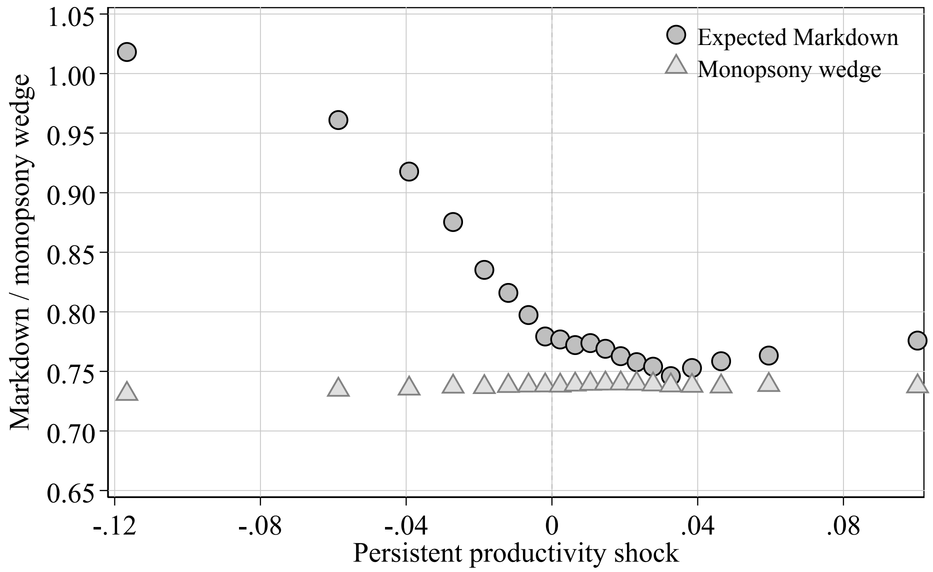
**Productivity shocks.** Finally, Figure 9 decomposes the relationship between markdowns and productivity shocks. The monopsony wedge is nearly flat across productivity shocks, confirming that the asymmetric response documented in Figure 5 is driven almost entirely by the cost wedge.

Overall, these results illustrate that the two components play distinct roles: monopsony power shapes the cross-sectional relationship between markdowns and market structure, while the cost wedge governs the dynamic response to productivity shocks. In the next section, we estimate the net marginal benefit function to determine how much of the variation in the cost wedge reflects endogenous responses to shocks versus fixed heterogeneity or measurement error.

## 6.3 Understanding Dynamic Labor Costs

We next ask whether the cost wedge is primarily a residual object or whether it is systematically related to the firm states that enter the model. If the cost wedge is explained by observed firm states and persistent firm heterogeneity, the dynamic component of the markdown reflects systematic features of firms' labor-demand problem rather than noise.

Figure 9: EXPECTED MARKDOWNS, MONOPSONY WEDGE, AND PRODUCTIVITY SHOCKS




*Notes:* Figure 9 plots the expected markdown and monopsony wedge across bins of the persistent productivity shock.

Our goal is to estimate the dynamic net marginal benefit term $B^{\phi}_{jt} \equiv \partial\overline{V}_{jt+1}/\partial L_{jt} - \partial\Phi_{jt}/\partial L_{jt}$ that enters the cost wedge defined in Section 6.2. To do so, we impose some additional structure on the cost function. We assume that the labor-relevant component of the cost function, $\Phi^L_{jt}$, depends on changes (in levels) in the labor input $\Delta L_{jt}$ and workforce composition terms $\Delta\widetilde{Z}_{jt}$, as well as the levels of labor $L_{jt}$, job amenities $u_{jt}$, workforce composition $\widetilde{Z}_{jt}$ and other predetermined firm characteristics $Z_{jt}$. We also allow firms to differ in their fundamental "cost efficiency" of employing labor, $\omega^{\phi}_j$, which captures the fact that some firms may simply be more efficient at managing or training labor than others.

With an additional assumption on measurement error in expected MRPL, we can obtain the following observation equation:

$$\widehat{B}^{\phi}_{jt} = D^{\phi}(\Delta L_{jt}, \Delta\tilde{Z}_{jt}, \overline{Z}_{jt}, \overline{L}_{jt+1}, L_{jt}, \tilde{Z}_{jt}, u_{jt}) + \omega^{\phi}_j + \xi^{ME}_{jt}. \tag{14}$$

Here $\widehat{B}^{\phi}_{jt}$ is the empirical counterpart of $B^{\phi}_{jt}$, computed as the residual wedge between the estimated firm average wage and estimated MRPL after accounting for the monopsony wedge: $\widehat{B}^{\phi}_{jt} \equiv \overline{W}_{jt}/\mu^{\varepsilon}_{jt} - \overline{MRPL}_{jt}$. Under the maintained measurement-error convention, $B^{\phi}_{jt} = D^{\phi}(\cdot) + \omega^{\phi}_j$ and $\widehat{B}^{\phi}_{jt} = B^{\phi}_{jt} + \xi^{ME}_{jt}$.

We estimate Equation (14) using both a parsimonious parametric specification and a flexible second-degree polynomial, proxying for $\overline{L}_{jt+1}$ using the firm's information set in period $t$.[35] We

[35] See Appendix E.5 for derivation and details.

Table 5: COST WEDGE AND ESTIMATED COST COMPONENTS

| *Panel A. Distribution of the cost wedge* | | | | | |
|---|---|---|---|---|---|
| Variable | Mean | Std. Dev. | p10 | p50 | p90 |
| $\mu^{\phi}_{jt} - 1$ | 0.10 | 0.36 | -0.26 | 0.05 | 0.47 |

| *Panel B. Variance decomposition* | | | | | |
|---|---|---|---|---|---|
| Object | Specification | Observables ($D^{\phi}$) (%) | Firm effect ($\omega^{\phi}_{j}$) (%) | Residual ($\xi^{ME}_{jt}$) (%) | Model ($D^{\phi} + \omega^{\phi}_{j}$) (%) |
| $\widehat{B}^{\phi}_{jt}$ | Parametric | 52 | 34 | 14 | 86 |
| | Polynomial | 64 | 24 | 12 | 88 |
| $\mu^{\phi}_{jt}$ | Parametric | 51 | 31 | 18 | 82 |
| | Polynomial | 52 | 31 | 17 | 83 |

*Notes:* Panel A reports selected moments of the cost wedge. Panel B reports the share of variance allocated to the fitted observable component, the firm fixed effect, and the residual. In all cases the covariance terms are split equally across the components involved. The empirical net marginal benefit object is $\widehat{B}^{\phi}_{jt} \equiv \overline{W}_{jt}/\mu^{\varepsilon}_{jt} - \overline{MRPL}_{jt}$. The estimated cost wedge satisfies $\mu^{\phi}_{jt} - 1 = \widehat{B}^{\phi}_{jt}/\overline{MRPL}_{jt}$. The model share is the sum of the observable and firm-effect components. The parametric specification follows the cost-function structure described in Appendix E.5, while the polynomial specification estimates the same relationship using a complete second-degree polynomial in the same variables.

then decompose the variance of the empirical object $\widehat{B}^{\phi}_{jt}$ and of the estimated cost wedge, $\mu^{\phi}_{jt} = 1 + \widehat{B}^{\phi}_{jt}/\overline{MRPL}_{jt}$, into the components due to observable firm state variables ($D^{\phi}$), time-invariant firm heterogeneity $\omega^{\phi}_{j}$, and the residual $\xi^{ME}_{jt}$.

Table 5 reports the main results. The variance decomposition shows that dispersion in the cost wedge is mainly systematic and driven by model-relevant objects. For the empirical net marginal benefit object $\widehat{B}^{\phi}_{jt}$, observed state variables explain 52% of the variance in the parametric specification and 64% in the polynomial specification. Adding the firm effect raises the explained share to 86–88%. The same conclusion holds after scaling by expected MRPL to form the estimated cost wedge. Observed states and firm effects together account for more than 80% of the variance of $\mu^{\phi}_{jt}$.

This decomposition then allows us to examine what drives dispersion and dynamics in the cost-related component of the markdown. In Figure 9, we show that markdown dynamics in response to productivity shocks act entirely through the cost wedge. In Appendix Figure G.4, we further decompose the response of the cost wedge to productivity shocks and find that it is driven entirely by within-firm variation in firm/worker characteristics rather than fixed firm characteristics.

## 6.4 Sources of Variation in the Cost Wedge

The previous subsection shows that the cost wedge $\mu_{jt}^{\phi}$ exhibits substantial and systematic variation. We now examine which economic mechanisms account for this variation. Guided by the model, we partition the firm states into blocks corresponding to distinct components of $\mu_{jt}^{\phi}$.

Formally, we estimate regressions of the form

$$\mu_{jt}^{\phi} = \alpha + X_{jt}^{\text{dyn}}\beta_{\text{dyn}} + X_{jt}^{\Delta}\beta_{\Delta} + X_{jt}^{\text{stock}}\beta_{\text{stock}} + X_{jt}^{\text{amen}}\beta_{\text{amen}} + X_{jt}^{\text{comp}}\beta_{\text{comp}} + X_{jt}^{\text{fin}}\beta_{\text{fin}} + \varepsilon_{jt}, \quad (15)$$

where each block of variables corresponds to a distinct economic channel in the model.

The dynamic block $X_{jt}^{\text{dyn}}$ includes productivity shocks and firm states that affect the expected future value of labor, specifically $\eta_{jt}$ (separately for positive and negative realizations), lagged productivity $\omega_{jt-1}$, capital, and investment. The adjustment-flow block $X_{jt}^{\Delta}$ includes changes in employment and workforce composition ($\widetilde{Z}_{jt}$), capturing realized contemporaneous adjustment costs. The stock block $X_{jt}^{\text{stock}}$ includes levels of employment, workforce composition, firm size, and firm age. The amenities block $X_{jt}^{\text{amen}}$ includes the firm-level amenity measure estimated in Section 6.1. The market structure block $X_{jt}^{\text{comp}}$ includes firm-level labor market share measures, capturing correlations of market structure with the cost wedge. The financial block $X_{jt}^{\text{fin}}$ includes leverage-based controls.

This specification is motivated by the structure of the cost function in Section 6.3, with each block corresponding directly to a component of $\partial V_{jt+1}/\partial L_{jt}$ or $\partial \Phi_{jt}/\partial L_{jt}$. We quantify the contribution of each block using a dominance decomposition based on regressions of $\mu_{jt}^{\phi}$ on all subsets of the state variables. This procedure averages the marginal contribution of each block to $R^2$ across all orderings of the remaining blocks and therefore provides an order-invariant measure of explanatory importance.

Table 6 yields three main results. First, variation in the cost wedge is predominantly associated with dynamic factors and persistent organizational features. As a fraction of the explained variation ($R^2 = 0.38$), the contribution of dynamics (45.9%) is nearly identical to that of static costs and organization (46.0%). In contrast, market structure and financial conditions together account for less than 10% of explained variation.

Second, forward-looking firm states play a central role. Variables that proxy for the expected future value of labor account for 44% of explained variation, by far the largest single component. This is consistent with $\mu_{jt}^{\phi}$ reflecting the dynamic value of labor and with labor hoarding and continuation-value effects in the model. By contrast, variables capturing contemporaneous adjustment flows contribute little, accounting for less than 2% of variation. This indicates that the cost wedge is not primarily driven by realized short-run hiring and firing flows, but rather by the lack of adjustment

Table 6: Sources of Variation in the Cost Wedge

| Block / component | Contribution to $R^2$ | Share (%) |
|---|---|---|
| *1. Dynamics* | | |
| Future value of labor | 0.165 | 44.0 |
| Workforce adjustment flows | 0.007 | 1.9 |
| *Subtotal* | 0.172 | 45.9 |
| *2. Static costs and organization* | | |
| Workforce stocks | 0.074 | 19.7 |
| Amenities | 0.099 | 26.3 |
| *Subtotal* | 0.173 | 46.0 |
| *3. Market structure* | | |
| Labor-market structure | 0.028 | 7.6 |
| *4. Financial conditions* | | |
| Financial variables | 0.002 | 0.6 |
| **Total** | **0.376** | **100.0** |

*Notes:* Table 6 reports general dominance statistics from regressions of the cost wedge, $\mu^{\phi}_{jt}$, on grouped firm state variables. Dynamics include variables related to the expected future value of labor and workforce adjustment. Static costs and organization include workforce composition, firm size, and amenities. Market structure includes labor market share variables. Financial conditions include leverage-based controls.

in the face of dynamic uncertainty and constraints.

Third, workforce stocks and amenities together account for 46% of explained variation, implying that firms with different organizational structures and job characteristics face significant variation in the marginal cost of maintaining a workforce.

Table 7 shows that the contribution of the future-value block is driven primarily by negative productivity shocks. Negative shocks alone account for more than one-third of the explanatory power of this block, substantially more than positive shocks or productivity. Capital, investment, and lagged productivity also contribute significantly, but none individually match the importance of negative shocks. This asymmetry is consistent with the theoretical example in Section 2.2 and Appendix Figure G.4, where negative shocks account for much of the variation in markdowns and the cost wedge while positive shocks have little effect. In the model, adjustment costs and continuation values can make firms retain labor after negative shocks, raising wages and markdowns relative to the static benchmark.

Taken together, these results reinforce the interpretation of the cost wedge as a dynamic object varying primarily with forward-looking firm states and persistent organizational features, rather than with static monopsony power. The asymmetric role of negative shocks further supports the view that the cost wedge captures economically relevant adjustment constraints.

Table 7: DECOMPOSITION OF THE FUTURE-VALUE COMPONENT

| Component | Contribution to $R^2$ | Share within block (%) |
|---|---|---|
| Negative productivity shocks ($\eta_{jt}^{-}$) | 0.061 | 37.0 |
| Capital and investment | 0.047 | 28.4 |
| Lagged productivity ($\omega_{jt-1}$) | 0.045 | 27.3 |
| Positive productivity shocks ($\eta_{jt}^{+}$) | 0.012 | 7.3 |
| **Total: future-value block** | **0.165** | **100.0** |

*Notes:* Table 7 reports variable-level dominance statistics within the future-value block. Contributions are normalized to sum to the total explanatory power of the block. Shares are expressed relative to the future-value block.

# 7 Implications for Misallocation and Markdown Accounting

The results above imply that observed markdowns are not sufficient statistics for labor market power, which in turn has important implications for the relation between measured markdowns, misallocation, and welfare. In this section, we quantify this point using three simple accounting exercises. Each exercise starts from the model decomposition

$$\overline{W}_{jt} = \overline{\mu}_{jt}\overline{MRPL}_{jt} = \mu_{jt}^{\varepsilon}\mu_{jt}^{\phi}\overline{MRPL}_{jt}, \tag{16}$$

where $\mu_{jt}^{\varepsilon}$ is the monopsony wedge and $\mu_{jt}^{\phi}$ is the dynamic cost wedge. Note that all of these exercises are local/static approximations in that they hold fixed the estimated revenue production function, firm states, and market environment, and ask how observed wage–MRPL gaps would be interpreted in standard wedge-accounting calculations. Appendix F gives the derivations and implementation details.

We first construct a Harberger-style revenue-loss index from the dispersion of markdowns. Let $z_{jt} = -\log \overline{\mu}_{jt}$, and let $\widetilde{z}_{jt}$ denote the deviation of the log markdown from its loss-weighted market-year mean, as defined in Appendix F. We focus on within-market dispersion in the wedge to abstract away from the common (or aggregate) markdown, which affects the division of surplus between firms and workers, but does not by itself reallocate labor across firms. The local loss index relevant for misallocation is therefore

$$\mathcal{L} = \frac{1}{2}\sum_{jt} \frac{\alpha_{jt}^{L}\overline{R}_{jt}}{\kappa_{jt}}\widetilde{z}_{jt}^{2}, \tag{17}$$

where $\kappa_{jt}$ is the local curvature of labor demand and $\alpha_{jt}^{L}\overline{R}_{jt} = \overline{MRPL}_{jt}L_{jt}$ scales the local loss by the marginal revenue value of labor at the firm. Since we can write $z_{jt} = z_{jt}^{\varepsilon} + z_{jt}^{\phi}$, the index

Table 8: LOCAL REVENUE LOSS FROM MARKDOWN DISPERSION

| Specification | Full Loss | Monopsony Share | Dynamic Share |
|---|---|---|---|
| CD | 0.76 | 3.98 | 96.02 |
| ACT | 5.87 | 3.97 | 96.03 |
| FLEX | 6.00 | 4.47 | 95.53 |

*Notes:* Table 8 reports the Harberger-style local revenue-loss index implied by within-market markdown dispersion. Full Loss is reported as a percentage of aggregate revenue. Monopsony and Dynamic shares are Shapley allocations of the monopsony and dynamic cost-wedge components of markdown dispersion. Markets are defined as municipality-industry-year cells. CD uses a Cobb-Douglas approximation $\kappa_j = 1 - \alpha^L_{jt}$. ACT uses the estimated labor-demand curvature implied by the nonparametric production function. FLEX allows intermediate inputs to adjust optimally when labor is reallocated.

can be decomposed into monopsony, dynamic-cost, and covariance terms. We report shares using a Shapley allocation that assigns one-half of the covariance term to each component.

Table 8 reports the results under three curvature measures. The level of the implied local revenue loss depends on curvature: a Cobb–Douglas approximation gives a loss of 0.76% of revenue, while the other two curvature measures implied by the estimated production function give losses close to 6% of revenue. The decomposition, however, is nearly invariant across specifications. Monopsony accounts for only 4.0–4.5% of the within-market markdown-dispersion loss, while the dynamic cost wedge accounts for roughly 96%. A wedge-accounting exercise that treats observed markdown dispersion as coming from monopsony power would attribute almost all of the implied loss to the wrong source.

Second, we decompose the aggregate wage–MRPL gap. Multiplying Equation (16) by ability-adjusted labor gives

$$L_{jt}(\overline{MRPL}_{jt} - \overline{W}_{jt}) = L_{jt}\overline{MRPL}_{jt}(1 - \mu^{\varepsilon}_{jt}) - L_{jt}\overline{MRPL}_{jt}\mu^{\varepsilon}_{jt}(\mu^{\phi}_{jt} - 1). \tag{18}$$

The first term on the right-hand side of the equation is the static monopsony benchmark. The second term is the contribution of dynamic costs and continuation values. Unlike the misallocation exercise above, this measure depends both on the aggregate markdown and on the correlation between markdowns and firm size. We report the aggregate gap both relative to the (potential) wage bill that would arise if firms paid workers their marginal product, $\sum_{jt} L_{jt}\overline{MRPL}_{jt}$, and relative to the actual wage bill, $\sum_{jt} L_{jt}\overline{W}_{jt}$.

Table 9 shows that the aggregate wage–MRPL gap is large: it is 38.6% of the potential wage bill and 62.9% of the actual wage bill. This implies that if workers were paid their marginal product, ceteris paribus, total payments to labor would increase by nearly 63%. Another way to interpret the potential wage bill gap is as the economy-wide mean markdown gap, $\mathbb{E}[1 - \overline{\mu}_{jt}]$, weighted

Table 9: AGGREGATE WAGE–MRPL GAP ACCOUNTING

| | Observed gap | | Share of observed gap | |
|---|---|---|---|---|
| Sample | Potential wage bill (%) | Actual wage bill (%) | Monopsony (%) | Dynamic cost (%) |
| All firms | 38.62 | 62.91 | 71.69 | 28.31 |
| Size $<$ 100 | 22.02 | 28.25 | 121.90 | -21.90 |
| Size $\geq$ 100 | 45.75 | 84.33 | 61.29 | 38.71 |

*Notes:* The observed gap is $\sum_{jt} L_{jt}(\overline{MRPL}_{jt} - \overline{W}_{jt})$. "Potential wage bill" denotes $\sum_{jt} L_{jt}\overline{MRPL}_{jt}$, the wage bill that would obtain if all workers were paid their marginal revenue product. "Actual wage bill" denotes $\sum_{jt} L_{jt}\overline{W}_{jt}$. The decomposition uses $\overline{W}_{jt} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}\overline{MRPL}_{jt}$, so that the monopsony component is $L_{jt}\overline{MRPL}_{jt}(1-\mu^{\varepsilon}_{jt})$ and the dynamic component is $-L_{jt}\overline{MRPL}_{jt}\mu^{\varepsilon}_{jt}(\mu^{\phi}_{jt}-1)$. A negative dynamic share means dynamic forces offset the static monopsony benchmark.

by $L_{jt}\overline{MRPL}_{jt}$. Static monopsony accounts for 71.7% of this aggregate gap, while the dynamic component accounts for the remaining 28.3%, though this decomposition depends on firm size. Among smaller firms, the dynamic component is negative, meaning that dynamic forces offset the static monopsony benchmark: these firms pay higher wages and employ more labor than the static monopsony benchmark would predict. Among larger firms, the dynamic component is positive and accounts for 38.7% of the observed gap. This pattern is consistent with the central result of the paper. Monopsony is important for the level of the markdown, but dynamic employment costs and continuation values materially change the mapping from wage gaps to labor market power.

Finally, we ask whether observed markdowns are useful policy targets. We rank firms by three objects: the full (inverse) markdown, the (inverse) monopsony wedge, and the absolute dynamic cost wedge $|\log \mu^{\phi}_{jt}|$. The full target selects firms with the widest observed wage–MRPL gaps. We then ask what fraction of those targeted firm-years are also selected by the monopsony or dynamic-cost target. Here, the target is the top decile of firms in each distribution. We report both economy-wide rankings and rankings within labor markets.

Table 10 shows that observed markdowns are a noisy measure of both underlying sources of the wedge. In the economy-wide ranking, only 23.6% of full-markdown targets are also monopsony targets, and 35.8% are dynamic-cost targets. We also weight firms by the marginal-product value of their labor input (the "potential wage bill") which is the relevant exposure measure for wage-gap and local misallocation accounting. This weighting lowers the overlap with monopsony to 18.1% while increasing the overlap with the dynamic cost wedge. Within-market targeting raises the unweighted overlap, especially for monopsony, but the potential-wage-bill weighted overlaps remain small. Weighting target firm-years by their potential wage bill, only 23.8% overlap with the within-market monopsony target and only 20.1% overlap with the within-market dynamic-cost

Table 10: TARGETING FIRMS BY OBSERVED MARKDOWNS AND STRUCTURAL COMPONENTS

| Conditioning target | Comparison target | Unweighted (%) | Potential wage bill weighted (%) |
|---|---|---|---|
| *Panel A. Economy-wide top-decile targets* | | | |
| Full markdown | Monopsony wedge | 23.56 | 18.12 |
| Full markdown | Dynamic cost wedge | 35.82 | 58.68 |
| *Panel B. Within-market top-decile targets* | | | |
| Full markdown | Monopsony wedge | 58.50 | 23.79 |
| Full markdown | Dynamic cost wedge | 44.13 | 20.07 |

*Notes:* Entries report the share of full-markdown target firm-years that are also selected by the comparison target. The full-markdown target selects firms with the widest observed wage–MRPL gaps. The monopsony target selects firms with the widest monopsony component. The dynamic cost target selects firms with the largest $|\log \mu^{\phi}_{jt}|$. Economy-wide targets are defined using economy-wide deciles. Within-market targets are defined within cells based on four-digit industry, municipality, and year. Potential-wage-bill weights are $L_{jt}\overline{MRPL}_{jt}$.

target. Thus, screening based on the observed markdown gap does not reliably identify the firms with the greatest labor market power or the strongest dynamic labor-demand constraints.

Taken together, these exercises reinforce the main message of the paper. While static monopsony explains a substantial share of the aggregate wage gap, the dynamic cost wedge drives nearly all of the cross-firm dispersion in markdowns. Unlike monopsony wedges, however, the dynamic component should not automatically be interpreted as a welfare distortion. Firms may be optimally hoarding labor or delaying hiring in the face of real adjustment costs and uncertainty, so the resulting markdown dispersion often reflects efficient dynamic responses rather than misallocation. Standard accounting exercises that interpret observed markdowns purely as employer market power will therefore misattribute the source of these gaps and potentially overstate the associated welfare losses.

# 8 Conclusion

This paper develops a framework for studying wage setting when firms face both upward-sloping labor supply and dynamic employment costs. Combining matched employer–employee data with firm accounts, we jointly recover worker ability, firm productivity, MRPL, and wage markdowns, and then use the firm's dynamic problem to separate the markdown into a static monopsony wedge and a cost wedge reflecting non-wage employment costs and the continuation value of labor. A central feature of the approach is that the production-side estimates can be recovered before imposing the structural wage-setting model, allowing us to first establish the empirical properties of markdowns and firm dynamics and only then quantify their underlying sources.

Two empirical findings are especially important. First, accounting for latent worker ability materially changes the measured distributions of productivity and markdowns. Treating workers as homogeneous substantially inflates the dispersion of both objects, mechanically assigning differences in workforce quality to firm productivity and MRPL. Second, the resulting markdown distribution exhibits a distinctly dynamic structure. Markdowns rise sharply following negative productivity shocks but respond little to positive shocks, labor adjusts much less than intermediate inputs, and episodes in which wages exceed current MRPL are temporary and concentrated among firms with low profits, high leverage, and elevated exit. These patterns point to an important role for the costs and continuation values associated with maintaining a workforce.

The structural decomposition produces a sharp quantitative result. Static monopsony provides a good benchmark for the level of wage markdowns: the monopsony wedge implied by the estimated labor-supply elasticities lies close to the center of the production-side markdown distribution, and monopsony accounts for roughly 72% of the aggregate wage–MRPL gap. But monopsony contributes very little to the dispersion of markdowns across firms. The cost wedge accounts for nearly all of that variation, and its movements are systematically related to firm states, workforce structure, amenities, and, especially, variables associated with the expected future value of labor. The broad conclusion is therefore not that monopsony is unimportant. Rather, monopsony and dynamic employment costs govern different margins of wage setting: the former is central for the average division of surplus between firms and workers, while the latter is central for understanding why wage–MRPL gaps differ so much across firms and over time.

This distinction also matters for how markdown heterogeneity enters aggregate accounting. In standard local misallocation calculations, roughly 96% of the within-market revenue-loss index generated by markdown dispersion is associated with the dynamic cost component rather than the monopsony component. These accounting losses need not correspond one-for-one to welfare losses, since dynamic wedges may reflect costly adjustment, labor hoarding, or the value of preserving a workforce under uncertainty. More generally, policies that increase workers' outside options and policies that reduce the costs of hiring, retaining, or reorganizing labor operate through different components of firms' wage-setting problem. Understanding their incidence therefore requires knowing not only how large wage–MRPL gaps are, but also why they arise.

The broader lesson is that labor market power and firm dynamics should be studied jointly. Labor supply determines how much wage-setting power firms have, while employment costs and continuation values determine how that power is exercised as firms respond to shocks and manage their workforces over time. Bringing these forces together provides a more complete account of wage determination, firm adjustment, and the cross-firm distribution of labor income.

## References


Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. Econometrica 67(2).

Agostinelli, F., D. Ferraro, G. Sorrenti, and L. Treuren (2025). Employment relationships, wage setting, and labor market power. Working Paper 34439, National Bureau of Economic Research.

Andersen, T. M. (2021). The danish labor market, 2000–2020. IZA World of Labor.

Andersen, T. M. and M. Svarer (2007). Flexicurity—labour market performance in denmark. CESifo Economic Studies 53(3).

Asker, J., A. Collard-Wexler, and J. De Loecker (2014). Dynamic inputs and resource (mis) allocation. Journal of Political Economy 122(5).

Azar, J. and I. Marinescu (2024). Monopsony power in the labor market: From theory to policy. Annual Review of Economics 16.

Bagger, J., B. J. Christensen, and D. T. Mortensen (2014). Productivity and wage dispersion: Heterogeneity or misallocation. Technical report, Working paper.

Bagger, J. and R. Lentz (2019). An empirical model of wage dispersion with sorting. The Review of Economic Studies 86(1).

Berger, D., K. Herkenhoff, and S. Mongey (2022). Labor market power. American Economic Review 112(4).

Berger, D., K. Herkenhoff, and S. Mongey (2026). Labor market power: From micro evidence to macro consequences. Journal of Economic Perspectives 40(1).

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. Econometrica: Journal of the Econometric Society, 841–890.

Bhuller, M., L. Delgado-Prieto, S. Hermo, and L. Lorentzen (2025). Wage-setting constraints and firm responses to demand shocks. Working Paper.

Brooks, W. J., J. P. Kaboski, Y. A. Li, and W. Qian (2021). Exploitation of labor? classical monopsony power and labor's share. Journal of Development Economics 150.

Caldwell, S., A. Dube, and S. Naidu (2026, August). Monopsony makes it big. Working Paper 35608, 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(S1).

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of west german wage inequality. The Quarterly Journal of Economics 128(3).

Chan, M., K. Kroft, E. Mattana, and I. Mourifié (2024). An empirical framework for matching with imperfect competition. Technical report, National Bureau of Economic Research.

Chan, M., S. Salgado, and M. Xu (2026). Heterogeneous passthrough from tfp to wages. Journal of Labor Economics.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. Handbook of econometrics 6.

Dahl, C. M., D. Le Maire, and J. R. Munch (2013). Wage dispersion and decentralization of wage bargaining. Journal of Labor Economics 31(3).

De Loecker, J. (2011). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. Econometrica 79(5).

Demirer, M. (2026). Production function estimation with factor-augmenting technology: An application to markups. Econometrica.

Dhyne, E., A. K. Kikkawa, T. Komatsu, M. Mogstad, and F. Tintelnot (2025). Firm responses and wage effects of foreign demand shocks with fixed labor costs and monopsony. American Economic Review 115(12).

Doraszelski, U. and J. Jaumandreu (2018). Measuring the bias of technological change. Journal of Political Economy 126(3).

Edmond, C., V. Midrigan, and D. Y. Xu (2015). Competition, markups, and the gains from international trade. American Economic Review 105(10).

Engbom, N., C. Moser, and J. Sauermann (2023). Firm pay dynamics. Journal of Econometrics 233(2).

Faia, E., B. Lochner, and B. Schoefer (2026). Monopsony, markdowns, and minimum wages. Working Paper 34699, National Bureau of Economic Research.

Fox, J. T. and V. Smeets (2011). Does input quality drive measured differences in firm productivity? International Economic Review 52(4).

Gandhi, A., S. Navarro, and D. A. Rivers (2020). On the identification of gross output production functions. Journal of Political Economy 128(8).

Hamermesh, D. S. (1989). Labor demand and the structure of adjustment costs. The American Economic Review 79(4).

Han, Q. (2024). Estimating labor market monopsony power from a forward-looking perspective. Economics Letters 244.

Hellerstein, J. K. and D. Neumark (2007). Production function and wage equation estimation with heterogeneous labor: Evidence from a new matched employer-employee data set. In Hard-to-measure goods and services: Essays in honor of Zvi Griliches. University of Chicago Press.

Hopenhayn, H. and R. Rogerson (1993). Job turnover and policy evaluation: A general equilibrium analysis. Journal of political Economy 101(5).

Klette, T. J. and Z. Griliches (1996). The inconsistency of common scale estimators when output prices are unobserved and endogenous. Journal of Applied Econometrics 11(4).

Kline, P. (2025). Labor market monopsony: Fundamentals and frontiers. In Handbook of Labor Economics, Volume 6, Chapter 8. Elsevier.

Lachowska, M., A. Mas, R. Saggio, and S. A. Woodbury (2023). Do firm effects drift? evidence from washington administrative data. Journal of Econometrics 233(2).

Lamadon, T., M. Mogstad, and B. Setzler (2022). Imperfect competition, compensating differentials, and rent sharing in the us labor market. American Economic Review 112(1).

Petrin, A. and J. Sivadasan (2013). Estimating lost output from allocative inefficiency, with an application to chile and firing costs. Review of Economics and Statistics 95(1).

Rubens, M. and B. S. Silveira (2026). Monopsony and backloaded compensation: Theory and evidence from public accountants. American Economic Review: Insights 8(1).

Seegmiller, B. (2025). Valuing labor market power: The role of productivity advantages. Working Paper, Kellogg School of Management.

Sokolova, A. and T. A. Sorensen (2021). Monopsony in labor markets: A meta-analysis. ILR Review 74(1).

Syverson, C. (2011). What determines productivity? Journal of Economic Literature 49(2).

Yeh, C., C. Macaluso, and B. Hershbein (2022). Monopsony in the us labor market. American Economic Review 112(7).

# Supplemental Online Appendix

# NOT FOR PUBLICATION

# A Model Details

In this section, we provide additional model details and derivations.

## A.1 Details of the First-Order Conditions

This appendix derives the first-order conditions underlying Equations (5), (6), and (7). The labor supply function is

$$L_{jt} = L\left(\overline{W}_{jt}, Z_{jt}, \widetilde{Z}_{j,t-1}, u_{jt}, I_{gt}\right),$$

where $\overline{W}_{jt}$ is the expected wage per unit of ability-adjusted labor. Assume that this labor supply function is continuous and strictly increasing in $\overline{W}_{jt}$. Let $G_j(\cdot)$ denote the corresponding inverse labor supply curve:

$$\overline{W}_{jt} = G_j\left(L_{jt}, Z_{jt}, \widetilde{Z}_{j,t-1}, u_{jt}, I_{gt}\right).$$

The realized wage satisfies

$$W_{jt} = W_{jt}^c f^c(\epsilon_{jt}), \qquad \overline{W}_{jt} = E_{\epsilon_{jt}}\left[W_{jt} \mid \mathcal{I}_{jt}, W_{jt}^c\right] = W_{jt}^c \mathcal{E}^c,$$

with

$$\mathcal{E}^c \equiv E_{\epsilon_{jt}}[f^c(\epsilon_{jt})], \qquad \mathcal{E} \equiv E_{\epsilon_{jt}}[e^{\epsilon_{jt}}].$$

The firm's period-$t$ information set is

$$\mathcal{I}_{jt} = \left\{K_{jt}, L_{j,t-1}, \widetilde{Z}_{j,t-1}, \omega_{j,t-1}, \eta_{jt}, Z_{jt}, P_t, u_{jt}, I_{gt}, \Phi_j\right\}.$$

Current persistent productivity is

$$\omega_{jt} = h(\omega_{j,t-1}) + \eta_{jt},$$

and total revenue productivity is $\nu_{jt} = \omega_{jt} + \epsilon_{jt}$, where $\epsilon_{jt} \notin \mathcal{I}_{jt}$. Define the expected discounted continuation value as

$$\overline{V}_{j,t+1} \equiv \beta E_{\epsilon_{jt}, \eta_{jt+1}, P_{t+1}, I_{g,t+1}}\left[V_{j,t+1}(\mathcal{I}_{j,t+1}) \mid \mathcal{I}_{jt}, L_{jt}, \widetilde{Z}_{jt}, K_{jt+1}\right].$$

After substituting the inverse labor supply curve, the firm's problem is

$$\begin{aligned} V_{jt}(\mathcal{I}_{jt}) = \max_{L_{jt}, K_{jt}^I, M_{jt}, \widetilde{Z}_{jt}} \quad & R(K_{jt}, L_{jt}, M_{jt}) e^{\omega_{jt}} \mathcal{E} - G_j\left(L_{jt}, Z_{jt}, \widetilde{Z}_{j,t-1}, u_{jt}, I_{gt}\right) L_{jt} \\ & - \Phi\left(L_{jt}, L_{j,t-1}, K_{j,t+1}, K_{jt}, Z_{jt}, \widetilde{Z}_{jt}, \widetilde{Z}_{j,t-1}, \Phi_j\right) \\ & - P_t^K K_{jt}^I - P_t^M M_{jt} + \overline{V}_{j,t+1} \end{aligned} \tag{19}$$

subject to

$$K_{j,t+1} = (1-\delta)K_{jt} + K_{jt}^I.$$

**Materials.** Materials are flexible, do not enter the non-wage cost function, and are not a state variable. The FOC for $M_{jt}$ is therefore

$$\frac{\partial R(K_{jt}, L_{jt}, M_{jt})}{\partial M_{jt}} e^{\omega_{jt}} \mathcal{E} = P_t^M.$$

This is Equation (5) in the main text. The left-hand side is the expected marginal revenue product of materials. This is the only firm optimality condition used for production-function identification.

**Labor.** Let

$$\overline{MRPL}_{jt} \equiv \frac{\partial R(K_{jt}, L_{jt}, M_{jt})}{\partial L_{jt}} e^{\omega_{jt}} \mathcal{E}$$

denote the expected marginal revenue product of ability-adjusted labor. The labor FOC is

$$0 = \overline{MRPL}_{jt} - \left[ G_j(.) + L_{jt} \frac{\partial G_j(.)}{\partial L_{jt}} \right] - \frac{\partial \Phi_{jt}}{\partial L_{jt}} + \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}}. \tag{20}$$

The firm-level labor supply elasticity with respect to the expected wage is

$$\varepsilon^L_{\overline{W}jt} \equiv \frac{\partial \log L(\overline{W}_{jt}, Z_{jt}, \widetilde{Z}_{j,t-1}, u_{jt}, I_{gt})}{\partial \log \overline{W}_{jt}}.$$

Since $G_j(\cdot)$ is the inverse labor supply curve,

$$L_{jt} \frac{\partial G_j(.)}{\partial L_{jt}} = \frac{\overline{W}_{jt}}{\varepsilon^L_{\overline{W}jt}}.$$

Using $\overline{W}_{jt} = G_j(.)$, Equation (20) becomes

$$\overline{W}_{jt} \left( 1 + \frac{1}{\varepsilon^L_{\overline{W}jt}} \right) = \overline{MRPL}_{jt} + \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}}.$$

Hence

$$\overline{W}_{jt} = \frac{\varepsilon^L_{\overline{W}jt}}{1 + \varepsilon^L_{\overline{W}jt}} \left( \overline{MRPL}_{jt} + \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}} \right). \tag{21}$$

Because $W_{jt} = \overline{W}_{jt} f^c(\epsilon_{jt}) / \mathcal{E}^c$, the realized wage is

$$W_{jt} = \frac{\varepsilon^L_{\overline{W}jt}}{1 + \varepsilon^L_{\overline{W}jt}} \left( \overline{MRPL}_{jt} + \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}} \right) \frac{f^c(\epsilon_{jt})}{\mathcal{E}^c}, \tag{22}$$

which is Equation (6) in the main text.

Define the net marginal benefit, or continuation-value, of labor as

$$B^{\phi}_{jt} \equiv \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}}.$$

Then Equation (21) can be written compactly as

$$\overline{W}_{jt} = \frac{\varepsilon^L_{\overline{W}jt}}{1 + \varepsilon^L_{\overline{W}jt}} \left( \overline{MRPL}_{jt} + B^{\phi}_{jt} \right).$$

The realized marginal revenue product is

$$MRPL_{jt} = \frac{\partial R(K_{jt}, L_{jt}, M_{jt})}{\partial L_{jt}} e^{\omega_{jt}+\epsilon_{jt}},$$

so

$$\overline{MRPL}_{jt} = MRPL_{jt} \frac{\mathcal{E}}{e^{\epsilon_{jt}}}.$$

The realized markdown is therefore

$$\mu_{jt} \equiv \frac{W_{jt}}{MRPL_{jt}} = \frac{\varepsilon^{L}_{\overline{W}jt}}{1+\varepsilon^{L}_{\overline{W}jt}} \left(1 + \frac{B^{\phi}_{jt}}{\overline{MRPL}_{jt}}\right) \frac{f^{c}(\epsilon_{jt})}{\mathcal{E}^{c}} \frac{\mathcal{E}}{e^{\epsilon_{jt}}}, \tag{23}$$

which is Equation (7) in the main text. The expected markdown relevant for the firm's employment decision is

$$\overline{\mu}_{jt} \equiv \frac{\overline{W}_{jt}}{\overline{MRPL}_{jt}} = \frac{\varepsilon^{L}_{\overline{W}jt}}{1+\varepsilon^{L}_{\overline{W}jt}} \left(1 + \frac{B^{\phi}_{jt}}{\overline{MRPL}_{jt}}\right). \tag{24}$$

**Workforce characteristics.** The direct choice of $\widetilde{Z}_{jt}$ represents the workforce state implemented through recruiting, screening, training, retention, and organizational policies. By construction, $\Phi_{jt}$ includes the minimized current cost of implementing this state. Current production depends on worker composition only through total effective labor $L_{jt}$, and current aggregate labor supply depends on the inherited state $\widetilde{Z}_{j,t-1}$ rather than the contemporaneous composition choice. Hence, holding $L_{jt}$ fixed, the first-order condition is

$$\frac{\partial \Phi_{jt}}{\partial \widetilde{Z}_{jt}} = \frac{\partial \overline{V}_{j,t+1}}{\partial \widetilde{Z}_{jt}}.$$

The condition equates the current marginal implementation cost of changing the workforce state to its expected discounted continuation value.

**Capital investment.** Since $K_{j,t+1} = (1-\delta)K_{jt} + K^{I}_{jt}$, the FOC for investment is

$$\frac{\partial \overline{V}_{j,t+1}}{\partial K_{j,t+1}} = P^{K}_{t} + \frac{\partial \Phi_{jt}}{\partial K_{j,t+1}}. \tag{25}$$

If current non-wage costs do not include capital adjustment or installation costs, then

$$\frac{\partial \Phi_{jt}}{\partial K_{j,t+1}} = 0$$

and the condition reduces to

$$\frac{\partial \overline{V}_{j,t+1}}{\partial K_{j,t+1}} = P^{K}_{t}.$$

**Envelope conditions.** If lagged labor enters the period-$t$ problem only through the non-wage cost function, then

$$\frac{\partial V_{jt}(\mathcal{I}_{jt})}{\partial L_{j,t-1}} = -\frac{\partial \Phi_{jt}}{\partial L_{j,t-1}}. \tag{26}$$

The envelope condition for predetermined capital is

$$\begin{aligned}\frac{\partial V_{jt}(\mathcal{I}_{jt})}{\partial K_{jt}} &= \frac{\partial R(K_{jt}, L_{jt}, M_{jt})}{\partial K_{jt}} e^{\omega_{jt}} \mathcal{E} - \frac{\partial \Phi_{jt}}{\partial K_{jt}} + (1-\delta)\left(\frac{\partial \overline{V}_{j,t+1}}{\partial K_{j,t+1}} - \frac{\partial \Phi_{jt}}{\partial K_{j,t+1}}\right) \\ &= \frac{\partial R(K_{jt}, L_{jt}, M_{jt})}{\partial K_{jt}} e^{\omega_{jt}} \mathcal{E} - \frac{\partial \Phi_{jt}}{\partial K_{jt}} + (1-\delta) P_t^K, \end{aligned} \tag{27}$$

where the second line uses the investment FOC (25). Since materials are fully flexible and not a state variable,

$$\frac{\partial V_{jt}}{\partial M_{j,t-1}} = 0.$$

---

In summary:

$L_{jt}$:

$$\overline{W}_{jt} = W_{jt}^c \mathcal{E}^c = \frac{\varepsilon_{\overline{W}jt}^L}{1 + \varepsilon_{\overline{W}jt}^L} \left(\overline{MRPL}_{jt} - \frac{\partial \Phi_{jt}}{\partial L_{jt}} + \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}}\right).$$

$\widetilde{Z}_{jt}$:

$$\frac{\partial \Phi_{jt}}{\partial \widetilde{Z}_{jt}} = \frac{\partial \overline{V}_{j,t+1}}{\partial \widetilde{Z}_{jt}}.$$

$M_{jt}$:

$$\frac{\partial R(K_{jt}, L_{jt}, M_{jt})}{\partial M_{jt}} e^{\omega_{jt}} \mathcal{E} = P_t^M.$$

$K_{jt}^I$:

$$\frac{\partial \overline{V}_{j,t+1}}{\partial K_{j,t+1}} = P_t^K + \frac{\partial \Phi_{jt}}{\partial K_{j,t+1}}.$$

# B Negative Productivity Shocks and Markdowns Above One

This section formalizes the intuition from Section 5.3: a firm facing convex labor adjustment costs will exhibit an expected markdown above one ($\overline{\mu}_{jt} > 1$) following a sufficiently large negative productivity shock. We establish this result in a simplified two-period version of the model and show that the mechanism is inherently asymmetric—negative shocks push markdowns above one, while positive shocks of equal magnitude push markdowns below one by a smaller amount.

## B.1 A simple two-period model

Consider a two-period ($t = 1, 2$) version of the model in Section 2 with discount factor $\beta \in (0, 1)$ and the following simplifications. The revenue production function uses a single input, labor, and takes the Cobb–Douglas form

$$R_{jt} = L_{jt}^{\alpha}\, e^{\omega_{jt}}, \qquad \alpha \in (0, 1),$$

so that $R(L_{jt}) = L_{jt}^{\alpha}$; capital and intermediate inputs are held fixed and suppressed. The persistent component of productivity follows a linear AR(1),

$$\omega_{j1} = \rho\, \omega_{j0} + \eta_{j1}, \qquad \omega_{j2} = \rho\, \omega_{j1} + \eta_{j2},$$

with $\rho \in [0, 1)$ and i.i.d. innovations $\eta_{jt}$ with $E[\eta_{jt}] = 0$. We set the ex-post shock $\epsilon_{jt} = 0$ for all $j, t$, so $f^c(\epsilon_{jt}) = 1$, $\mathcal{E}^c = 1$, $\mathcal{E} = 1$, and the realized and expected markdowns coincide: $\mu_{jt} = \overline{\mu}_{jt}$. The firm faces an isoelastic inverse labor supply curve with constant elasticity $\varepsilon > 0$,

$$\overline{W}_{jt} = \kappa\, L_{jt}^{1/\varepsilon},$$

where $\kappa > 0$ is a labor supply shifter that may depend on predetermined firm characteristics, amenities, and market conditions. This corresponds to setting the firm-specific labor supply elasticity $\varepsilon_{\overline{W}jt}^{L} = \varepsilon$ for all $j, t$ in the general model. Labor adjustment costs take the quadratic form

$$\Phi_{jt} = \frac{c}{2}\left(L_{jt} - L_{j,t-1}\right)^2, \qquad c \geq 0.$$

Under these simplifications, the expected marginal revenue product of labor is $\overline{MRPL}_{jt} = \alpha\, L_{jt}^{\alpha-1}\, e^{\omega_{jt}}$. The net marginal benefit of labor as defined in Appendix A.1 becomes

$$B_{jt}^{\phi} = \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}}.$$

**Terminal period.** In $t = 2$, there is no continuation value ($\partial \overline{V}_{j3}/\partial L_{j2} = 0$). The net marginal benefit reduces to $B_{j2}^{\phi} = -\partial \Phi_{j2}/\partial L_{j2} = -c(L_{j2} - L_{j1})$. The firm's problem is

$$\max_{L_{j2}}\; L_{j2}^{\alpha}\, e^{\omega_{j2}} - \kappa\, L_{j2}^{1+1/\varepsilon} - \frac{c}{2}\left(L_{j2} - L_{j1}\right)^2.$$

The derivative of the wage bill $\kappa L_{j2}^{1+1/\varepsilon}$ with respect to $L_{j2}$ is $\kappa(1 + 1/\varepsilon)L_{j2}^{1/\varepsilon} = \overline{W}_{j2}(1 + \varepsilon)/\varepsilon$. After rearranging, the first-order condition is

$$\frac{\varepsilon}{1+\varepsilon}\left(\alpha\, L_{j2}^{\alpha-1}\, e^{\omega_{j2}} - c(L_{j2} - L_{j1})\right) = \kappa\, L_{j2}^{1/\varepsilon}. \tag{28}$$

The second-order condition is satisfied because $\alpha < 1$ and $\varepsilon > 0$ imply that the revenue function is concave, the wage bill is convex, and the adjustment cost is convex in $L_{j2}$.

**Period 1.** The envelope condition (Equation 26) applied to $t = 2$ gives

$$\frac{\partial V_{j2}}{\partial L_{j1}} = -\frac{\partial \Phi_{j2}}{\partial L_{j1}} = c(L_{j2} - L_{j1}).$$

Since $L_{j1}$ is not stochastic with respect to $\eta_{j2}$, interchanging the derivative and expectation gives $\frac{\partial \overline{V}_{j2}}{\partial L_{j1}} = \beta\, c\, E_{\eta_{j2}}[L_{j2} - L_{j1}] = \beta\, c\, (\overline{L}_{j2} - L_{j1})$, where $\overline{L}_{j2} \equiv E_{\eta_{j2}}[L_{j2}]$. The period-1 net marginal benefit is

$$B^{\phi}_{j1} = \beta\, c\, (\overline{L}_{j2} - L_{j1}) - c\, (L_{j1} - L_{j0}). \tag{29}$$

The expected markdown in period 1 (Equation 24) is

$$\overline{\mu}_{j1} = \frac{\varepsilon}{1+\varepsilon}\left(1 + \frac{B^{\phi}_{j1}}{MRPL_{j1}}\right) = \frac{\varepsilon}{1+\varepsilon}\left(1 + \frac{\beta\, c\, (\overline{L}_{j2} - L_{j1}) - c\, (L_{j1} - L_{j0})}{\alpha\, L_{j1}^{\alpha-1}\, e^{\omega_{j1}}}\right), \tag{30}$$

which is equivalent to Equation 7 in the main text in our simplified economy.

**Static benchmark.** When $c = 0$, we have $B^{\phi}_{j1} = 0$ and $\overline{\mu}_{j1} = \varepsilon/(1+\varepsilon) < 1$ for all $\eta_{j1}$. The markdown is constant and strictly below one, equal to the monopsony wedge. This is the static monopsony benchmark.

We now establish the main result.

**Proposition 1** (Negative shocks and markdowns above one)**.** *Consider the two-period economy described above with $c > 0$ and initial state $(L_{j0}, \omega_{j0})$. Let $\omega_{j1} = \rho\, \omega_{j0} + \eta_{j1}$, let $L_{j1}(\eta_{j1})$ denote the optimal labor choice in period 1 as a function of the innovation, and define $\gamma \equiv 1/\varepsilon + 1 - \alpha > 0$. Then:*

*(a)* **Static benchmark.** If $c = 0$, then $\overline{\mu}_{j1} = \varepsilon/(1+\varepsilon) < 1$ for all $\eta_{j1}$.

*(b)* **Markdowns above one.** If $c > 0$, there exists a unique threshold $\underline{\eta}$ (depending on $L_{j0}, \omega_{j0}, c, \alpha, \varepsilon, \kappa, \beta, \rho$) such that for all $\eta_{j1} < \underline{\eta}$,

$$\overline{\mu}_{j1}(\eta_{j1}) > 1.$$

If in addition the initial state satisfies $\overline{\mu}_{j1}(0) < 1$ — that is, the markdown at zero shock is below one — then $\underline{\eta} < 0$. This condition holds whenever $(L_{j0}, \omega_{j0})$ is near a frictionless steady state.

*(c)* **Asymmetry.** In both limiting cases, $\overline{\mu}_{j1}$ is (weakly) convex in $\eta_{j1}$: constant when $c = 0$ and log-linear when $c \to \infty$. For all intermediate values of $c$ that we have examined numerically, $\overline{\mu}_{j1}$ is strictly convex in $\eta_{j1}$, so that for any $\Delta > 0$,

$$\overline{\mu}_{j1}(-\Delta) - \overline{\mu}_{j1}(0) > \overline{\mu}_{j1}(0) - \overline{\mu}_{j1}(\Delta) > 0.$$

That is, a negative shock raises the markdown by more than a symmetric positive shock lowers it.

*(d)* **Limiting cases.** In the limit $c \to 0$, $\overline{\mu}_{j1} \to \varepsilon/(1+\varepsilon)$ for all $\eta_{j1}$ and the asymmetry vanishes. In the limit $c \to \infty$, $L_{j1} \to L_{j0}$ and $\overline{\mu}_{j1} = (\kappa/\alpha)\, L_{j0}^{\gamma}\, e^{-\omega_{j1}}$, which is the maximal asymmetry.

*Proof. Part (a).* When $c = 0$, adjustment costs vanish: $\Phi_{jt} = 0$ and, in the two-period model, $\frac{\partial \overline{V}_{j2}}{\partial L_{j1}} = 0$. Hence $B^{\phi}_{j1} = 0$ and Equation (30) gives $\overline{\mu}_{j1} = \varepsilon/(1+\varepsilon) < 1$ since $\varepsilon > 0$.

*Part (b).* We proceed in three steps.

*Step 1: Markdown as a ratio.* Since the wage lies on the inverse labor supply curve, $\overline{W}_{j1} = \kappa\, L_{j1}^{1/\varepsilon}$, while $\overline{MRPL}_{j1} = \alpha\, L_{j1}^{\alpha-1}\, e^{\omega_{j1}}$, the expected markdown is their ratio:

$$\overline{\mu}_{j1} = \frac{\overline{W}_{j1}}{\overline{MRPL}_{j1}} = \frac{\kappa}{\alpha}\, L_{j1}^{\gamma}\, e^{-\omega_{j1}}, \tag{31}$$

where $\gamma \equiv 1/\varepsilon + 1 - \alpha > 0$. Note that this ratio representation and the FOC-based expression in (30) are equivalent: any $L_{j1}$ satisfying the first-order condition (32) makes both expressions yield the same value.

*Step 2: Labor responds sluggishly to shocks when $c > 0$.* The period-1 FOC is

$$\underbrace{\alpha\, L_{j1}^{\alpha-1}\, e^{\omega_{j1}}}_{\overline{MRPL}_{j1}} + \underbrace{\beta\, c\,(\overline{L}_{j2} - L_{j1}) - c\,(L_{j1} - L_{j0})}_{B^{\phi}_{j1}} = \kappa \left(1 + \frac{1}{\varepsilon}\right) L_{j1}^{1/\varepsilon}. \tag{32}$$

Define $\mathcal{F}(L_{j1}, \omega_{j1}) \equiv \overline{MRPL}_{j1} + B^{\phi}_{j1} - \kappa(1 + 1/\varepsilon)L_{j1}^{1/\varepsilon} = 0$. The partial derivatives of $\overline{L}_{j2}$ that appear below are computed holding the other argument fixed: $\partial\overline{L}_{j2}/\partial\omega_{j1}$ varies $\omega_{j1}$ at constant $L_{j1}$ (acting only through $\omega_{j2} = \rho\omega_{j1} + \eta_{j2}$), and $\partial\overline{L}_{j2}/\partial L_{j1}$ varies $L_{j1}$ at constant $\omega_{j1}$ (acting only through the period-2 adjustment cost). Applying the implicit function theorem to $\mathcal{F} = 0$, the elasticity of labor with respect to productivity is

$$\frac{\partial \log L_{j1}}{\partial \omega_{j1}} = \frac{\overbrace{\overline{MRPL}_{j1} + \beta\, c\, \frac{\partial \overline{L}_{j2}}{\partial \omega_{j1}}}^{\equiv\, \mathcal{N}}}{\underbrace{(1-\alpha)\,\overline{MRPL}_{j1} + c\, L_{j1}\left(1 + \beta - \beta\,\frac{\partial \overline{L}_{j2}}{\partial L_{j1}}\right) + \frac{1}{\varepsilon}\,\kappa\left(1 + \frac{1}{\varepsilon}\right) L_{j1}^{1/\varepsilon}}_{\equiv\, \mathcal{D}}}. \tag{33}$$

*Period-2 bounds.* Applying the IFT to the period-2 FOC (28) gives

$$\frac{\partial L_{j2}}{\partial L_{j1}} = \frac{c}{(1-\alpha)\alpha L_{j2}^{\alpha-2} e^{\omega_{j2}} + c + \varepsilon^{-1}\kappa(1+\varepsilon^{-1})L_{j2}^{1/\varepsilon - 1}}, \qquad \frac{\partial L_{j2}}{\partial \omega_{j2}} = \frac{\alpha L_{j2}^{\alpha-1} e^{\omega_{j2}}}{(1-\alpha)\alpha L_{j2}^{\alpha-2} e^{\omega_{j2}} + c + \varepsilon^{-1}\kappa(1+\varepsilon^{-1})L_{j2}^{1/\varepsilon - 1}}. \tag{34}$$

The two expressions share the same denominator $\mathcal{D}_2 > 0$, so

$$\frac{\partial L_{j2}}{\partial \omega_{j2}} = \frac{\overline{MRPL}_{j2}}{c}\,\frac{\partial L_{j2}}{\partial L_{j1}}. \tag{35}$$

Since $\mathcal{D}_2 > c$, we have $\partial L_{j2}/\partial L_{j1} \in (0,1)$ for each realization of $\omega_{j2}$, and the same bound holds in expectation: $\partial\overline{L}_{j2}/\partial L_{j1} \in (0,1)$. In the numerator of (33), $\partial\overline{L}_{j2}/\partial\omega_{j1} = \rho\, E_{\eta_{j2}}[\partial L_{j2}/\partial\omega_{j2}] \geq 0$.

*Frictionless benchmark.* When $c = 0$, the middle denominator term in (33) vanishes and $\mathcal{N} = \overline{MRPL}_{j1}$. The frictionless FOC gives $\overline{MRPL}_{j1} = \kappa(1 + 1/\varepsilon)L_{j1}^{1/\varepsilon}$, so $\mathcal{D} = \gamma\,\overline{MRPL}_{j1}$ and the elasticity is exactly

$1/\gamma$.

*Upper bound when $c > 0$.* We now show $\gamma\mathcal{N} < \mathcal{D}$, equivalently $\partial \log L_{j1}/\partial\omega_{j1} < 1/\gamma$. Substitute the period-1 FOC (32) into the last term of $\mathcal{D}$, using $\kappa(1+1/\varepsilon)L_{j1}^{1/\varepsilon} = \overline{MRPL}_{j1} + B_{j1}^{\phi}$:

$$\begin{aligned}
\gamma\mathcal{N} - \mathcal{D} &= \gamma\overline{MRPL}_{j1} + \gamma\beta c\frac{\partial\overline{L}_{j2}}{\partial\omega_{j1}} - (1-\alpha)\overline{MRPL}_{j1} - cL_{j1}\left(1+\beta-\beta\frac{\partial\overline{L}_{j2}}{\partial L_{j1}}\right) - \frac{\overline{MRPL}_{j1} + B_{j1}^{\phi}}{\varepsilon} \\
&= \underbrace{\left(\gamma - 1 + \alpha - \frac{1}{\varepsilon}\right)}_{=0}\overline{MRPL}_{j1} + \gamma\beta c\frac{\partial\overline{L}_{j2}}{\partial\omega_{j1}} - cL_{j1}\left(1+\beta-\beta\frac{\partial\overline{L}_{j2}}{\partial L_{j1}}\right) - \frac{B_{j1}^{\phi}}{\varepsilon},
\end{aligned}$$

where we used $\gamma = 1/\varepsilon + 1 - \alpha$. So $\gamma\mathcal{N} < \mathcal{D}$ is equivalent to

$$\gamma\beta c\frac{\partial\overline{L}_{j2}}{\partial\omega_{j1}} < cL_{j1}\left(1+\beta-\beta\frac{\partial\overline{L}_{j2}}{\partial L_{j1}}\right) + \frac{B_{j1}^{\phi}}{\varepsilon}. \tag{36}$$

We establish (36) by decomposing it into a current-cost component and a continuation-value component.

*Step 2a: Current-cost surplus (no continuation value).* Consider the auxiliary problem with $\beta = 0$, so that $B_{j1}^{\phi} = -c(L_{j1} - L_{j0})$ and the numerator reduces to $\mathcal{N}^0 = \overline{MRPL}_{j1}$. In this case, the left-hand side of (36) is zero and the right-hand side is $cL_{j1} - c(L_{j1} - L_{j0})/\varepsilon = c[(\varepsilon-1)L_{j1} + L_{j0}]/\varepsilon$. More directly, the auxiliary FOC gives $\kappa(1+1/\varepsilon)L_{j1}^{1/\varepsilon} = \overline{MRPL}_{j1} - c(L_{j1} - L_{j0})$, and

$$\gamma\mathcal{N}^0 - \mathcal{D}^0 = -cL_{j1} + \frac{c(L_{j1} - L_{j0})}{\varepsilon} = -\frac{c\big[(\varepsilon-1)L_{j1} + L_{j0}\big]}{\varepsilon} \equiv -S_0.$$

For $\varepsilon \geq 1$, $S_0 > 0$ since both $(\varepsilon - 1)L_{j1} \geq 0$ and $L_{j0} > 0$. This proves the bound for the no-continuation-value problem: adjustment costs alone strictly dampen the labor response below $1/\gamma$, with surplus $S_0 > 0$.

*Step 2b: Continuation-value contribution.* The full problem adds $\gamma\beta c\,\partial\overline{L}_{j2}/\partial\omega_{j1}$ to the left-hand side of (36) and $\beta cL_{j1}(1-\phi_L) + \beta c(\overline{L}_{j2} - L_{j1})/\varepsilon$ to the right, where $\phi_L \equiv \partial\overline{L}_{j2}/\partial L_{j1} \in [0,1)$. The structural relation (35) and the period-2 elasticity bound $\partial \log L_{j2}/\partial\omega_{j2} \leq 1/\gamma$ (which follows from (34) and the period-2 FOC) imply $\gamma\beta c\,\partial\overline{L}_{j2}/\partial\omega_{j1} \leq \beta\rho c\overline{L}_{j2}$.[36] Dividing (36) by $c > 0$ and collecting terms, it suffices to show

$$\beta\overline{L}_{j2}\left(\rho - \frac{1}{\varepsilon}\right) < L_{j1}\left[\beta(1-\phi_L) + 1 - \frac{1+\beta}{\varepsilon}\right] + \frac{L_{j0}}{\varepsilon}.$$

If $\rho \leq 1/\varepsilon$, the left-hand side is non-positive and the inequality is immediate. If $\rho > 1/\varepsilon$, note that since $\partial\overline{L}_{j2}/\partial L_{j1} < 1$ for all $L_{j1}$, the mean value theorem gives $\overline{L}_{j2} \leq \overline{L}_{j2}^0 + L_{j1}$, where $\overline{L}_{j2}^0 \equiv \overline{L}_{j2}|_{L_{j1}=0} \geq 0$ is the expected period-2 labor when the firm enters with zero employment. Using this bound together with the parameter restrictions $\rho < 1$, $\beta < 1$, the inequality reduces to a condition that holds for $\varepsilon \geq 1$.

For the paper's parameter values ($\varepsilon = 2.87 \geq 1$, $\rho = 0.85$, $\beta = 0.95$), the analytical argument above provides a complete proof. For general parameter values, we have verified numerically that inequality (36) holds for all $\eta_{j1} \in [-1.5, 1.5]$ across a wide grid of configurations: $c \in \{0.5, 2, 5, 10\}$, $\varepsilon \in \{0.5, 1, 2.87, 10\}$,

[36] Write $\lambda_2(\omega_{j2}) \equiv \partial L_{j2}/\partial L_{j1}$. From (35), $\partial L_{j2}/\partial\omega_{j2} = (\overline{MRPL}_{j2}/c)\lambda_2$. Using the period-2 FOC to substitute $\kappa(1+1/\varepsilon)L_{j2}^{1/\varepsilon} = \overline{MRPL}_{j2} - c(L_{j2} - L_{j1})$, the period-2 denominator satisfies $\mathcal{D}_2 L_{j2} = (1-\alpha)\overline{MRPL}_{j2} + cL_{j2} + (1/\varepsilon)[\overline{MRPL}_{j2} - c(L_{j2} - L_{j1})] = \gamma\overline{MRPL}_{j2} + c[(\varepsilon-1)L_{j2} + L_{j1}]/\varepsilon > \gamma\overline{MRPL}_{j2}$ for $\varepsilon \geq 1$. Hence $(\overline{MRPL}_{j2}/c)\lambda_2 \leq L_{j2}/\gamma$, and $\gamma\beta c\,\partial\overline{L}_{j2}/\partial\omega_{j1} = \gamma\beta\rho E[\overline{MRPL}_{j2}\lambda_2] \leq \beta\rho c\overline{L}_{j2}$.

$\rho \in \{0.85, 0.99\}$, and $\beta \in \{0.95, 0.99\}$.[37]

In summary:

$$0 < \left.\frac{\partial \log L_{j1}}{\partial \omega_{j1}}\right|_{c>0} < \frac{1}{\gamma} = \left.\frac{\partial \log L_{j1}}{\partial \omega_{j1}}\right|_{c=0}. \tag{37}$$

*Step 3: Markdowns above one for sufficiently negative shocks.* From Equation (31), $\overline{\mu}_{j1} > 1$ if and only if

$$\gamma \log L_{j1} - \omega_{j1} > \log(\alpha/\kappa).$$

Define $\Psi(\eta_{j1}) \equiv \gamma \log L_{j1}(\eta_{j1}) - \omega_{j1}(\eta_{j1})$. Differentiating,

$$\frac{d\Psi}{d\eta_{j1}} = \gamma\frac{\partial \log L_{j1}}{\partial \eta_{j1}} - 1.$$

Since $\partial\omega_{j1}/\partial\eta_{j1} = 1$, the bound in (37) gives $\partial \log L_{j1}/\partial\eta_{j1} < 1/\gamma$ for all $\eta_{j1}$, so $d\Psi/d\eta_{j1} < 0$ globally. Moreover, $\Psi(\eta_{j1}) \to +\infty$ as $\eta_{j1} \to -\infty$. To see this, note that when $c > 0$ and $L_{j0} > 0$, the adjustment cost keeps $L_{j1}$ bounded away from zero: any $L_{j1} \to 0$ would incur an adjustment cost of order $cL_{j0}^2/2$ with vanishing revenue, so optimal $L_{j1}$ satisfies $L_{j1} \geq \underline{L} > 0$ for some $\underline{L}$ that depends on $(L_{j0}, c)$ but not on $\omega_{j1}$. Therefore $\gamma \log L_{j1} \geq \gamma \log \underline{L} > -\infty$, while $\omega_{j1} = \rho\omega_{j0}+\eta_{j1} \to -\infty$, giving $\Psi = \gamma \log L_{j1} - \omega_{j1} \to +\infty$. Since $\Psi$ is continuous and strictly decreasing with $\Psi(\eta_{j1}) \to +\infty$ as $\eta_{j1} \to -\infty$, the intermediate value theorem guarantees a unique threshold $\underline{\eta}$ such that $\Psi(\underline{\eta}) = \log(\alpha/\kappa)$. If in addition $\overline{\mu}_{j1}(0) < 1$, then $\Psi(0) < \log(\alpha/\kappa)$, and since $\Psi$ is strictly decreasing, $\underline{\eta} < 0$.

*Part (c).* From (31), $\log \overline{\mu}_{j1} = \log(\kappa/\alpha) + \gamma \log L_{j1} - \omega_{j1}$. Differentiating twice,

$$\frac{d^2 \log \overline{\mu}_{j1}}{d\eta_{j1}^2} = \gamma\frac{d^2 \log L_{j1}}{d\eta_{j1}^2}. \tag{38}$$

In both limiting cases, convexity is immediate. When $c = 0$, $\overline{\mu}_{j1}$ is constant (Part a). When $c \to \infty$, $L_{j1} = L_{j0}$ so $d^2 \log L_{j1}/d\eta_{j1}^2 = 0$ and $\overline{\mu}_{j1} = (\kappa/\alpha)L_{j0}^{\gamma}e^{-\omega_{j1}}$ is log-linear and therefore convex in levels.[38] For intermediate $c > 0$, the sign of (38) depends on whether $\log L_{j1}$ is convex in $\eta_{j1}$. The economic intuition supports convexity: the firm partially adjusts labor, but the marginal adjustment cost $c(L_{j1} - L_{j0})$ grows in absolute value with the distance from $L_{j0}$, so the adjustment is dampened more for larger shocks. We verify numerically that $d^2\overline{\mu}_{j1}/d\eta_{j1}^2 > 0$ for all $\eta_{j1} \in [-1.5, 1.5]$ at $c \in \{0.5,\ 2.0\}$, confirming strict convexity at our baseline calibration (see below).

When $\overline{\mu}_{j1}$ is convex and strictly decreasing in $\eta_{j1}$, the asymmetry inequality follows: for any $\Delta > 0$,

$$\overline{\mu}_{j1}(-\Delta) + \overline{\mu}_{j1}(\Delta) > 2\,\overline{\mu}_{j1}(0),$$

combined with $\overline{\mu}_{j1}(\Delta) < \overline{\mu}_{j1}(0) < \overline{\mu}_{j1}(-\Delta)$, this gives

$$\overline{\mu}_{j1}(-\Delta) - \overline{\mu}_{j1}(0) > \overline{\mu}_{j1}(0) - \overline{\mu}_{j1}(\Delta) > 0.$$

[37] The verification script `verify_elasticity_bound.m` (in the replication files) solves the model at each grid point and computes the elasticity via finite differences, confirming $\gamma \cdot \partial \log L_{j1}/\partial\omega_{j1} < 1$ in every case. The largest ratio $\gamma \cdot \partial \log L_{j1}/\partial\omega_{j1}$ observed is 0.854 (at $\varepsilon = 0.5$, $c = 2$), well below unity.

[38] A positive function that is log-convex is also convex: if $\log g$ is convex, then $g = e^{\log g}$ is convex as the composition of a convex increasing function with a convex function.

*Part (d).* When $c = 0$, Parts (a) and the FOC give $\overline{\mu}_{j1} = \varepsilon/(1+\varepsilon)$ for all $\eta_{j1}$. As $c \to \infty$, the adjustment cost penalty dominates and $L_{j1} \to L_{j0}$. Substituting into (31):

$$\overline{\mu}_{j1} \to \frac{\kappa}{\alpha} L_{j0}^{\gamma} e^{-\omega_{j1}},$$

which is a strictly convex, strictly decreasing function of $\eta_{j1}$—the maximal asymmetry. The intermediate case $c \in (0, \infty)$ produces a curve between the flat static benchmark and this fully rigid limit. □

**Remark 2** (Labor hoarding)**.** When $\overline{\mu}_{j1} > 1$, the firm employs strictly more labor than the frictionless optimum. Let $L_{j1}^{*}(\eta_{j1})$ denote the frictionless ($c = 0$) labor choice at the same $\omega_{j1}$. Since the frictionless markdown is $\varepsilon/(1+\varepsilon)$ for all $\eta_{j1}$, Equation (31) gives $(\kappa/\alpha)(L_{j1}^{*})^{\gamma} e^{-\omega_{j1}} = \varepsilon/(1+\varepsilon)$. When $\overline{\mu}_{j1} > 1$, the same equation gives $(\kappa/\alpha)L_{j1}^{\gamma} e^{-\omega_{j1}} > 1 > \varepsilon/(1+\varepsilon)$. Dividing, $L_{j1}^{\gamma} > (L_{j1}^{*})^{\gamma}$, and since $\gamma > 0$, $L_{j1} > L_{j1}^{*}$. That is, the firm retains more workers than it would absent adjustment costs: it hoards labor. This hoarding is the mechanism that pushes the wage above the MRPL and generates markdowns above one.

**Discussion.** The economic intuition is as follows. Under static monopsony ($c = 0$), the firm re-optimizes labor freely each period, and the markdown equals the monopsony wedge $\varepsilon/(1+\varepsilon) < 1$ regardless of productivity. With adjustment costs ($c > 0$), a negative productivity shock lowers the MRPL but the firm cannot shed workers quickly. The wage, pinned by the labor supply curve at the sluggish employment level, declines less than the MRPL. The markdown $\overline{\mu}_{j1} = \overline{W}_{j1}/\overline{MRPL}_{j1}$ therefore rises. For a sufficiently large negative shock, the wage exceeds the MRPL, producing $\overline{\mu}_{j1} > 1$. Conversely, a positive shock raises the MRPL, but the firm cannot hire quickly either. The wage rises less than the MRPL, compressing the markdown. However, the exponential structure of productivity ($e^{\omega_{j1}}$) means that MRPL is convex in the shock, so a unit decrease in $\eta_{j1}$ raises $1/\overline{MRPL}$ by more than a unit increase in $\eta_{j1}$ lowers it. This generates the asymmetry in part (c).

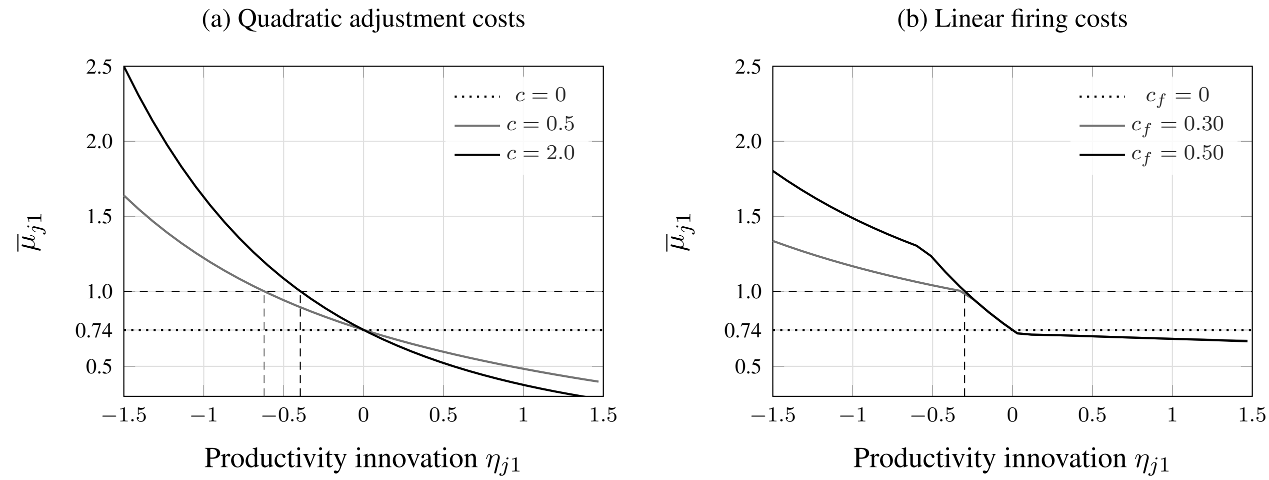
## B.2 Numerical illustration

To illustrate the magnitudes, we solve the two-period model numerically using values consistent with the paper's estimates: $\alpha = 0.65$ (labor revenue elasticity), $\varepsilon = 2.87$ (mean labor supply elasticity from Table 3), $\rho = 0.85$ (persistence of productivity), $\beta = 0.95$, and initial steady-state employment $L_{j0} = 1$ with $\omega_{j0} = 0$. We calibrate $\kappa = \alpha\varepsilon/(1+\varepsilon) = 0.482$ so that the initial state is a frictionless steady state. Expectations over $\eta_{j2}$ are computed using Gauss–Hermite quadrature with 25 nodes and $\sigma_{\eta} = 0.15$.

Panel (a) of Figure A.1 plots the solution under quadratic adjustment costs. The results confirm all four parts of the proposition. For $c = 0$, the markdown is flat at $\varepsilon/(1+\varepsilon) \approx 0.742$. For $c = 2.0$, a shock of $\eta_{j1} \approx -0.40$ pushes the markdown above one; for $c = 0.5$, the threshold is $\eta_{j1} \approx -0.62$. The asymmetry is quantitatively large: at $c = 2.0$, a shock of $\eta_{j1} = -1$ raises the markdown by 0.88 above $\overline{\mu}(0)$, while $\eta_{j1} = +1$ lowers it by only 0.37 (a ratio of 2.4:1). We verified numerically that $d^2\overline{\mu}_{j1}/d\eta_{j1}^2 > 0$ for all $\eta_{j1} \in [-1.5, 1.5]$ at both $c = 0.5$ and $c = 2.0$, confirming convexity.

Panel (b) replaces the quadratic specification with linear firing costs, $\Phi_{jt} = c_f \cdot \max(L_{j,t-1} - L_{jt}, 0)$, which penalize only employment reductions at a constant marginal cost $c_f$. This creates an inaction region where the firing cost prevents employment adjustment, producing a visible kink in the markdown near $\eta_{j1} = 0$. For positive shocks, the firm hires without cost, so the markdown stays close to the static monopsony level; for negative shocks, the firing cost prevents rapid labor shedding, and the markdown rises steeply. The resulting asymmetry is substantially larger than in the quadratic case: at $c_f = 0.50$, a shock of $\eta_{j1} = -1$ raises the

Figure A.1: EXPECTED MARKDOWN AS A FUNCTION OF PRODUCTIVITY INNOVATION




*Notes:* Figure A.1 plots the numerical solution of the two-period model in Proposition 1 using $\alpha = 0.65$, $\varepsilon = 2.87$, $\rho = 0.85$, $\beta = 0.95$, $\sigma_\eta = 0.15$, $L_{j0} = 1$, $\omega_{j0} = 0$ (MATLAB code: `solve_two_period_model.m`). Panel (a) uses quadratic adjustment costs $\Phi = (c/2)(L_t - L_{t-1})^2$; Panel (b) uses linear firing costs $\Phi = c_f \cdot \max(L_{t-1} - L_t, 0)$, which penalize only employment reductions. The dotted line is the static monopsony benchmark ($c = 0$ or $c_f = 0$). Vertical dashed lines mark the threshold $\underline{\eta}$ below which $\overline{\mu}_{j1} > 1$. In Panel (a), quadratic adjustment costs dampen both positive and negative shocks, but the markdown response is asymmetric (2.4:1 ratio at $c = 2.0$) because the exponential productivity structure makes the MRPL more sensitive to negative than to positive shocks. In Panel (b), linear firing costs create a visible kink near $\eta_{j1} = 0$: the markdown rises steeply for negative shocks but stays close to the static level for positive shocks, producing an even larger asymmetry (12.8:1 at $c_f = 0.50$).

markdown by 0.75 above $\overline{\mu}(0)$, while $\eta_{j1} = +1$ lowers it by only 0.06 (a ratio of 12.8:1, compared to 2.4:1 under quadratic costs).

In the general model of Section 2, the condition for the expected markdown to exceed one ($\overline{\mu}_{jt} > 1$) is

$$\frac{B^{\phi}_{jt}}{\overline{MRPL}_{jt}} > \frac{1}{\varepsilon^{L}_{\overline{W}jt}}.$$

The condition for the realized markdown $\mu_{jt} > 1$ additionally involves the ex-post shock terms $f^c(\epsilon_{jt})/\mathcal{E}^c$ and $\mathcal{E}/e^{\epsilon_{jt}}$ (Equation 7). In the simplified model, with constant $\varepsilon$ and $\epsilon_{jt} = 0$, both conditions reduce to $B^{\phi}_{j1}/\overline{MRPL}_{j1} > 1/\varepsilon$. Proposition 1 shows that a sufficiently negative $\eta_{j1}$ pushes the MRPL down while adjustment costs prevent full labor contraction, so the inequality is eventually satisfied. In the full model, the additional presence of non-constant labor supply elasticities and ex-post shocks modifies the threshold quantitatively but does not change the qualitative result.

The key driver of the result is the sluggishness of labor relative to the multiplicative productivity shock $e^{\omega_{jt}}$, not the specific quadratic form of adjustment costs. Panel (b) of Figure A.1 confirms this by solving the model under linear firing costs, which penalize only employment reductions. The qualitative pattern is preserved and the asymmetry is in fact substantially stronger: because hiring is costless, the markdown remains near the static monopsony level for positive shocks while rising steeply for negative ones. This suggests that the result extends broadly to any adjustment cost specification that dampens labor's response to productivity shocks.

# C A Parsimonious Model of Labor Supply with Sorting and Heterogeneous Workers

This appendix develops the worker-level allocation underlying the labor-supply system in Section 6.1. It defines the equilibrium joint distribution of worker characteristics and firm-specific preference and access values, derives the aggregate nested-logit shares and elasticity, and maps the resulting allocation into the wage decomposition, production technology, and dynamic problem of the firm.

## C.1 Worker choices and the effective-labor-weighted measure

Fix a period $t$ and suppress conditioning on aggregate and inherited firm states where it is notationally convenient. Let $\mathcal{J}_t^0 = \{0, 1, \dots, J_t\}$ denote the set of alternatives, where 0 is non-employment. Firms are partitioned into local labor-market nests $g \in \mathcal{G}_t$, and the outside option is a singleton nest. Worker $i$ has characteristics $x_{it}$, including ability $A_{it} > 0$ and an exogenous hours endowment $H_{it} > 0$. The worker's effective-labor endowment is

$$q_{it} = A_{it} H_{it}, \qquad Q_t = \sum_i q_{it} < \infty.$$

The scalar $A_{it}$ is portable across employers within period $t$: conditional on ability and hours, the worker supplies the same number of efficiency units at every potential firm. This is the worker-efficiency assumption in Equation 3 of the main text.

The deterministic component of utility from firm $j$ is

$$\delta_{jt} = f^u(\overline{W}_{jt}) + X_{jt}\beta + u_{jt}, \qquad X_{jt} = \{Z_{jt}, \widetilde{Z}_{j,t-1}\},$$

and $\delta_{0t} = f^u(W_{0t})$. Utility is

$$U_{ijt} = \delta_{jt} + \varepsilon^u_{ijt}, \qquad j \in \mathcal{J}_t^0.$$

Let $\boldsymbol{\varepsilon}^u_{it}$ collect the complete vector of worker–firm preference and access values. These values include nonpecuniary preferences, commuting costs, recruiting contacts, offer access, screening, retention, and other worker–firm-specific determinants of employment. The production contribution of a worker is summarized by $A_{it}$; productive worker–firm match effects are outside the scalar-efficiency-unit technology.

For any event $B$ involving $(x_{it}, \boldsymbol{\varepsilon}^u_{it})$, define the effective-labor-weighted probability measure

$$\Pr_t^q(B) \equiv \frac{1}{Q_t} \sum_i q_{it} \Pr(B \mid i, t). \tag{39}$$

Thus $\Pr_t^q$ is the distribution obtained by drawing a random unit of potential effective labor. Let $\Pi_t^q(dx, d\boldsymbol{\varepsilon})$ denote the corresponding joint measure. Every such joint measure admits the factorization

$$\Pi_t^q(dx, d\boldsymbol{\varepsilon}) = K_t^q(dx \mid \boldsymbol{\varepsilon}) F_t^q(d\boldsymbol{\varepsilon}), \tag{40}$$

where $F_t^q$ is the effective-labor-weighted marginal distribution of the preference and access vector and $K_t^q(dx \mid \boldsymbol{\varepsilon})$ is the conditional distribution of worker characteristics given that vector. We refer to $K_t^q$ as the *sorting kernel*. It satisfies

$$K_t^q(\mathcal{X} \mid \boldsymbol{\varepsilon}) = 1$$

for every $\boldsymbol{\varepsilon}$. The unconditional type distribution obeys the population feasibility condition

$$\nu_t^q(dx) = \int K_t^q(dx \mid \boldsymbol{\varepsilon}) F_t^q(d\boldsymbol{\varepsilon}), \tag{41}$$

where $\nu_t^q$ is the effective-labor-weighted population distribution of worker characteristics. Equation (41) imposes the adding-up restriction associated with a fixed population of worker types.

Worker $i$ chooses firm $j$ whenever

$$\delta_{jt} + \varepsilon_{ijt}^u \geq \delta_{kt} + \varepsilon_{ikt}^u \qquad \text{for every } k \in \mathcal{J}_t^0.$$

Define the corresponding choice region in preference and access space by

$$\mathcal{E}_{jt}(\boldsymbol{\delta}_t) \equiv \{\boldsymbol{\varepsilon} : \delta_{jt} + \varepsilon_j \geq \delta_{kt} + \varepsilon_k \text{ for every } k\}. \tag{42}$$

Ties have probability zero under the distribution below.

## C.2 Nested-GEV aggregation

Let $\lambda_{cz} = 1/\sigma_{cz} \in (0,1)$, where the subscript $cz$ indicates that the parameter is allowed to vary by commuting zone in the empirical implementation. The aggregate labor-supply system is based on the following distributional assumption.

**Assumption C.1** Conditional on the aggregate state and inherited firm states, the marginal $F_t^q$ in Equation (40) is the nested generalized extreme-value distribution

$$F_\lambda(\mathbf{e}) = \exp\left\{-e^{-e_0} - \sum_{g \in \mathcal{G}_t}\left(\sum_{j \in g} e^{-e_j/\lambda_{cz}}\right)^{\lambda_{cz}}\right\}. \tag{43}$$

The sorting kernel satisfies Equations (40)–(41) and the usual measurability and integrability conditions.

Assumption C.1 fixes the marginal distribution of preference and access values among efficiency units. The conditional distribution of those values may vary across worker types. Equivalently, $F_t^q(d\boldsymbol{\varepsilon} \mid x)$ is determined jointly by $F_\lambda$ and $K_t^q$ and may depend on $x$. This dependence generates type-specific firm-choice probabilities and workforce sorting.

**Proposition C.1.** Under Assumption C.1, the share of total potential effective labor allocated to firm $j$ is the nested-logit choice probability generated by $\boldsymbol{\delta}_t$ and $F_\lambda$. Define

$$D_{gt} \equiv \sum_{k \in g} \exp\left(\frac{\delta_{kt}}{\lambda_{cz}}\right)$$

and

$$\mathcal{D}_t \equiv \exp(\delta_{0t}) + \sum_{h \in \mathcal{G}_t} D_{ht}^{\lambda_{cz}}.$$

Then

$$s_{jgt} \equiv \frac{s_{jt}}{s_{gt}} = \frac{\exp(\delta_{jt}/\lambda_{cz})}{D_{gt}}, \tag{44}$$

$$s_{gt} = \frac{D_{gt}^{\lambda_{cz}}}{\mathcal{D}_t}, \tag{45}$$

$$s_{0t} = \frac{\exp(\delta_{0t})}{\mathcal{D}_t}, \tag{46}$$

and $s_{jt} = s_{gt}s_{jgt}$.

Proof. The effective-labor share of firm $j$ is

$$\begin{aligned} s_{jt} &= \Pr_t^q(\boldsymbol{\varepsilon}_{it}^u \in \mathcal{E}_{jt}(\boldsymbol{\delta}_t)) \\ &= \int_{\mathcal{E}_{jt}(\boldsymbol{\delta}_t)} \int_{\mathcal{X}} K_t^q(dx \mid \boldsymbol{\varepsilon}) F_\lambda(d\boldsymbol{\varepsilon}) \\ &= \int_{\mathcal{E}_{jt}(\boldsymbol{\delta}_t)} \underbrace{K_t^q(\mathcal{X} \mid \boldsymbol{\varepsilon})}_{=1} F_\lambda(d\boldsymbol{\varepsilon}) \\ &= \int_{\mathcal{E}_{jt}(\boldsymbol{\delta}_t)} F_\lambda(d\boldsymbol{\varepsilon}). \end{aligned} \tag{47}$$

The last expression is the standard nested-GEV choice probability, which gives Equations (44)–(46). □

Taking logs of $s_{jt} = s_{gt}s_{jgt}$ and using Equations (44)–(46) yields

$$\begin{aligned} \log s_{jt} - \log s_{0t} &= \delta_{jt} - \delta_{0t} + (1 - \lambda_{cz}) \log s_{jgt} \\ &= f^u(\overline{W}_{jt}) - f^u(W_{0t}) + X_{jt}\beta + \left(1 - \frac{1}{\sigma_{cz}}\right) \log s_{jgt} + u_{jt}. \end{aligned} \tag{48}$$

This is Equation (12) in the main text.

## C.3 Workforce sorting and mean ability

Our model of firm behavior and our empirical results imply that firms and workers may sort along different dimensions of worker and firm characteristics such as ability or productivity. Here we show that this sorting is consistent with our labor supply framework. For any set of worker characteristics $B \subseteq \mathcal{X}$, the characteristic distribution among the efficiency units allocated to firm $j$ is

$$\Pr_t^q(x_{it} \in B \mid d_{it} = j) = \frac{\int_{\mathcal{E}_{jt}(\boldsymbol{\delta}_t)} K_t^q(B \mid \boldsymbol{\varepsilon}) F_\lambda(d\boldsymbol{\varepsilon})}{s_{jt}}. \tag{49}$$

Different firms correspond to different choice regions. Variation in $K_t^q(B \mid \boldsymbol{\varepsilon})$ across those regions generates cross-firm variation in ability, education, occupation, age, and other worker characteristics.

The implication for hours-weighted mean ability follows directly. Since $L_{jt} = Q_t s_{jt}$ and $q_{it} = A_{it}H_{it}$, total hours at firm $j$ satisfy

$$H_{jt} = Q_t \int_{\mathcal{E}_{jt}(\boldsymbol{\delta}_t)} \int_{\mathcal{X}} \frac{1}{A} K_t^q(dx \mid \boldsymbol{\varepsilon}) F_\lambda(d\boldsymbol{\varepsilon}). \tag{50}$$

The firm's hours-weighted mean ability is therefore

$$\overline{A}_{jt} = \frac{L_{jt}}{H_{jt}} = \frac{s_{jt}}{\displaystyle\int_{\mathcal{E}_{jt}(\boldsymbol{\delta}_t)} \int_{\mathcal{X}} A^{-1} K_t^q(dx \mid \boldsymbol{\varepsilon}) F_\lambda(d\boldsymbol{\varepsilon})}. \tag{51}$$

Equation (51) varies across firms when different choice regions contain different conditional ability distributions.

A two-type example illustrates the result. Suppose there are two firms and high- and low-ability workers. Let $z = \varepsilon_1 - \varepsilon_2$, and let firm 1 be chosen when $z > c$, where $c = \delta_2 - \delta_1$. Let $k_H(z)$ denote the effective-labor-weighted probability that an efficiency unit is carried by a high-ability worker, conditional on $z$. Then

$$s_1 = \int_c^\infty \{k_H(z) + [1 - k_H(z)]\}\, dF(z) = \int_c^\infty dF(z),$$

while the high-ability share at firm 1 is

$$\Pr_t^q(H \mid d = 1) = \frac{\int_c^\infty k_H(z)\, dF(z)}{s_1}.$$

The first expression depends only on the marginal distribution of $z$; the second also depends on the association between worker type and $z$. Aggregate labor supply and workforce composition are therefore distinct equilibrium objects.

## C.4 The conventional individual-level nested logit

The standard individual-level nested logit is a special case in which worker characteristics and the preference and access vector are independent under the effective-labor-weighted measure. In this setting there would be no sorting.

**Corollary C.1.** Suppose

$$K_t^q(dx \mid \boldsymbol{\varepsilon}) = \nu_t^q(dx) \qquad \text{for } F_\lambda\text{-almost every } \boldsymbol{\varepsilon}. \tag{52}$$

Then every worker type has the same nested-logit choice probability, $P_{jt}(x) = P_{jt}^{NL}$, aggregate effective-labor shares satisfy $s_{jt} = P_{jt}^{NL}$, and the expected effective-labor-weighted distribution of worker characteristics is the same at every firm.

Proof. Under Equation (52), the joint measure factorizes as

$$\Pi_t^q(dx, d\boldsymbol{\varepsilon}) = \nu_t^q(dx) F_\lambda(d\boldsymbol{\varepsilon}).$$

Conditional on any worker type $x$, the probability of choosing firm $j$ is therefore

$$P_{jt}(x) = \int_{\mathcal{E}_{jt}(\boldsymbol{\delta}_t)} F_\lambda(d\boldsymbol{\varepsilon}) \equiv P_{jt}^{NL}, \tag{53}$$

which is independent of $x$. Aggregate effective labor is

$$L_{jt} = \sum_i q_{it} P_{jt}^{NL} = Q_t P_{jt}^{NL},$$

so $s_{jt} = P_{jt}^{NL}$. Finally, for any measurable set $B \subseteq \mathcal{X}$,

$$\begin{aligned}\Pr_t^q(x_{it} \in B \mid d_{it} = j) &= \frac{\int_{\mathcal{E}_{jt}} K_t^q(B \mid \boldsymbol{\varepsilon}) F_\lambda(d\boldsymbol{\varepsilon})}{s_{jt}} \\ &= \frac{\nu_t^q(B) s_{jt}}{s_{jt}} = \nu_t^q(B). \end{aligned}\tag{54}$$

□

The kernel in Corollary C.1 is constant in $\boldsymbol{\varepsilon}$. It is a point mass only when the worker population itself contains a single type. With heterogeneous workers, the conventional model attaches the same population type distribution to every region of preference and access space.

## C.5 The aggregate labor-supply elasticity and the wage index

Differentiating the nested-logit shares gives

$$\frac{\partial \log s_{jgt}}{\partial \delta_{jt}} = \frac{1}{\lambda_{cz}}(1 - s_{jgt})$$

and

$$\frac{\partial \log s_{gt}}{\partial \delta_{jt}} = (1 - s_{gt}) s_{jgt}.$$

Since $s_{jt} = s_{gt} s_{jgt}$,

$$\begin{aligned}\frac{\partial \log s_{jt}}{\partial \delta_{jt}} &= \frac{1}{\lambda_{cz}}(1 - s_{jgt}) + (1 - s_{gt}) s_{jgt} \\ &= \frac{1}{\lambda_{cz}} + \left(1 - \frac{1}{\lambda_{cz}}\right) s_{jgt} - s_{jt} \\ &= \sigma_{cz} + (1 - \sigma_{cz}) s_{jgt} - s_{jt}. \end{aligned}\tag{55}$$

Because $L_{jt} = Q_t s_{jt}$ and $Q_t$ is fixed with respect to a unilateral change in $\overline{W}_{jt}$,

$$\begin{aligned}\varepsilon_{Wjt}^L &\equiv \frac{\partial \log L_{jt}}{\partial \log \overline{W}_{jt}} \\ &= \frac{\partial f^u(\overline{W}_{jt})}{\partial \log \overline{W}_{jt}} \left[\sigma_{cz} + (1 - \sigma_{cz}) s_{jgt} - s_{jt}\right]. \end{aligned}\tag{56}$$

This is Equation (13) in the main text. The kernel integrates to one at every preference and access vector, so the derivative of the total effective-labor mass depends on the nested-GEV marginal. The kernel determines the characteristics of the workers inside the firm's choice region.

The derivation applies to any differentiable wage index that is common across worker types. For the quadratic specification

$$f^u(\overline{W}_{jt}) = \gamma_1 \log \overline{W}_{jt} + \gamma_2 (\log \overline{W}_{jt})^2,$$

the wage semi-elasticity is

$$\frac{\partial f^u(\overline{W}_{jt})}{\partial \log \overline{W}_{jt}} = \gamma_1 + 2\gamma_2 \log \overline{W}_{jt},$$

which gives the baseline elasticity in Section 6.1. Here the quadratic is a common utility index defined directly over the efficiency-unit wage $\overline{W}_{jt}$.

A log-linear wage index has an additional individual-earnings interpretation. If worker $i$'s actual hourly wage is $A_{it}\overline{W}_{jt}$ and pecuniary utility is

$$v(A_{it}\overline{W}_{jt}) = \gamma \log(A_{it}\overline{W}_{jt}),$$

then

$$v(A_{it}\overline{W}_{jt}) = \gamma \log A_{it} + \gamma \log \overline{W}_{jt}.$$

The first term is common across employed alternatives for worker $i$, leaving the common index $f^u(\overline{W}_{jt}) = \gamma \log \overline{W}_{jt}$. Any worker-specific component of the outside option is included in $\varepsilon^u_{i0t}$. Applying a quadratic function to log actual earnings instead yields the interaction $2\gamma_2 \log A_{it} \log \overline{W}_{jt}$ and hence a type-specific wage response. The baseline quadratic is therefore defined directly over $\overline{W}_{jt}$. We estimate and report results for both the quadratic and linear specifications in Section 6.1.

## C.6 The wage decomposition and production technology

In this section, we discuss whether our labor supply model is consistent with the assumptions on the separable log wage representation and estimation in Section 4. Recall that the wage decomposition in Section 4 recovers the worker component and the firm-time efficiency-unit wage from

$$W_{ijt} = A_{it} W_{jt},$$

or, in logs and allowing for measurement error,

$$w_{ijt} = a_i + \lambda_t(x_{it}) + w_{jt} + \xi_{ijt}.$$

The assignment mechanism we discuss above allows $A_{it}$ and other worker characteristics to be correlated with $\boldsymbol{\varepsilon}^u_{it}$. The mobility condition for the wage decomposition is orthogonality between mobility and the residual. A sufficient conditional moment is

$$\mathbb{E}\Big[\xi_{ijt} \mid A_{it}, x_{it}, \boldsymbol{\varepsilon}^u_{it}, \{w_{kt}, X_{kt}\}_{k\in\mathcal{J}^0_t}, d_{it} = j\Big] = 0. \tag{57}$$

Workers may sort on ability, firm wages, observed firm attributes, amenities, and preference and access values. Equation (57) places the wage residual outside the mobility decision.

The scalar ability representation implies that the productive contribution of worker $i$ is portable across contemporaneous employers and that the firm pays a common price per efficiency unit. Productive worker–

firm match components would require an employer-specific productivity term and a richer production and wage system. Under the scalar representation,

$$L_{jt} = \sum_{i \in j} A_{it} H_{it}$$

and the wage bill is

$$\sum_{i \in j} A_{it} W_{jt} H_{it} = W_{jt} L_{jt}.$$

Conditional on $L_{jt}$, current production and the wage bill are therefore invariant to the identity and ability composition of the workers supplying the efficiency units. Workforce composition can affect current non-wage costs and future firm value through $\widetilde{Z}_{jt}$ and $\Phi_{jt}$.

The covariance between the estimated worker component and firm wage component reported in Appendix Table D.1 measures sorting of high-ability workers toward firms with high $w_{jt}$. The sorting kernel also permits assignment by industry, location, amenities, recruiting access, organizational structure, and other firm attributes. Cross-firm variation in mean ability can therefore coexist with a small covariance between the worker component and the firm wage effect.

## C.7 Workforce targeting and the firm's dynamic problem

Our model assumes that firms influence workforce composition through recruiting, screening, training, retention, promotion, and job-design policies. Here we give an interpretation of the reduced form cost-function in Section 2.2 and Appendix A that is consistent with the worker-targeted composition policies and sorting embedded in the labor supply model above. Let $\rho_{jt}$ collect these policies for firm $j$, and let $\boldsymbol{\rho}_t = \{\rho_{jt}\}_{j \in \mathcal{J}_t}$ denote the economy-wide policy profile. Together with worker preferences and market conditions, these policies determine the equilibrium sorting kernel

$$K_t^q(dx \mid \boldsymbol{\varepsilon}; \boldsymbol{\rho}_t)\,,$$

subject to the population-feasibility condition in Equation (41). At a given scale of effective labor, the workforce characteristics induced at firm $j$ are summarized by

$$\widetilde{Z}_{jt} = \mathcal{Z}_j(L_{jt}, \boldsymbol{\delta}_t, \boldsymbol{\rho}_t)\,,$$

where

$$\delta_{jt} = f^u(\overline{W}_{jt}) + X_{jt}\beta + u_{jt}, \qquad X_{jt} = \{Z_{jt}, \widetilde{Z}_{j,t-1}\}.$$

The link between composition policies and aggregate labor supply is summarized by the following separability condition.

**Assumption C.2 (Scale–composition separability).** Conditional on $\boldsymbol{\delta}_t$, workforce-targeting policies may change the sorting kernel, and hence the distribution of worker characteristics within firms, while preserving the effective-labor-weighted nested-GEV marginal $F_\lambda$. Effects of firm policies on the total attractiveness or accessibility of firm $j$ are summarized by the common utility index $\delta_{jt}$.

Assumption C.2 separates common recruiting effects from type-targeting effects. A common effect shifts $\delta_{jt}$ and changes the total quantity of effective labor supplied to the firm. A type-targeting effect changes the association between worker characteristics and preference and access vectors, thereby changing $\widetilde{Z}_{jt}$

conditional on the firm's effective-labor share.

The cost function $\Phi_{jt}$ in the firm's problem is a general reduced-form non-wage cost. To give its dependence on workforce composition a primitive interpretation, let

$$\widehat{\Phi}_{jt}\Big(L_{jt}, L_{j,t-1}, K_{j,t+1}, K_{jt}, Z_{jt}, \widetilde{Z}_{jt}, \widetilde{Z}_{j,t-1}, \rho_{jt}; \boldsymbol{\rho}_{-j,t}, \Phi_j\Big)$$

denote the firm's total non-wage cost before optimizing over the policies $\rho_{jt}$. This primitive cost may include labor and capital adjustment, hiring and dismissal, recruiting, screening, training, retention, management, organization, congestion, and other non-wage costs. The reduced-form cost in the main text can be written as

$$\begin{aligned}&\Phi_{jt}\Big(L_{jt}, L_{j,t-1}, K_{j,t+1}, K_{jt}, Z_{jt}, \widetilde{Z}_{jt}, \widetilde{Z}_{j,t-1}, \overline{\Phi}_j\Big)\\ &\quad = \inf_{\rho_{jt}} \Big\{ \widehat{\Phi}_{jt}\Big(L_{jt}, L_{j,t-1}, K_{j,t+1}, K_{jt}, Z_{jt}, \widetilde{Z}_{jt}, \widetilde{Z}_{j,t-1}, \rho_{jt}; \boldsymbol{\rho}_{-j,t}, \Phi_j\Big) : \mathcal{Z}_j\big(L_{jt}, \boldsymbol{\delta}_t, (\rho_{jt}, \boldsymbol{\rho}_{-j,t})\big) = \widetilde{Z}_{jt} \Big\}.\end{aligned} \tag{58}$$

Thus, optimizing over workforce policies is one component of the reduced-form cost function. It includes recruiting, screening, training, retention, organization, congestion, hiring and dismissal, and costs associated with changing the number or type of employees. Under the usual regularity conditions, the envelope theorem maps derivatives of $\Phi_{jt}$ into the corresponding derivatives of the minimized primitive implementation cost.

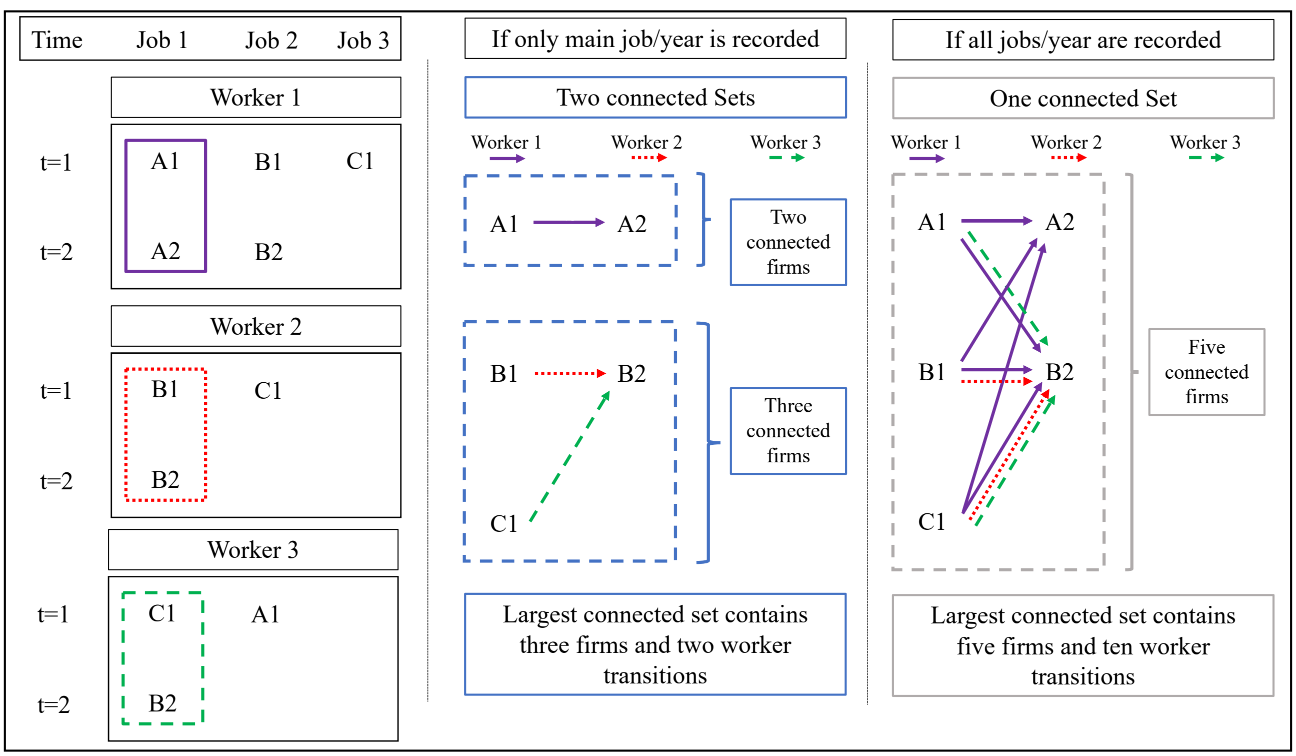
# D AKM Estimation, Connected Sets, and Ability Measurement

The identification of the coefficients on the individual covariates, $\Gamma_t$, the individual fixed effects $\alpha_i$, and the firm-time fixed effects $\psi_{j(i,t)t}$ in the two-way fixed effects model with time-varying firm effects follows Chan et al. (2026). Given the connected set, the standard identifying restriction is that worker mobility is orthogonal to the residual $\xi_{ijt}$ in the wage equation. A related concern is that the model may face limited mobility bias. The richness of our data helps mitigate this issue, as illustrated in Figure D.1.

Figure D.1 provides a graphical representation of how we construct the connected set in our model with firm-time fixed effects and multiple observations per worker within a year. The left panel shows a theoretical set of jobs across two years for three different workers. In this example, worker 1 has three jobs in period 1, working at firms A, B, and C. In period 2, worker 1 has only two jobs—working at firms A and B—and so on. Workers' main jobs—those that provide the largest income in a year—are identified by solid purple, dotted red, and dashed green boxes. Because we estimate firm effects separately for each year, we treat each firm-year observation as a separate firm. Thus, A1 and A2 refer to firm A in periods 1 and 2, respectively.

The middle panel shows the network graph of all five firm-year nodes if we were to consider only each worker's primary job (as is typically done in the literature). For example, worker 1 moves from A1 in period 1 to A2 in period 2, while worker 3 moves from C1 to B2. The result is two connected sets; the first, with firms A1 and A2, and the second (the largest connected set) with firms B1, B2 and C1. Each firm is connected to the rest of the set with just a single worker transition.

Figure D.1: CONNECTED SETS USING MULTIPLE JOBS




*Notes:* Figure D.1 illustrates how multiple job observations increase connectedness in the two-way fixed effects model with firm-year fixed effects.

The right panel of Figure D.1 shows the network graph when we consider all worker-firm pair information available in a year. In this case, each first-period job worked by an individual is connected to every second-period job worked by that same individual. For example, worker 1 is employed in firms A1, B1, and C1 in period 1, and A2 and B2 in period 2, leading to a full set of connections between period 1 and period 2 firms as depicted by the multiple solid lines in the right panel of Figure D.1. This leads to a connected set including all five firm-year nodes in the sample. Moreover, each firm-year node in this larger connected set is connected by at least 3 worker transitions to the rest of the set, strengthening the identification of the firm and worker fixed effects.

A common concern in AKM fixed-effects regressions is that there may be many firm-time pairs that are weakly connected to the largest connected set. For example, in our data set, we find that roughly 5.2% of all firm-time observations have only one transition connecting these firms to the largest connected set, with another 13.3% of firm-time observations having only two connections. As noted by Andrews et al. (2008), if firm or firm-time fixed effects are identified using a small number of workers who move across firms, the AKM estimates may be biased, overstating the role of firms relative to the role of sorting in accounting for the variation in labor earnings. Notice, however, that using multiple job observations for workers helps to reduce the extent of this limited mobility bias: if we include only one job per worker, we find that 7.0% of firm-time observations have only one link (versus 5.2% in our baseline sample), and 20.1% have only two links to the largest connected set (versus 13.3% using workers' top three job observations).

Limited mobility bias is most directly a concern for the estimated firm-time effects and the sorting covariance. Our main use of the AKM estimates is to recover the worker component, $\alpha_i + X_{it}\Gamma_t$, which is used to construct ability-adjusted labor input for the production-function estimation. The larger connected set obtained from multiple job observations improves the precision of both worker and firm components. Chan et al. (2026) further address the limited mobility bias problem by removing firm-time observations with low numbers of connections and estimating the parameters on several restricted connected sets. Their sensitivity tests suggest that limited mobility bias in this two-way fixed effects analysis using Danish administrative data is small and not quantitatively important for the objects used in our production-function estimation.

As is standard in the literature, we can decompose the variance of the log hourly wages as follows

$$Var\left(w_{ijt}\right) = \underbrace{Var\left(\alpha_i + X_{it}\Gamma_t\right)}_{\text{Worker Component}} + \underbrace{Var\left(\psi_{j(i,t)t}\right)}_{\text{Firm Component}} + \underbrace{2 \times Cov\left(\alpha_i + X_{it}\Gamma_t, \psi_{j(i,t)t}\right)}_{\text{Wage Sorting Component}} + \underbrace{Var\left(\xi_{ijt}\right)}_{\text{Residual}}, \tag{59}$$

where the first and second components capture the variation in log hourly wages accounted for by heterogeneity across workers and firms, respectively. The third component accounts for the variation in log hourly wages that can be attributed to the sorting of high-ability workers—as measured by $\alpha_i + X_{it}\Gamma_t$—employed by high-wage firms—as measured by $\psi_{j(i,t)t}$. The results are shown in Table D.1. We find that around one-half of the variance of log hourly wages is accounted for by workers' observed and unobserved characteristics and about 16% is accounted for by firms' time-varying characteristics. The estimated AKM wage-sorting covariance is close to zero. This means that the estimated worker component is nearly uncorrelated with the estimated firm-time wage component. However, it does not imply that firms employ similar ability distributions or that worker ability is unrelated to other firm attributes. Workers may sort across firms by industry, location, amenities, recruiting access, organizational structure, or other characteristics even when the AKM worker–firm wage covariance is small. For example, high-ability (worker component) workers may choose to work at high-productivity firms with potentially higher firm wage components. However, if high productivity firms also pay workers a lower share of their MRPL (as we document), this may counteract any positive correlation between the firm and worker components.

Table D.1: VARIANCE DECOMPOSITION OF LOG HOURLY WAGES USING AKM

| | 1991–2000 | 2001–2010 | Pooled |
|---|---|---|---|
| *Variance components* | | | |
| Total variance, $Var(w_{ijt})$ | 0.287 | 0.315 | 0.302 |
| Worker heterogeneity, $Var(\alpha_i + X_{it}\Gamma_t)$ | 0.138 | 0.155 | 0.148 |
| Firm heterogeneity, $Var(\psi_{j(i,t)t})$ | 0.045 | 0.052 | 0.049 |
| Wage sorting, $2 \times Cov(\alpha_i + X_{it}\Gamma_t, \psi_{j(i,t)t})$ | -0.001 | -0.003 | -0.002 |
| Residual, $Var(\xi_{ijt})$ | 0.106 | 0.110 | 0.108 |
| *Connected set and fit* | | | |
| Largest connected set | 98.0% | 99.0% | 99.0% |
| Adjusted $R^2$ | 62.0% | 63.0% | 63.0% |

*Notes:* Table D.1 decomposes the variance of log hourly wages using the AKM estimates, as in Equation (59), for two time intervals and for the pooled sample. "Largest connected set" reports the share of the sample contained in the largest connected set. The pooled sample contains 27.3 million worker-job-year observations, 4.3 million unique workers, 3.0 million unique firm-years, and 0.45 million unique firms. Variance components may not sum exactly to total variance because of rounding.

# E Additional Estimation Details and Results

## E.1 Revenue Production Function Estimation

This subsection provides the identification and estimation details underlying Section 4.2. Our approach builds on the non-parametric identification strategy of Gandhi, Navarro, and Rivers (2020) (GNR), modified to accommodate ability-adjusted labor and a labor input that is chosen dynamically.

**Intermediate-input share equation.** Recall that the firm's FOC with respect to intermediate inputs is

$$\frac{\partial R(K_{jt}, L_{jt}, M_{jt})}{\partial M_{jt}} e^{\omega_{jt}} \mathcal{E} = P_t^M,$$

where $\mathcal{E} \equiv \mathbb{E}[e^{\epsilon_{jt}}]$. Multiplying both sides by $M_{jt}/R_{jt}$, using

$$R_{jt} = R(K_{jt}, L_{jt}, M_{jt}) e^{\omega_{jt}+\epsilon_{jt}},$$

and then taking logs gives

$$s_{jt} = \ln \mathcal{E} + \ln D(k_{jt}, l_{jt}, m_{jt}) - \epsilon_{jt} \equiv \ln D^{\mathcal{E}}(k_{jt}, l_{jt}, m_{jt}) - \epsilon_{jt}, \tag{60}$$

where

$$s_{jt} \equiv \ln\left(\frac{P_t^M M_{jt}}{R_{jt}}\right), \qquad D(k_{jt}, l_{jt}, m_{jt}) \equiv \frac{\partial r(k_{jt}, l_{jt}, m_{jt})}{\partial m_{jt}},$$

and $D^{\mathcal{E}} \equiv \mathcal{E} D$.[39]

Because period-$t$ inputs are chosen conditional on $\mathcal{I}_{jt}$ and $\mathbb{E}[\epsilon_{jt} \mid \mathcal{I}_{jt}] = 0$, we have

$$\mathbb{E}[\epsilon_{jt} \mid k_{jt}, l_{jt}, m_{jt}] = 0.$$

[39] As in GNR, we assume that $R(K_{jt}, L_{jt}, M_{jt})$ admits the log representation $r(k_{jt}, l_{jt}, m_{jt})$.

Equation (60) therefore identifies $\ln D^{\mathcal{E}}(k_{jt}, l_{jt}, m_{jt})$ from the conditional mean of the intermediate-input share. Given

$$\epsilon_{jt} = \ln D^{\mathcal{E}}(k_{jt}, l_{jt}, m_{jt}) - s_{jt},$$

the residual identifies the distribution of $\epsilon_{jt}$ and hence

$$\mathcal{E} = \mathbb{E}[e^{\epsilon_{jt}}].$$

The intermediate-input elasticity is then recovered as

$$D(k_{jt}, l_{jt}, m_{jt}) = \frac{D^{\mathcal{E}}(k_{jt}, l_{jt}, m_{jt})}{\mathcal{E}}.$$

We estimate Equation (60) by non-linear least squares using a second-degree polynomial sieve.[40]

**Recovering the revenue production function.** Because

$$D(k_{jt}, l_{jt}, m_{jt}) = \frac{\partial r(k_{jt}, l_{jt}, m_{jt})}{\partial m_{jt}},$$

the log revenue production function can be written as

$$r(k_{jt}, l_{jt}, m_{jt}) = \mathcal{D}(k_{jt}, l_{jt}, m_{jt}) - \Psi(k_{jt}, l_{jt}),$$

where

$$\mathcal{D}(k_{jt}, l_{jt}, m_{jt}) \equiv \int D(k_{jt}, l_{jt}, m_{jt})\, dm_{jt}$$

and $\Psi(k_{jt}, l_{jt})$ is an unknown integration function. Observed log revenue satisfies

$$r_{jt} = r(k_{jt}, l_{jt}, m_{jt}) + \omega_{jt} + \epsilon_{jt}.$$

Define

$$\widetilde{r}_{jt} \equiv r_{jt} - \epsilon_{jt} - \mathcal{D}(k_{jt}, l_{jt}, m_{jt}).$$

It follows that

$$\widetilde{r}_{jt} = -\Psi(k_{jt}, l_{jt}) + \omega_{jt}.$$

Using the productivity process $\omega_{jt} = h(\omega_{j,t-1}) + \eta_{jt}$ gives the second-stage estimating equation

$$\widetilde{r}_{jt} = -\Psi(k_{jt}, l_{jt}) + h(\widetilde{r}_{j,t-1} + \Psi(k_{j,t-1}, l_{j,t-1})) + \eta_{jt}. \tag{61}$$

The variable $\widetilde{r}_{jt}$ is observable given the first-stage estimates of $\epsilon_{jt}$ and $\mathcal{D}(k_{jt}, l_{jt}, m_{jt})$. We estimate $\Psi$ and $h$ jointly using complete polynomial sieves and the moment restrictions described below.[41]

**Labor endogeneity and instruments.** The current labor input $l_{jt}$ is endogenous in Equation (61). Firms observe the persistent productivity innovation $\eta_{jt}$ before choosing labor, so $l_{jt}$ is generally correlated

[40] We follow GNR in using a second-degree polynomial sieve to recover $D(k_{jt}, l_{jt}, m_{jt})$. This specification also gives a closed-form solution for its integral with respect to $m_{jt}$. See GNR and Chen (2007) for further details.

[41] We allow $h$ to vary across two-digit industries but assume that it is common across firms within an industry. We estimate $\Psi$ and $h$ as complete second- and third-degree polynomials, respectively.

with $\eta_{jt}$. The timing assumptions of the model imply

$$\mathbb{E}\Big[\eta_{jt} \mid k_{jt}, l_{j,t-1}, k_{j,t-1}, Z_{jt}, \widetilde{Z}_{j,t-1}\Big] = 0.$$

We can therefore use functions of

$$\Big\{k_{jt}, l_{j,t-1}, k_{j,t-1}, Z_{jt}, \widetilde{Z}_{j,t-1}, \widetilde{r}_{j,t-1}\Big\}$$

as instruments for Equation (61).

Lagged workforce composition plays a particularly important role. By assumption, $\widetilde{Z}_{j,t-1}$ does not enter the current revenue production function directly, providing the exclusion restriction. At the same time, because labor is costly to adjust and employ, current employment depends on lagged workforce composition through the firm's dynamic optimization problem. Thus, $\widetilde{Z}_{j,t-1}$ has predictive power for $l_{jt}$ beyond that contained in $(k_{jt}, l_{j,t-1}, k_{j,t-1}, Z_{jt})$.[42]

This is the principal identification difference from GNR, where labor is predetermined and therefore does not require additional instruments. It also differs from much of the production-function-based market-power literature, where labor is treated as fully flexible; see, for example, Yeh et al. (2022).

**Measurement and construction of productivity and MRPL.** We measure inputs and output following standard practice in the production-function literature (Syverson, 2011). Revenues are deflated using industry-level price indices, the capital stock is constructed using the perpetual-inventory method, and intermediate inputs are measured as total expenditures on materials, energy, and other non-labor inputs.

Our main departure from standard practice is the construction of labor as the ability-weighted sum of total hours,

$$L_{jt} = \sum_{i \in j} A_{it} H_{ijt}.$$

This measure has two advantages. First, it does not impose homogeneous ability across workers or perfect competition in the labor market when estimating the production function. Second, it accounts for endogenous changes and cross-sectional differences in the ability composition of the workforce and removes them from measured firm productivity.

The estimates provide the non-parametric revenue production function, the revenue elasticities of each input, and the firm-level distributions of $\omega_{jt}$, $\eta_{jt}$, and $\epsilon_{jt}$. In particular, the firm-level MRPL can be constructed analytically as

$$MRPL_{jt} = \frac{\partial r(k_{jt}, l_{jt}, m_{jt})}{\partial l_{jt}} \frac{R_{jt}}{L_{jt}},$$

which gives the markdown

$$\mu_{jt} = \frac{W_{jt}}{MRPL_{jt}}.$$

Because worker $i$ contributes $A_{it}$ efficiency units per hour, the corresponding worker-level marginal revenue product is

$$MRPL_{ijt} = A_{it} MRPL_{jt}.$$

[42] Our second-stage moments use capital, squared capital, and lags of mean worker experience, worker age, and market share, together with lags of the shares of workers who have a college degree, hold management positions, are white-collar workers, or were employed by the firm in the previous period.

Our productivity measures are revenue-productivity rather than quantity-productivity terms and therefore combine variation in production efficiency with variation in output demand, as discussed in Section 4.2.

## E.2 Ex-post Wage Adjustment Function Estimation

To construct our markdown term free of the ex-post shock in Section 5.1, we estimate the ex-post wage adjustment function $f^c(\epsilon_{jt})$ as a flexible function of $\epsilon_{jt}$. Specifically, we directly estimate the firm-level wages $w_{jt}$ as a second-degree polynomial function of the ex-post shock $\epsilon_{jt}$. Since the ex-post shocks are independent of the firm's information set in period $t$, we have $\mathbb{E}[w^c_{jt}|\epsilon_{jt}] = \mathbb{E}[w^c_{jt}]$. Hence, projecting $w_{jt}$ on a polynomial in $\epsilon_{jt}$ identifies $\log f^c(\epsilon_{jt})$ up to an additive constant. Table E.1 shows the results. The first column shows the coefficient on $\epsilon_{jt}$ and the second column shows the coefficient on $\epsilon^2_{jt}$. The magnitudes and signs of the coefficients imply that $f^c$ is increasing and concave over the empirical support of $\epsilon_{jt}$. In columns 3 and 4, we show the mean and standard deviation of $f^c(\epsilon_{jt})$. We consider several other specifications for $f^c$, including more flexible polynomials and specifications with interactions for positive and negative shocks. The results remain largely unaffected. We can then construct the expected wage, $\overline{W}_{jt}$, the expected MRPL, $\overline{MRPL}_{jt}$, and the expected markdown, $\overline{\mu}_{jt}$. We report the distributions of these variables (in logs) along with $\epsilon_{jt}$ and $f^c(\epsilon_{jt})$ in Table E.2.

Table E.1: Estimation of Function $f^c(\epsilon_{jt})$

| Coeff. on $\epsilon_{jt}$ | Coeff. on $\epsilon^2_{jt}$ | Mean of $f^c(\epsilon_{jt})$ | Std. Dev. of $f^c(\epsilon_{jt})$ |
|---|---|---|---|
| 0.396 | -0.323 | 0.990 | 0.109 |

*Notes:* Table E.1 reports the estimated coefficients of the ex-post shock function. The first two columns report coefficients from the quadratic approximation to $\log f^c(\epsilon_{jt})$. The last two columns report moments of the implied distribution of $f^c(\epsilon_{jt})$; the mean corresponds to $\mathcal{E}^c$.

Table E.2: Ex-post Shocks and Expected Values

| Statistic | $\overline{w}_{jt}$ | $\overline{mrpl}_{jt}$ | $\overline{\mu}_{jt}$ | $\epsilon_{jt}$ | $f^c(\epsilon_{jt})$ | $\log f^c(\epsilon_{jt})$ |
|---|---|---|---|---|---|---|
| Mean | 6.60 | 6.86 | 0.81 | -0.01 | 0.99 | -0.02 |
| Std. Dev. | 0.25 | 0.23 | 0.27 | 0.19 | 0.11 | 0.11 |
| p10 | 6.30 | 6.58 | 0.54 | -0.21 | 0.87 | -0.14 |
| p25 | 6.47 | 6.75 | 0.65 | -0.11 | 0.93 | -0.07 |
| p50 | 6.62 | 6.88 | 0.77 | -0.02 | 0.99 | -0.01 |
| p75 | 6.75 | 6.99 | 0.92 | 0.08 | 1.05 | 0.05 |
| p90 | 6.88 | 7.09 | 1.10 | 0.21 | 1.12 | 0.12 |
| p99 | 7.25 | 7.42 | 1.83 | 0.57 | 1.29 | 0.25 |
| Observations | | | 374,000 | | | |

*Notes:* Table E.2 reports moments of the ex-ante expected firm wage, MRPL and markdown, along with the ex-post shock and wage adjustment function. Expected wages and MRPL are in logs, while expected markdowns are in levels. The number of observations is rounded to the nearest thousand. Percentiles are calculated as the mean of adjacent milliles.

## E.3 Alternate Labor Supply Elasticity Estimation

To complement the structural model of labor supply we posit and estimate in Section 6.1, here we leverage the structure and timing assumptions of our model to directly estimate a reduced-form model of labor supply

Table E.3: ESTIMATION OF REDUCED FORM LABOR SUPPLY ELASTICITY

| | Log-linear | | Polynomial | |
|---|---|---|---|---|
| | OLS | IV | OLS | IV |
| $E(\varepsilon_W^L)$ | 0.053 | 1.781 | 0.019 | 2.746 |
| | (0.006) | (0.044) | – | – |
| $R^2$ | 0.527 | 0.411 | 0.539 | 0.529 |
| Observations | 374,000 | | | |

*Notes:* Table E.3 reports estimates from the labor supply elasticity procedure described in Section E.3. The instruments used in the IV regressions are $\omega_{jt-1}$ and $\eta_{jt}$. The log-linear columns report coefficients with standard errors in parentheses. The polynomial columns report means of the estimated elasticity distribution. The number of observations is rounded to the nearest thousand.

responses to wages using a natural set of demand shifters as instruments, similar to our baseline approach. Specifically, we estimate the following model equation:

$$L_{jt} = L(\overline{W}_{jt}, Z_{jt}) \tag{62}$$

As a complementary diagnostic, we estimate Equation (62) in two ways. First, we use a log-linear form to approximate the supply function, where we regress the firm's labor input in logs $\ell_{jt}$ on logs of the ability price $w_{jt}$, components of $Z_{jt}$ as above, and a set of time indicators. We do this both using OLS and 2SLS. Our instruments are exogenous variables which shift firm-level demand for labor – namely $\eta_{jt}$ and $\omega_{jt-1}$. The implicit assumption is that the demand shifters (in this case, idiosyncratic persistent productivity shocks and lagged levels) are not correlated with unobserved supply shifters, and therefore shift the demand curve alone – allowing us to trace out the supply curve with observed changes in wages and labor inputs. This gives us a single estimate of the average labor supply elasticity for our firms. Our second approach is to approximate the supply curve with a complete second-degree polynomial in (log) wages and $Z_{jt}$. We also do this using OLS and an IV approach. For the latter, we first predict each endogenous term in the polynomial supply curve (levels, squares, and interaction terms involving $w_{jt}$) using a complete second degree polynomial in the demand shifters $\{\eta_{jt}, \omega_{jt-1}\}$ and $Z_{jt}$, then estimate the supply curve with the predicted polynomial terms. This gives us a distribution of labor supply elasticities.

The results are shown in Table E.3. Columns 1 and 2 show the results from the log-linear approximation. Without instrumenting for wages (using OLS), we find a labor supply elasticity parameter of 0.053. Once we use our demand shifters to instrument for the wage, our estimated parameter increases to 1.781. Using the polynomial approximation, we estimate the equation, take derivatives with respect to the log wage, and find a mean labor supply elasticity of 0.019 (without demand shifters) and 2.746 (with demand shifters). This closely matches the mean labor supply elasticity we estimate using the structural nested logit model of labor supply in Section 6.1. Both the linear and polynomial results are consistent with the rest of the literature on estimating labor supply.

## E.4 Markdowns and Labor-Supply Elasticities in a Monopsony Model

One potential concern with the variance decomposition in Section 6 is that the relatively small contribution of the monopsony wedge could arise because the estimated labor supply model generates insufficient heterogeneity in firm-level elasticities. We assess this concern using a simple reverse-mapping exercise that does not rely on the functional form of our estimated nested-logit labor-supply model. Under the pure static-monopsony benchmark, the dynamic cost wedge is equal to one, and the expected markdown must satisfy

$$\mu_{jt} = \frac{\varepsilon^L_{W_{jt}}}{1 + \varepsilon^L_{W_{jt}}}.$$

Hence, for any observed markdown below one, the labor-supply elasticity that would be required to rationalize the markdown entirely through static monopsony is

$$\varepsilon^R_{jt} = \frac{\mu_{jt}}{1 - \mu_{jt}}.$$

For markdowns weakly above one, no positive finite labor-supply elasticity can rationalize the observation under the static benchmark.

Panel A of Table E.4 applies this inversion to selected percentiles of the expected markdown distribution. At the 10th percentile, an expected markdown of 0.54 requires an elasticity of 1.17, while the median markdown of 0.77 requires an elasticity of 3.35. The required elasticity then rises very rapidly as the markdown approaches one: the 75th-percentile markdown of 0.92 requires an elasticity of 11.50. At the 90th percentile, the expected markdown is 1.10, which lies outside the admissible range of the static-monopsony model. For comparison, our baseline estimated elasticity distribution has a mean of 2.87, a standard deviation of 0.47, and a 95th percentile of 3.71. Thus, reproducing the observed dispersion in markdowns through static monopsony alone would require substantially more dispersion in firm-level labor-supply elasticities than is present in our baseline estimates or in the alternative labor-supply specifications considered above.

Panel B performs a related exercise using within-firm markdown movements following large negative productivity innovations. We rank firm-year observations by the persistent productivity innovation ($\eta_{jt}$) and consider observations immediately below the 15th, 10th, 5th, and 1st percentiles of the shock distribution.[43] For each cutoff, we report the mean expected markdown among the 500 observations immediately below the cutoff, together with the mean and standard deviation of the one-period change in the markdown. We then use the static-monopsony mapping above to calculate the labor-supply elasticities that would be required before and after the shock if the observed markdown movement were generated entirely by a change in monopsony power.

For example, around the fifth-percentile cutoff ($\eta_{jt} = -0.0931$), the mean expected markdown is approximately 0.99 and its mean one-period change is 0.13. Using the rounded values, the corresponding pre-shock markdown is therefore approximately 0.86. A pure static-monopsony interpretation would require the firm's labor-supply elasticity to increase from approximately 6.33 before the shock to 120.21 after the shock, a nearly twentyfold increase in a single period. These magnitudes are far outside both the estimated cross-sectional variation and reasonable reasonable elasticity estimates in the literature. Generally, the results in Panel B show that rationalizing the response of markdowns to large negative firm-side productivity shocks through static monopsony alone requires very large and rapidly changing labor-supply elasticities. This exercise does not by itself identify the source of the residual wedge, but it shows that the limited contribution of monopsony to markdown dispersion is not simply a mechanical consequence of the particular labor-supply specification used in our baseline decomposition.

[43] For confidentiality reasons, percentiles are calculated as the mean of adjacent milliles.

Table E.4: Required Labor-Supply Elasticities under Pure Static Monopsony

| | Productivity Shock Cutoff $\eta$ | Expected Markdown $\mu$ | $\Delta$ Expected Markdown $\Delta\mu$ | Required Elasticity $\varepsilon^R$ | Baseline Estimated Elasticity $\varepsilon_W^L$ |
|---|---|---|---|---|---|
| *Panel A. Selected percentiles of the expected markdown distribution* | | | | | |
| p10 | — | 0.54 | — | 1.17 | 2.27 |
| p50 | — | 0.77 | — | 3.35 | 2.84 |
| p75 | — | 0.92 | — | 11.50 | 3.13 |
| p90 | — | 1.10 | — | N/A | 3.45 |
| *Panel B. Large negative productivity shocks* | | | | | |
| p15 | -0.0404 | 0.91 | 0.09 (0.24) | $4.70 \rightarrow 10.70$ | $2.86 \rightarrow 2.90$ |
| p10 | -0.0588 | 0.98 | 0.12 (0.34) | $6.50 \rightarrow 70.81$ | $2.85 \rightarrow 2.87$ |
| p5 | -0.0931 | 0.99 | 0.13 (0.25) | $6.33 \rightarrow 120.21$ | $2.80 \rightarrow 2.80$ |
| p1 | -0.1902 | 1.15 | 0.25 (0.27) | $9.19 \rightarrow$ N/A | $2.77 \rightarrow 2.73$ |

*Notes:* Panel A reports the labor-supply elasticity that would be required to rationalize selected percentiles of the expected markdown distribution entirely through static monopsony. Under the static-monopsony benchmark, $\mu = \varepsilon/(1+\varepsilon)$, so that the required elasticity is $\varepsilon^R = \mu/(1-\mu)$ for $\mu < 1$. Markdowns weakly greater than one cannot be rationalized by any positive finite labor-supply elasticity. The final column reports the corresponding percentile of the baseline estimated labor-supply elasticity distribution as a distributional benchmark; markdown and elasticity percentiles need not refer to the same firms. Panel B focuses on firms experiencing large negative persistent productivity innovations. For each cutoff, the sample consists of the 500 firm-year observations immediately below the indicated percentile of the productivity shock distribution. $\Delta\mu_t = \mu_t - \mu_{t-1}$; standard deviations of $\Delta\mu_t$ are reported in parentheses. The required-elasticity column reports the static-monopsony elasticities implied by the mean pre- and post-shock markdowns, $\varepsilon_{t-1}^R \rightarrow \varepsilon_t^R$. The values are calculated using unrounded moments (columns 2 and 3 report rounded moments). The final column reports the mean baseline-model labor-supply elasticity for the same firm-year observations.

## E.5 Dynamic Cost Wedge Estimation

This subsection estimates the dynamic component underlying the cost wedge. Recall that we allow firms to face a general cost function that may include fixed costs, adjustment costs in labor and capital, as well as non-wage variable costs that may depend on the level of labor or capital. The remaining model object we want to estimate is the dynamic net marginal benefit term,

$$B_{jt}^{\phi} \equiv \left( \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}}{\partial L_{jt}} \right).$$

In order to do this, we impose some additional structure on the cost function. We assume that the labor-relevant component of the cost function, $\Phi_{jt}^{L}$, depends on changes (in levels) in the labor input $\Delta L_{jt}$ and workforce composition terms $\Delta \widetilde{Z}_{jt}$, as well as the levels of labor $L_{jt}$, job amenities $u_{jt}$, workforce composition $\widetilde{Z}_{jt}$ and other predetermined firm characteristics $Z_{jt}$. We also allow firms to differ in their fundamental marginal cost of employing labor, $\omega_j^{\Phi}$, which captures the fact that some firms may be more or less efficient at managing, training, or retaining labor.

We then take two approaches to estimating the dynamic net marginal benefit term. First, we take a parametric approach and assume that $\Phi_{jt}^{L}$ takes the following form:

$$\Phi_{jt}^{L} = \alpha_1 (L_{jt} - L_{j,t-1})^2 + L_{jt} \big[ (\widetilde{Z}_{jt} - \widetilde{Z}_{j,t-1})^2 \Gamma_1 + \widetilde{Z}_{jt} \Gamma_2 + Z_{jt} \Gamma_3 + \alpha_2 u_{jt} + \omega_j^{\Phi} \big]. \tag{63}$$

The derivative of the current cost function with respect to $L_{jt}$, holding workforce composition fixed, is then

$$\frac{\partial \Phi_{jt}^{L}}{\partial L_{jt}} = 2\alpha_1 (L_{jt} - L_{j,t-1}) + (\widetilde{Z}_{jt} - \widetilde{Z}_{j,t-1})^2 \Gamma_1 + \widetilde{Z}_{jt} \Gamma_2 + Z_{jt} \Gamma_3 + \alpha_2 u_{jt} + \omega_j^{\Phi}. \tag{64}$$

The other term in the dynamic component is $\partial \overline{V}_{j,t+1} / \partial L_{jt}$, where

$$\overline{V}_{j,t+1} \equiv \beta E_t \left[ V_{j,t+1}(\mathcal{I}_{j,t+1}) \right].$$

The state variable $L_{jt}$ enters the next-period problem as lagged labor. Under the parametric cost function above, it affects next-period costs through the adjustment-cost term $\alpha_1 (L_{j,t+1} - L_{jt})^2$. By the envelope condition,

$$\frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} = -\beta E_t \left[ \frac{\partial \Phi_{j,t+1}^{L}}{\partial L_{jt}} \right] = 2\beta \alpha_1 (\overline{L}_{j,t+1} - L_{jt}), \tag{65}$$

where

$$\overline{L}_{j,t+1} \equiv E_t [L_{j,t+1}].$$

Combining Equations (64) and (65), we can write the net marginal benefit term as

$$\begin{aligned} B_{jt}^{\phi} &\equiv \frac{\partial \overline{V}_{j,t+1}}{\partial L_{jt}} - \frac{\partial \Phi_{jt}^{L}}{\partial L_{jt}} \\ &= 2\beta \alpha_1 (\overline{L}_{j,t+1} - L_{jt}) - 2\alpha_1 (L_{jt} - L_{j,t-1}) \\ &\quad - (\widetilde{Z}_{jt} - \widetilde{Z}_{j,t-1})^2 \Gamma_1 - \widetilde{Z}_{jt} \Gamma_2 - Z_{jt} \Gamma_3 - \alpha_2 u_{jt} - \omega_j^{\Phi}. \end{aligned} \tag{66}$$

Equivalently,

$$B^{\phi}_{jt} = 2\alpha_1 \left[\beta \overline{L}_{j,t+1} - (1+\beta)L_{jt} + L_{j,t-1}\right] - (\widetilde{Z}_{jt} - \widetilde{Z}_{j,t-1})^2\Gamma_1 - \widetilde{Z}_{jt}\Gamma_2 - Z_{jt}\Gamma_3 - \alpha_2 u_{jt} - \omega^{\Phi}_{j}.$$

We can estimate the dynamic net marginal benefit term using our estimates of the labor supply elasticity, wage, and MRPL. We assume that expected MRPL contains classical measurement error, such that

$$\overline{MRPL}^{true}_{jt} = \overline{MRPL}_{jt} + \xi^{ME}_{jt}.$$

The expected wage first-order condition is then

$$\overline{W}_{jt} = \frac{\varepsilon^{L}_{\overline{W}jt}}{1+\varepsilon^{L}_{\overline{W}jt}} \left(\overline{MRPL}_{jt} + \xi^{ME}_{jt} + B^{\phi}_{jt}\right). \tag{67}$$

Let

$$\mu^{\varepsilon}_{jt} \equiv \frac{\varepsilon^{L}_{\overline{W}jt}}{1+\varepsilon^{L}_{\overline{W}jt}}.$$

Rearranging gives the estimating equation

$$\frac{\overline{W}_{jt}}{\mu^{\varepsilon}_{jt}} - \overline{MRPL}_{jt} = B^{\phi}_{jt} + \xi^{ME}_{jt}. \tag{68}$$

Under the assumptions above, and after absorbing the sign of the primitive firm marginal-cost shifter into a firm fixed effect, we can rewrite this equation as

$$B^{\phi}_{jt} = D^{\phi}(\Delta L_{jt}, \Delta\widetilde{Z}_{jt}, Z_{jt}, \overline{L}_{j,t+1}, L_{jt}, \widetilde{Z}_{jt}, u_{jt}) + \omega^{\phi}_{j} + \xi^{ME}_{jt}. \tag{69}$$

Here $D^{\phi}(\cdot)$ is the observable component implied by Equation (66), and $\omega^{\phi}_{j} \equiv -\omega^{\Phi}_{j}$ is the corresponding firm-specific net-benefit fixed effect. We estimate Equation (69) with a linear OLS regression with firm fixed effects. Our second approach to estimating the dynamic cost wedge is to relax the parametric assumptions and estimate $D^{\phi}(\cdot)$ as a complete second-degree polynomial in the same variables.

One final issue is that expected labor next period, $\overline{L}_{j,t+1}$, is not observed in period $t$. However, it is a function of the period $t$ information set, $\mathcal{I}_{jt}$, plus the period $t$ choice variables $K^{I}_{jt}$, $L_{jt}$, and $\widetilde{Z}_{jt}$, and the exogenous distributions of future shocks. Thus we include the period $t$ information set in the regressions as a proxy for expected future labor. In particular, we additionally include TFP shocks and levels, the capital stock, and the net investment rate, as these are potentially strong determinants of the expected future value of labor while not directly entering current marginal labor costs in the parametric specification.

Estimating Equation (69) provides the firm-level distribution of predicted net marginal benefits and the latent firm-level component $\omega^{\phi}_{j}$. In Table E.5 we report moments of each component and the share of total variation accounted for by each of the empirical net marginal benefit term, $B^{\phi}_{jt}$. Overall, both the parametric and polynomial specifications predict the variation very well. Most of the variation in the estimated net marginal benefit term is accounted for by the component explained by the observed state variables in the model, $D^{\phi}(\cdot)$, rather than by the latent firm component or the residual. This is true under both the parsimonious parametric specification (52%) and the more flexible polynomial specification (64%), and is even more pronounced in the latter. The estimated model (the predicted and fixed-effect components) accounts for 86–88% of the

Table E.5: ESTIMATED NET MARGINAL BENEFIT COMPONENTS

| | Observed | Parametric | | | Polynomial | | |
|---|---|---|---|---|---|---|---|
| Statistic | (1) $B^{\phi}_{jt}$ | (2) $D^{\phi}_{jt}$ | (3) $\omega^{\phi}_{j}$ | (4) $\xi^{ME}_{jt}$ | (5) $D^{\phi}_{jt}$ | (6) $\omega^{\phi}_{j}$ | (7) $\xi^{ME}_{jt}$ |
| Mean | 49 | 49 | 0 | 0 | 49 | 0 | 0 |
| Std. Dev. | 292 | 263 | 229 | 108 | 282 | 204 | 98 |
| p5 | -424 | -425 | -360 | -149 | -438 | -337 | -135 |
| p10 | -306 | -295 | -270 | -103 | -311 | -247 | -93 |
| p25 | -132 | -112 | -134 | -48 | -126 | -116 | -43 |
| p50 | 42 | 63 | 6 | -4 | 53 | 7 | 3 |
| p75 | 219 | 233 | 138 | 40 | 227 | 118 | 36 |
| p90 | 413 | 399 | 262 | 101 | 405 | 230 | 91 |
| p95 | 577 | 514 | 349 | 158 | 529 | 313 | 143 |
| Variance contribution | 100% | 52% | 34% | 14% | 64% | 24% | 12% |

*Notes:* Table E.5 reports moments of the estimated net marginal benefit components from $B^{\phi}_{jt} = D^{\phi}(\Delta L_{jt}, \Delta\widetilde{Z}_{jt}, Z_{jt}, \overline{L}_{j,t+1}, L_{jt}, \widetilde{Z}_{jt}, u_{jt}) + \omega^{\phi}_{j} + \xi^{ME}_{jt}$, where $B^{\phi}_{jt} \equiv \overline{W}_{jt}/\mu^{\varepsilon}_{jt} - \overline{MRPL}_{jt}$. Column (1) reports the observed left-hand-side object. Columns (2)–(4) report the fitted non-fixed-effect component, firm fixed effect, and residual from the parametric specification. Columns (5)–(7) report the corresponding objects from the flexible polynomial specification, in which $D^{\phi}(\cdot)$ is estimated as a complete second-degree polynomial in the same variables. The final row reports each component's contribution to $\mathrm{Var}(B^{\phi}_{jt})$. If the fitted observable and firm-effect components are not orthogonal, covariance terms are allocated equally across the corresponding components. Percentiles are calculated as the mean of adjacent milliles.

total variance in the empirical net marginal benefit term. In other words, the dynamic cost wedge is largely systematic rather than being a misspecification residual, and is closely linked to the same state variables that appear in the model, including labor adjustment, workforce composition, expected future labor, firm characteristics, and amenities. The polynomial specification increases the share of variation explained by observables in levels, while reducing the role of the fixed effect, suggesting that part of what appears as permanent heterogeneity under the parametric model is in fact due to nonlinearities in observables.

While $B^{\phi}_{jt}$ represents the level of the net marginal benefit of labor, it is the net marginal benefit relative to the firm's MRPL (i.e., the cost wedge $\mu^{\phi}_{jt}$) that is more economically meaningful. For example, larger firms may have larger marginal net benefits or costs, but be less distorted because they have even greater marginal labor productivity. We construct the estimated components of the cost wedge by dividing Equation (69) by $\overline{MRPL}_{jt}$ as

$$\mu^{\phi}_{jt} - 1 = \widetilde{D}^{\phi}_{jt} + \widetilde{\omega}^{\phi}_{jt} + \widetilde{\xi}^{ME}_{jt}, \tag{70}$$

where, for any component $X_{jt}$, we define $\widetilde{X}_{jt} \equiv X_{jt}/\overline{MRPL}_{jt}$. We report the moments and variance contributions of these cost-wedge terms in Table E.6.

## E.6 Additional Cost Wedge and Variance Decomposition Results

In this section we present additional results from the labor supply estimation and cost wedge decompositions, as well as robustness exercises using an alternative labor supply model.

First, we examine how our baseline cost wedge measure varies across industries. Table E.7 shows that cost wedges are highest in information, administrative services, professional services, and finance. These sec-

Table E.6: ESTIMATED COST WEDGE COMPONENTS

| | Observed | Parametric | | | Polynomial | | |
|---|---|---|---|---|---|---|---|
| Statistic | (1) $\mu^{\phi}_{jt} - 1$ | (2) $\widetilde{D}^{\phi}_{jt}$ | (3) $\widetilde{\omega}^{\phi}_{jt}$ | (4) $\widetilde{\xi}^{ME}_{jt}$ | (5) $\widetilde{D}^{\phi}_{jt}$ | (6) $\widetilde{\omega}^{\phi}_{jt}$ | (7) $\widetilde{\xi}^{ME}_{jt}$ |
| Mean | 0.10 | 0.08 | 0.01 | 0.00 | 0.08 | 0.01 | 0.00 |
| Std. Dev. | 0.36 | 0.31 | 0.25 | 0.14 | 0.31 | 0.24 | 0.13 |
| p5 | -0.34 | -0.42 | -0.36 | -0.16 | -0.41 | -0.36 | -0.14 |
| p10 | -0.26 | -0.29 | -0.27 | -0.11 | -0.30 | -0.26 | -0.10 |
| p25 | -0.12 | -0.12 | -0.14 | -0.05 | -0.13 | -0.12 | -0.04 |
| p50 | 0.05 | 0.06 | 0.01 | 0.00 | 0.05 | 0.01 | 0.00 |
| p75 | 0.24 | 0.24 | 0.15 | 0.04 | 0.25 | 0.12 | 0.04 |
| p90 | 0.47 | 0.47 | 0.28 | 0.11 | 0.48 | 0.25 | 0.10 |
| p95 | 0.68 | 0.67 | 0.39 | 0.18 | 0.66 | 0.35 | 0.16 |
| Variance contribution | 100% | 51% | 31% | 18% | 52% | 31% | 17% |

*Notes:* Table E.6 reports moments of the cost-wedge decomposition. For any component $X_{jt}$, define $\widetilde{X}_{jt} \equiv X_{jt}/\overline{MRPL}_{jt}$. Then $\mu^{\phi}_{jt} - 1 = \widetilde{D}^{\phi}_{jt} + \widetilde{\omega}^{\phi}_{jt} + \widetilde{\xi}^{ME}_{jt}$. Column (1) reports the observed distribution of $\mu^{\phi}_{jt} - 1$. Columns (2)–(4) report the corresponding components from the parametric specification, while columns (5)–(7) report the components from the flexible polynomial specification. The final row reports each component's contribution to $\mathrm{Var}(\mu^{\phi}_{jt})$. If the fitted observable and firm-effect components are not orthogonal, covariance terms are allocated equally across the corresponding components. Percentiles are calculated as the mean of adjacent milliles.

Table E.7: COST WEDGES BY INDUSTRY

| Industry | Mean | Std. Dev. | Industry | Mean | Std. Dev. |
|---|---|---|---|---|---|
| Information | 1.30 | 0.38 | Manufacturing | 1.08 | 0.29 |
| Administrative Services | 1.20 | 0.40 | Transportation | 1.06 | 0.33 |
| Professional Services | 1.19 | 0.36 | Real Estate | 1.04 | 0.38 |
| Finance | 1.18 | 0.39 | Other Services | 1.04 | 0.32 |
| Wholesale/Retail | 1.10 | 0.42 | Accommodation/Food | 0.96 | 0.30 |
| Construction | 1.09 | 0.27 | Mining | 0.83 | 0.38 |
| Total | 1.10 | 0.36 | | | |

*Notes:* Table E.7 reports industry-level means and standard deviations of the cost wedge, $\mu^{\phi}_{jt}$, by NACE-2 sector.

tors rely heavily on skilled workers, client relationships, and firm-specific knowledge, often with multiyear development time frames, so the shadow value of retaining labor is high. Cost wedges are lowest in accommodation and food and mining. In accommodation and food, labor services are more contemporaneous and firms can more easily adjust labor inputs, while safety, supervision, and regulatory costs may raise current non-wage marginal costs in the mining sector. The industry pattern therefore suggests that the cost wedge reflects persistent differences in workforce composition, expectations about future value of labor, adjustment frictions, and non-wage costs of labor.

To explore whether our variance decomposition result is driven solely by the chosen functional form, we calibrate the nested CES labor supply elasticity function used in Berger et al. (2022) to match our mean estimated labor supply elasticity of 2.87, obtain alternative measures of the monopsony and cost wedges, $\mu^{\varepsilon,ces}_{jt}$ and $\mu^{\phi,ces}_{jt}$ and repeat the variance decomposition from Section 6.2. In particular, we define $\varepsilon^{L,ces}_{W_{jt}} =$

Table E.8: ESTIMATED LABOR SUPPLY ELASTICITIES, MARKDOWNS, AND COST WEDGES

| *Panel A. Moments* | Labor wedge | Baseline LS specification | | | CES LS specification | | |
|---|---|---|---|---|---|---|---|
| Statistic | $\overline{\mu}_{jt}$ | $\varepsilon^{L}_{\overline{W}_{jt}}$ | $\mu^{\varepsilon}_{jt}$ | $\mu^{\phi}_{jt}$ | $\varepsilon^{L,\text{CES}}_{\overline{W}_{jt}}$ | $\mu^{\varepsilon,\text{CES}}_{jt}$ | $\mu^{\phi,\text{CES}}_{jt}$ |
| Mean | 0.81 | 2.87 | 0.74 | 1.10 | 2.87 | 0.70 | 1.22 |
| Std. Dev. | 0.27 | 0.47 | 0.03 | 0.36 | 1.14 | 0.14 | 0.56 |
| p5 | 0.48 | 2.14 | 0.68 | 0.66 | 0.51 | 0.34 | 0.64 |
| p10 | 0.54 | 2.27 | 0.69 | 0.74 | 0.90 | 0.47 | 0.73 |
| p25 | 0.65 | 2.57 | 0.72 | 0.88 | 2.11 | 0.68 | 0.89 |
| p50 | 0.77 | 2.84 | 0.74 | 1.05 | 3.28 | 0.77 | 1.09 |
| p75 | 0.92 | 3.13 | 0.76 | 1.24 | 3.83 | 0.79 | 1.37 |
| p90 | 1.10 | 3.45 | 0.78 | 1.47 | 4.02 | 0.80 | 1.85 |
| p95 | 1.25 | 3.71 | 0.79 | 1.68 | 4.07 | 0.80 | 2.32 |
| *Panel B. Variance decomposition of log expected markdown* | | | | | | | |
| Statistic | $\log(\overline{\mu}_{jt})$ | | $\log(\mu^{\varepsilon}_{jt})$ | $\log(\mu^{\phi}_{jt})$ | | $\log(\mu^{\varepsilon,\text{CES}}_{jt})$ | $\log(\mu^{\phi,\text{CES}}_{jt})$ |
| Share of variance (%) | 100.0 | – | 4.2 | 95.8 | – | 5.1 | 94.9 |

*Notes:* Panel A reports moments of the expected markdown, labor supply elasticity, monopsony wedge, and cost wedge. The expected markdown satisfies $\overline{\mu}_{jt} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}$. Panel B decomposes the variance of $\log(\overline{\mu}_{jt})$ into the labor-supply component and the dynamic cost component. Baseline columns use the nested-logit labor supply estimates. CES columns use the CES-based labor supply specification calibrated to match the mean baseline labor supply elasticity. Dashes indicate columns that do not enter the variance decomposition. The sample includes 374,000 firm-year observations, rounded to the nearest thousand. Percentiles are calculated as the mean of adjacent milliles.

$((\eta^{BHM})^{-1} + ((\theta^{BHM})^{-1} - (\eta^{BHM})^{-1}) * s^{BHM}_{gjt})^{-1}$, where the inside share is defined over effective labor (as in our baseline labor-supply model) at the industry-municipality-year level. We follow BHM here in using 3-digit industry codes. We choose the same degree of across-market substitution as in BHM, with $\theta^{BHM} = 0.42$, and choose the within-market substitutability parameter $\eta^{BHM}$ such that $E[\varepsilon^{L,ces}_{\overline{W}_{jt}}] = 2.87$. This gives us $\eta^{BHM} = 4.13$ and an alternative distribution of monopsony and cost wedges. The results are similar, with only 5.1% of the variance of markdowns attributed to variance in the monopsony wedge. We report the distributions and variance decomposition results from the CES calibration in columns 5-7 of Table E.8 and plot the distributions from both models in Appendix Figure E.1.

## E.7 Static Market Position and Dynamic Firm States

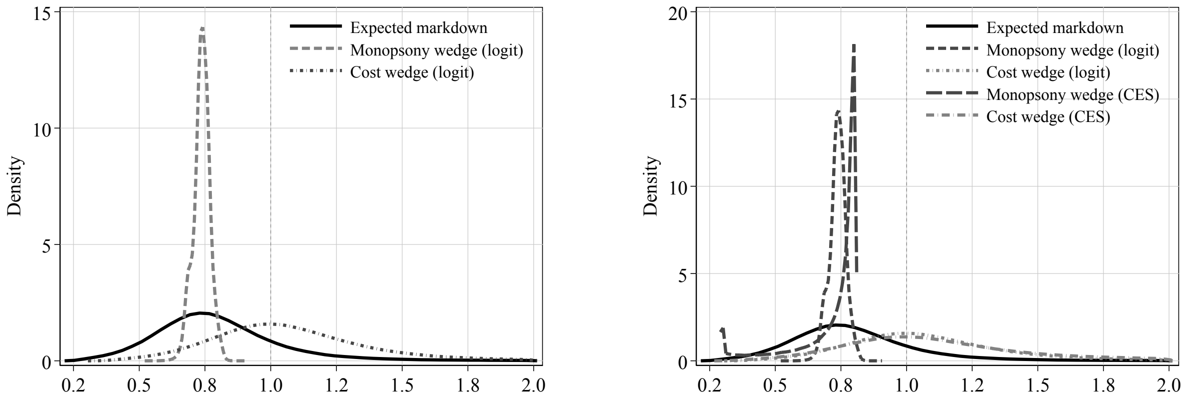
This section provides a reduced-form complement to the structural decomposition in Section 6.2. Rather than using the estimated labor-supply elasticity to construct the monopsony wedge, we ask how much markdown variation is predictable from observables associated with firms' static labor-market position and how much is predictable from firm states associated with the dynamic labor-demand problem.

Let

$$z_{jt} \equiv -\log \bar{\mu}_{jt},$$

and let $g(j,t)$ denote the municipality–four-digit-industry–year cell containing firm $j$ in year $t$. We first

Figure E.1: DENSITY PLOT OF MARKDOWN MEASURES




(a) Logit Wedge (b) Logit and CES

*Notes:* Figure E.1 plots the density of the ex-ante markdown ($\overline{W}_{jt}/\overline{MRPL}$), the logit monopsony wedge ($\mu^{\varepsilon}$), and the cost wedge ($\overline{W}_{jt}/(\mu^{\varepsilon}\overline{MRPL})$). It also plots the monopsony and cost wedges implied by a nested CES labor supply model.

decompose the total variation in $z_{jt}$ into between- and within-market-year components:

$$s_B = \frac{\sum_{jt}\left(\bar{z}_{g(j,t)} - \bar{z}\right)^2}{\sum_{jt}\left(z_{jt} - \bar{z}\right)^2}, \qquad s_W = \frac{\sum_{jt}\left(z_{jt} - \bar{z}_{g(j,t)}\right)^2}{\sum_{jt}\left(z_{jt} - \bar{z}\right)^2} = 1 - s_B.$$

We then estimate market-year fixed-effects regressions of the form

$$\begin{aligned} z_{jt} &= \gamma_{g(j,t)} + f_S(S_{jt}) + u_{jt}^{S}, \\ z_{jt} &= \gamma_{g(j,t)} + f_D(D_{jt}) + u_{jt}^{D}, \\ z_{jt} &= \gamma_{g(j,t)} + f_S(S_{jt}) + f_D(D_{jt}) + u_{jt}^{SD}, \end{aligned}$$

where $S_{jt}$ denotes a block of variables measuring static market position and $D_{jt}$ denotes a block of dynamic firm states. We repeat the same exercise for changes in the markdown.

The static block is deliberately broad. It includes firm shares of ability-adjusted labor and employment, corresponding within-market shares, wage-bill shares, firm size, market concentration, the number and size of firms in the local market, predetermined workforce characteristics, and interactions between firm shares and market structure. The workforce controls include lagged mean experience, age, ability, education, occupational composition, worker retention, and firm size. The dynamic block includes positive and negative persistent productivity innovations, lagged persistent productivity, lagged employment, employment growth, capital, and the investment rate.

We allow both blocks to enter flexibly. The linear specification enters the variables directly. The polynomial specification adds powers through order four for the principal continuous variables. The spline specification instead uses five-knot restricted cubic splines. For productivity innovations, the spline specification allows the fitted relationship to be nonlinear and asymmetric around zero.

Let $R_S^2$, $R_D^2$, and $R_{SD}^2$ denote the within-market $R^2$ from the static, dynamic, and joint specifications. We

Table E.9: STATIC MARKET POSITION AND DYNAMIC FIRM STATES

| | Levels | | | Changes | | |
|---|---|---|---|---|---|---|
| | Linear | Polynomial | Spline | Linear | Polynomial | Spline |
| *Variance decomposition* | | | | | | |
| Between-market share of total variance | 38.33 | 38.33 | 38.33 | 31.01 | 31.01 | 31.01 |
| Within-market share of total variance | 61.67 | 61.67 | 61.67 | 68.99 | 68.99 | 68.99 |
| *Within-market predictive fit* | | | | | | |
| Static block $R^2$ | 6.77 | 8.46 | 8.41 | 1.10 | 1.25 | 1.24 |
| Dynamic block $R^2$ | 12.28 | 15.99 | 15.59 | 11.68 | 13.60 | 11.73 |
| Both blocks $R^2$ | 22.59 | 26.01 | 23.80 | 14.81 | 16.91 | 13.84 |
| Partial $R^2$: dynamic \| static | 16.97 | 19.17 | 16.81 | 13.86 | 15.86 | 12.76 |
| Partial $R^2$: static \| dynamic | 11.75 | 11.94 | 9.73 | 3.54 | 3.83 | 2.39 |
| *General-dominance decomposition* | | | | | | |
| Static contribution to within $R^2$ | 8.54 | 9.25 | 8.31 | 2.11 | 2.28 | 1.67 |
| Static share of jointly explained variation | 37.79 | 35.54 | 34.91 | 14.27 | 13.49 | 12.09 |
| *Maximal monopsony attribution* | | | | | | |
| Within-market ceiling: $s_W R_S^2$ | 4.17 | 5.22 | 5.18 | 0.76 | 0.86 | 0.85 |
| Total ceiling: $s_B + s_W R_S^2$ | 42.51 | 43.55 | 43.52 | 31.77 | 31.88 | 31.87 |
| Within-market dominance allocation: $s_W G_S$ | 5.26 | 5.70 | 5.12 | 1.46 | 1.57 | 1.15 |
| Total dominance allocation: $s_B + s_W G_S$ | 43.60 | 44.04 | 43.46 | 32.47 | 32.59 | 32.17 |

Notes: Entries are percentages. The outcome is the log inverse expected markdown, $z_{jt} = -\log \bar{\mu}_{jt}$. Markets are municipality–four-digit-industry–year cells. The static block contains firm shares of ability-adjusted labor and employment, corresponding within-market shares, wage-bill shares, firm size, market-structure measures, predetermined workforce characteristics, and interactions between firm market position and market structure. The dynamic block contains persistent productivity innovations, lagged persistent productivity, lagged employment, employment growth, capital, and the investment rate. Polynomial specifications add powers through order four to the principal continuous variables; spline specifications use five-knot restricted cubic splines. Static, dynamic, and joint $R^2$s are calculated after removing market-year fixed effects. The partial $R^2$s measure the incremental explanatory power of one block conditional on the other. The general-dominance contribution averages the static block's incremental $R^2$ across the two possible block orderings, and the following row reports this contribution as a share of the joint $R^2$. Let $s_B$ and $s_W$ denote the between- and within-market variance shares, $R_S^2$ the static-only within $R^2$, and $G_S$ the static general-dominance contribution. The within-market and total ceilings are respectively $s_W R_S^2$ and $s_B + s_W R_S^2$. The final two rows replace $R_S^2$ by $G_S$.

also report the conditional partial $R^2$s

$$R^2_{D|S} = \frac{R^2_{SD} - R^2_S}{1 - R^2_S}, \qquad R^2_{S|D} = \frac{R^2_{SD} - R^2_D}{1 - R^2_D}.$$

Because the two blocks are not orthogonal, we additionally report a two-block general-dominance decomposition. The contribution assigned to the static block is

$$G_S = \frac{1}{2}\left[R^2_S + \left(R^2_{SD} - R^2_D\right)\right].$$

Thus, $G_S/R^2_{SD}$ measures the static block's share of the variation explained jointly by the two sets of observables.

Table E.9 shows a clear distinction between markdown levels and changes. In levels, 38.3% of total markdown variance occurs between market-year cells and 61.7% within them. The static market-position block explains 6.77–8.46% of within-market level dispersion across the three functional forms. Dynamic firm states are substantially more informative, explaining 12.28–15.99%, while the joint specifications explain 22.59–26.01%. Conditional on static market position, the dynamic block has a partial $R^2$ of 16.81–19.17%. Conditional on dynamic states, the corresponding partial $R^2$ of the static block is 9.73–11.94%. The general-dominance decomposition assigns 34.9–37.8% of jointly explained level variation to the static block and 62.2–65.1% to the dynamic block.

Static market position is much less informative about movements in markdowns. For changes, 31.0% of total variation occurs between market-year cells and 69.0% within them. The static block explains only 1.10–1.25% of within-market variation, compared with 11.68–13.60% for the dynamic block. The joint specifications explain 13.84–16.91%. Conditional on the static block, dynamic states have partial $R^2$s of 12.76–15.86%; conditional on dynamic states, the static partial $R^2$ is only 2.39–3.83%. The dominance decomposition accordingly assigns only 12.1–14.3% of jointly explained variation in markdown changes to the static block.

We can use these estimates to construct a deliberately generous accounting ceiling for the amount of markdown variation that could be attributed to monopsony. The within-market ceiling assigns all variation predicted by the static block to monopsony:

$$C_W^S = s_W R_S^2.$$

An even more generous total ceiling additionally attributes all between-market variation to monopsony:

$$C_T^S = s_B + s_W R_S^2.$$

These quantities are accounting bounds rather than structural variance decompositions.

For markdown levels, the within-market ceiling is 4.17–5.22% of total variance. This is close in magnitude to the 4–5% contribution of the monopsony wedge in the structural decomposition, despite the substantially broader set of variables included in the static block here. The total ceiling is much larger, 42.51–43.55%, but this number is driven primarily by the assumption that all 38.33 percentage points of between-market variation are due to monopsony. The additional contribution associated with the predictive content of static observables within markets is only 4.17–5.22 percentage points.

The distinction is even sharper for changes. Static market-position variables account for only 0.76–0.86% of total markdown variation through their within-market predictive content. The total ceiling is nevertheless 31.77–31.88% because the calculation mechanically attributes the entire 31.01% between-market component to monopsony. Using the static block's general-dominance contribution rather than its stand-alone $R^2$ produces similar conclusions: the within-market attribution is 5.12–5.70% for levels and only 1.15–1.57% for changes.

These exercises are predictive rather than causal. Firm market shares, workforce composition, and other measures of static market position may themselves respond to productivity shocks and dynamic employment decisions. The broad static block may therefore absorb variation that ultimately originates in firms' dynamic problem. The results nonetheless show that static market position has limited ability to account for within-market markdown dispersion and almost no ability to account for markdown movements. Dynamic firm states are substantially more informative in both dimensions, particularly for changes. This provides a reduced-form counterpart to the structural result that monopsony is important for the persistent level of the wage–MRPL gap, while dynamic employment considerations account for most of its cross-firm dispersion and movements over time.

# F Details on Misallocation and Markdown Accounting Exercises

This appendix gives the derivations and implementation details for the accounting exercises reported in Section 7. The exercises use the expected markdown and its structural decomposition,

$$\overline{\mu}_{jt} = \frac{\overline{W}_{jt}}{\overline{MRPL}_{jt}} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}, \tag{71}$$

where $\mu^{\varepsilon}_{jt}$ is the static monopsony component recovered from the labor-supply model and $\mu^{\phi}_{jt}$ is the dynamic cost wedge. Throughout, $L_{jt}$ denotes ability-adjusted labor input, $\overline{R}_{jt}$ is (expected) revenue, and $\alpha^{L}_{jt}$ is the labor revenue elasticity, so that $\overline{MRPL}_{jt}L_{jt} = \alpha^{L}_{jt}\overline{R}_{jt}$.

Note that these three exercises are partial-equilibrium accounting exercises. They do not solve for a new equilibrium allocation, do not model worker re-sorting, and do not impose that the dynamic cost wedge is necessarily inefficient. A cost wedge above one may reflect firms being constrained by socially costly adjustment frictions, but it may also reflect valuable labor hoarding or the continuation value of retaining workers. The exercises should therefore not be interpreted as welfare gains from eliminating dynamic costs.

## F.1 Local Harberger Revenue-Loss Index

The Harberger exercise converts dispersion in markdowns into a local revenue-loss index. Let

$$\tau_{jt} \equiv \frac{\overline{MRPL}_{jt}}{\overline{W}_{jt}} = \frac{1}{\overline{\mu}_{jt}}, \qquad z_{jt} \equiv \log \tau_{jt} = -\log \overline{\mu}_{jt}. \tag{72}$$

Using Equation (71),

$$z_{jt} = z^{\varepsilon}_{jt} + z^{\phi}_{jt}, \qquad z^{\varepsilon}_{jt} \equiv -\log \mu^{\varepsilon}_{jt}, \qquad z^{\phi}_{jt} \equiv -\log \mu^{\phi}_{jt}. \tag{73}$$

Consider a firm with local revenue function $\overline{R}_{jt}(L)$, holding productivity, capital, demand conditions, and other states fixed. The marginal revenue product is $\overline{MRPL}_{jt}(L) = \overline{R}'_{jt}(L)$. Define the local curvature of labor demand as

$$\kappa_{jt} \equiv -\frac{d\log \overline{MRPL}_{jt}}{d\log L_{jt}}. \tag{74}$$

A small change in labor therefore changes log marginal revenue product according to

$$d\log \overline{MRPL}_{jt} \simeq -\kappa_{jt}\, d\log L_{jt}. \tag{75}$$

Closing a local log markdown wedge $z_{jt}$ requires an approximate labor change

$$d\log L_{jt} \simeq \frac{z_{jt}}{\kappa_{jt}}, \qquad dL_{jt} \simeq L_{jt}\frac{z_{jt}}{\kappa_{jt}}. \tag{76}$$

For small wedges, $\overline{MRPL}_{jt} - \overline{W}_{jt} \simeq \overline{MRPL}_{jt}z_{jt}$. The Harberger triangle associated with moving labor locally toward the point at which $\overline{MRPL}_{jt} = \overline{W}_{jt}$ is therefore

$$
\begin{aligned}
Loss_{jt} &\simeq \frac{1}{2}(\overline{MRPL}_{jt} - \overline{W}_{jt})dL_{jt} && (77)\\
&\simeq \frac{1}{2}\overline{MRPL}_{jt} z_{jt}\left(L_{jt}\frac{z_{jt}}{\kappa_{jt}}\right) && (78)\\
&= \frac{1}{2}\frac{\overline{MRPL}_{jt}L_{jt}}{\kappa_{jt}} z_{jt}^2 = \frac{1}{2}\frac{\alpha_{jt}^L \overline{R}_{jt}}{\kappa_{jt}} z_{jt}^2. && (79)
\end{aligned}
$$

The preferred specification uses relative markdowns within labor markets. To isolate within-market dispersion, we remove the common market-year component using the same firm-level weights that enter the quadratic loss index. For a given curvature measure $\kappa_{jt}$, define the loss weight

$$
\varpi_{jt} \equiv \frac{\alpha_{jt}^L \overline{R}_{jt}}{\kappa_{jt}} = \frac{\overline{MRPL}_{jt}L_{jt}}{\kappa_{jt}}.
$$

Let $m(j,t)$ denote the market-year cell. We define the weighted market-year mean and the corresponding relative log wedge as

$$
\bar{z}_{m(j,t)} = \frac{\sum_{k\in m(j,t)} \varpi_{kt} z_{kt}}{\sum_{k\in m(j,t)} \varpi_{kt}}, \qquad \widetilde{z}_{jt} = z_{jt} - \bar{z}_{m(j,t)}. \tag{80}
$$

Thus, $\bar{z}_{m(j,t)}$ is the common market component that minimizes loss-weighted within-market dispersion. Because $\varpi_{jt}$ depends on $\kappa_{jt}$, the weighted market mean is recomputed separately for each curvature specification. The aggregate local revenue-loss index is then

$$
\mathcal{L} = \frac{1}{2}\sum_{jt} \varpi_{jt}\left(\widetilde{z}_{jt}\right)^2 = \frac{1}{2}\sum_{jt}\frac{\alpha_{jt}^L \overline{R}_{jt}}{\kappa_{jt}}\left(\widetilde{z}_{jt}\right)^2, \qquad \text{Loss} = 100 \times \frac{\mathcal{L}}{\sum_{jt}\overline{R}_{jt}}. \tag{81}
$$

We implement Equation (81) using three curvature measures. The Cobb–Douglas approximation is

$$
\kappa_{jt}^{CD} = 1 - \alpha_{jt}^L. \tag{82}
$$

The actual fixed-input curvature is

$$
\kappa_{jt}^{ACT} = -\frac{\partial \log \overline{MRPL}_{jt}}{\partial \log L_{jt}}, \tag{83}
$$

computed from the estimated revenue production function holding materials fixed. The flexible-materials curvature allows materials to adjust locally and is

$$
\kappa_{jt}^{FLEX} = -\left[\frac{\partial \log \overline{MRPL}_{jt}}{\partial \log L_{jt}} - \frac{\partial \log \overline{MRPL}_{jt}}{\partial \log M_{jt}}\frac{\partial \log MRPM_{jt}/\partial \log L_{jt}}{\partial \log MRPM_{jt}/\partial \log M_{jt}}\right]. \tag{84}
$$

The decomposition follows directly from Equation (73). Since

$$
\widetilde{z}_{jt} = \widetilde{z}_{jt}^{\varepsilon} + \widetilde{z}_{jt}^{\phi}, \tag{85}
$$

we have

$$
\left(\widetilde{z}_{jt}\right)^2 = \left(\widetilde{z}_{jt}^{\varepsilon}\right)^2 + \left(\widetilde{z}_{jt}^{\phi}\right)^2 + 2\widetilde{z}_{jt}^{\varepsilon}\widetilde{z}_{jt}^{\phi}. \tag{86}
$$

Thus total loss is the sum of a monopsony term, a dynamic cost term, and a covariance term:

$$\mathcal{L}^{\varepsilon} = \frac{1}{2}\sum_{jt} \frac{\alpha_{jt}^{L}\overline{R}_{jt}}{\kappa_{jt}} \left(\widetilde{z}_{jt}^{\varepsilon}\right)^2, \tag{87}$$

$$\mathcal{L}^{\phi} = \frac{1}{2}\sum_{jt} \frac{\alpha_{jt}^{L}\overline{R}_{jt}}{\kappa_{jt}} \left(\widetilde{z}_{jt}^{\phi}\right)^2, \tag{88}$$

$$\mathcal{L}^{cov} = \sum_{jt} \frac{\alpha_{jt}^{L}\overline{R}_{jt}}{\kappa_{jt}} \widetilde{z}_{jt}^{\varepsilon}\widetilde{z}_{jt}^{\phi}. \tag{89}$$

For the main tables, we allocate one-half of the covariance term to each component:

$$Share^{\varepsilon} = \frac{\mathcal{L}^{\varepsilon} + \frac{1}{2}\mathcal{L}^{cov}}{\mathcal{L}}, \qquad Share^{\phi} = \frac{\mathcal{L}^{\phi} + \frac{1}{2}\mathcal{L}^{cov}}{\mathcal{L}}. \tag{90}$$

These two shares sum to one. Table 8 reports the market-demeaned results for the three curvature measures.

## F.2 Aggregate Wage–MRPL Gap Accounting

The wage-gap exercise is an exact accounting identity. Starting from

$$\overline{W}_{jt} = \overline{\mu}_{jt}\overline{MRPL}_{jt} = \mu_{jt}^{\varepsilon}\mu_{jt}^{\phi}\overline{MRPL}_{jt}, \tag{91}$$

we can write the firm-level gap as

$$\begin{aligned} \overline{MRPL}_{jt} - \overline{W}_{jt} &= \overline{MRPL}_{jt}\left(1 - \mu_{jt}^{\varepsilon}\mu_{jt}^{\phi}\right) & (92)\\ &= \overline{MRPL}_{jt}\left(1 - \mu_{jt}^{\varepsilon}\right) - \overline{MRPL}_{jt}\mu_{jt}^{\varepsilon}\left(\mu_{jt}^{\phi} - 1\right). & (93) \end{aligned}$$

Multiplying by ability-adjusted labor input gives

$$L_{jt}(\overline{MRPL}_{jt} - \overline{W}_{jt}) = \underbrace{L_{jt}\overline{MRPL}_{jt}\left(1 - \mu_{jt}^{\varepsilon}\right)}_{\text{static monopsony component}} + \underbrace{L_{jt}\overline{MRPL}_{jt}\left[-\mu_{jt}^{\varepsilon}\left(\mu_{jt}^{\phi} - 1\right)\right]}_{\text{dynamic cost component}}. \tag{94}$$

Aggregating Equation (94) gives the entries in Table 9. We report the observed gap under two normalizations. The first is the potential wage bill,

$$PWB \equiv \sum_{jt} L_{jt}\overline{MRPL}_{jt}, \tag{95}$$

which is the wage bill that would obtain if each efficiency unit of labor were paid its marginal revenue product. The second is the actual wage bill,

$$WB \equiv \sum_{jt} L_{jt}\overline{W}_{jt}. \tag{96}$$

The component shares are computed relative to the observed aggregate gap. A negative dynamic component means that $\mu_{jt}^{\phi} > 1$ is quantitatively important, so dynamic continuation values or labor-hoarding motives partially offset the static monopsony wage gap.

## F.3 Firm Targeting Exercise

The targeting exercise asks whether the observed wage–MRPL wedge or markdown is a reliable screen for the structural source of the wedge. We define three target sets. The full-markdown target selects firm-years with the widest observed wage–MRPL gaps, equivalently the lowest values of $\overline{\mu}_{jt}$. The monopsony target selects firm-years with the widest static monopsony component, equivalently the lowest values of $\mu^{\varepsilon}_{jt}$. The dynamic-cost target selects firm-years with the largest values of $|\log \mu^{\phi}_{jt}|$.

We construct these targets in two ways. The economy-wide version uses unconditional deciles in the pooled sample. The within-market version ranks firms within cells defined by four-digit industry, municipality, and year, and selects the top decile within each cell. The main statistic is the conditional overlap

$$Overlap(r|full) = \Pr\left(T^{r}_{jt} = 1 \mid T^{full}_{jt} = 1\right), \qquad r \in \{\varepsilon, \phi\}, \tag{97}$$

where $T^{full}_{jt}$ is the full-markdown target and $T^{r}_{jt}$ is either the monopsony or dynamic-cost target. We also compute weighted overlaps,

$$Overlap_a(r|full) = \frac{\sum_{jt} a_{jt} T^{full}_{jt} T^{r}_{jt}}{\sum_{jt} a_{jt} T^{full}_{jt}}, \tag{98}$$

using the potential wage bill weight $a_{jt} = L_{jt}\overline{MRPL}_{jt}$.

This exercise is not intended to characterize an optimal policy rule. It asks a simpler diagnostic question: if a researcher or policymaker used the observed wage–MRPL wedge to identify firms with labor-market power or dynamic employment frictions, how often would that screen select the firms identified by the structural decomposition? The results in Table 10 show that the answer depends strongly on whether the target is monopsony power or the dynamic cost wedge, and on whether the ranking is economy-wide or within market.

# G Additional Results

## G.1 Tables

Table G.1: WORKER CHARACTERISTICS

| Statistic | Earnings | Annual wages | Hourly wage | Age |
|---|---|---|---|---|
| Mean | 51 | 56 | 38 | 38 |
| Std. Dev. | 37 | 117 | 162 | 13 |
| p10 | 6 | 10 | 17 | 20 |
| p25 | 25 | 38 | 27 | 28 |
| p50 | 51 | 54 | 35 | 38 |
| p75 | 67 | 69 | 44 | 48 |
| p90 | 88 | 91 | 58 | 57 |
| p99 | 165 | 177 | 149 | 66 |
| Observations | | 8,800,000 | | |

*Notes:* Table G.1 reports cross-sectional moments for workers in the sample. For comparison, annual earnings and annual wages are measured in thousands of 2010 USD using the Danish CPI and average DKK-USD exchange rate in 2010. Hourly wages are measured in 2010 USD, and age is measured in years. The number of observations is rounded to the nearest thousand.

Table G.2: FIRM CHARACTERISTICS

| Statistic | Employment | Revenue | Value added | Value added/worker | Firm age |
|---|---|---|---|---|---|
| Mean | 24 | 6,191 | 1,993 | 86 | 16 |
| Std. Dev. | 202 | 57,549 | 14,946 | 106 | 11 |
| p10 | 2 | 289 | 114 | 33 | 5 |
| p25 | 3 | 494 | 197 | 51 | 7 |
| p50 | 6 | 1,024 | 404 | 74 | 14 |
| p75 | 14 | 2,722 | 1,018 | 104 | 22 |
| p90 | 36 | 8,440 | 2,908 | 147 | 30 |
| p99 | 281 | 83,647 | 26,531 | 298 | 50 |
| Observations | | | 374,000 | | |

*Notes:* Table G.2 reports cross-sectional moments for firms in the estimation sample. For comparison, revenue, value added, and value added per worker are measured in thousands of 2010 USD using the Danish CPI and average DKK-USD exchange rate in 2010. The number of observations is rounded to the nearest thousand.

Table G.3: RETURNS TO SCALE AND OUTPUT ELASTICITIES: ABILITY-ADJUSTED VS. NON-ABILITY-ADJUSTED

| Statistic | $RTS_{jt}$ | $\varepsilon^Y_{K,jt}$ | $\varepsilon^Y_{L,jt}$ | $\varepsilon^Y_{M,jt}$ |
|---|---|---|---|---|
| *Panel A. Ability-adjusted (baseline)* | | | | |
| Mean | 0.95 | 0.05 | 0.35 | 0.54 |
| Std. Dev. | 0.05 | 0.02 | 0.14 | 0.17 |
| p10 | 0.88 | 0.03 | 0.16 | 0.33 |
| p50 | 0.95 | 0.05 | 0.36 | 0.54 |
| p90 | 1.01 | 0.07 | 0.53 | 0.77 |
| *Panel B. Non-ability-adjusted* | | | | |
| Mean | 0.99 | 0.06 | 0.40 | 0.53 |
| Std. Dev. | 0.08 | 0.13 | 0.19 | 0.16 |
| p10 | 0.89 | −0.10 | 0.16 | 0.33 |
| p50 | 0.99 | 0.06 | 0.39 | 0.52 |
| p90 | 1.10 | 0.23 | 0.65 | 0.74 |

*Notes:* Table G.3 reports cross-sectional moments of returns to scale and the revenue elasticities of the estimated revenue production function under the baseline ability-adjusted specification (Panel A) and the non-ability-adjusted specification (Panel B). Returns to scale are defined as $RTS_{jt} = \varepsilon^Y_{K,jt} + \varepsilon^Y_{L,jt} + \varepsilon^Y_{M,jt}$. In Panel A, labor input is measured using ability-adjusted hours, $L_{jt} = \sum_i A_{it} H_{it}$; in Panel B, the production function is estimated using total hours of employment as the labor input. The sample includes 374,000 firm-year observations, rounded to the nearest thousand. Percentiles are calculated as the mean of adjacent milliles.

Table G.4: MODEL ESTIMATES BY INDUSTRY

| | Wages and markdowns | | | Output elasticities | | | |
|---|---|---|---|---|---|---|---|
| Industry | $w_{jt}$ | $mrpl_{jt}$ | $\mu_{jt}$ | $RTS_{jt}$ | $\varepsilon^Y_{K,jt}$ | $\varepsilon^Y_{L,jt}$ | $\varepsilon^Y_{M,jt}$ |
| Mining and quarrying | 6.72 | 7.29 | 0.61 | 0.96 | 0.05 | 0.33 | 0.59 |
| Manufacturing | 6.63 | 6.88 | 0.80 | 0.97 | 0.05 | 0.38 | 0.53 |
| Construction | 6.60 | 6.81 | 0.84 | 0.94 | 0.05 | 0.38 | 0.51 |
| Wholesale and retail | 6.56 | 6.79 | 0.84 | 0.96 | 0.04 | 0.25 | 0.67 |
| Transportation and storage | 6.63 | 6.91 | 0.79 | 0.93 | 0.06 | 0.41 | 0.46 |
| Accommodation and food services | 6.49 | 6.84 | 0.74 | 0.92 | 0.06 | 0.35 | 0.52 |
| Information and communication | 6.69 | 6.80 | 0.94 | 0.96 | 0.06 | 0.45 | 0.45 |
| Finance and insurance | 6.57 | 6.77 | 0.87 | 0.96 | 0.06 | 0.43 | 0.47 |
| Real estate | 6.58 | 6.92 | 0.77 | 0.91 | 0.07 | 0.40 | 0.45 |
| Professional and technical services | 6.62 | 6.82 | 0.85 | 0.93 | 0.06 | 0.48 | 0.39 |
| Administrative and support services | 6.62 | 6.85 | 0.84 | 0.93 | 0.06 | 0.47 | 0.40 |
| Other services | 6.56 | 6.86 | 0.77 | 0.93 | 0.05 | 0.36 | 0.52 |

*Notes:* Table G.4 reports firm-year means of the baseline ability-adjusted model estimates by broad NACE Rev. 2 industry. The firm-level wage $w_{jt}$ and MRPL $mrpl_{jt}$ are reported in log real DKK, while the markdown $\mu_{jt}$ is reported in levels. Returns to scale are defined as $RTS_{jt} = \varepsilon^Y_{K,jt} + \varepsilon^Y_{L,jt} + \varepsilon^Y_{M,jt}$. The revenue elasticities are those of the estimated revenue production function. The underlying estimation sample includes 374,000 firm-year observations, rounded to the nearest thousand. Components may not sum exactly to the reported returns to scale because entries are rounded.

## Table G.5: Nested-Logit Labor Supply: Parameter Estimates

| | Linear wage index (1) | Quadratic wage index (2) |
|---|---|---|
| *Panel A. Wage index* | | |
| $\Delta_4 w_{jt}$ | 1.371 | $-2.143$ |
| | (0.408) | (2.063) |
| $\Delta_4 w_{jt}^2$ | — | 0.322 |
| | | (0.158) |
| *Panel B. Predetermined firm and workforce characteristics* | | |
| Lagged mean worker experience | 0.099 | 0.106 |
| | (0.006) | (0.010) |
| Lagged mean tenure | $-0.211$ | $-0.223$ |
| | (0.007) | (0.012) |
| Lagged mean worker age | $-0.050$ | $-0.065$ |
| | (0.017) | (0.027) |
| Lagged college share | 0.051 | $-0.002$ |
| | (0.017) | (0.036) |
| Lagged senior-position share | 0.014 | $-0.023$ |
| | (0.014) | (0.029) |
| Lagged white-collar share | 0.009 | 0.019 |
| | (0.008) | (0.013) |
| Lagged stayer share | 0.070 | 0.076 |
| | (0.007) | (0.011) |
| Firm age | 0.386 | 0.393 |
| | (0.007) | (0.012) |
| Log capital | 0.093 | 0.098 |
| | (0.006) | (0.009) |

*Panel C. Nesting coefficients,* $1-\sigma_{cz}^{-1}$

| | Linear | Quadratic | | Linear | Quadratic |
|---|---|---|---|---|---|
| CZ 1 | 0.214 | 0.219 | CZ 11 | 0.349 | 0.365 |
| | (0.009) | (0.015) | | (0.023) | (0.039) |
| CZ 2 | 0.299 | 0.305 | CZ 12 | 0.344 | 0.343 |
| | (0.022) | (0.036) | | (0.021) | (0.034) |
| CZ 3 | 0.294 | 0.296 | CZ 13 | 0.441 | 0.447 |
| | (0.021) | (0.034) | | (0.030) | (0.048) |
| CZ 4 | 0.499 | 0.508 | CZ 14 | 0.300 | 0.305 |
| | (0.045) | (0.073) | | (0.022) | (0.036) |
| CZ 6 | 0.266 | 0.267 | CZ 15 | 0.519 | 0.525 |
| | (0.054) | (0.088) | | (0.047) | (0.078) |
| CZ 7 | 0.256 | 0.258 | CZ 16 | 0.303 | 0.308 |
| | (0.015) | (0.024) | | (0.015) | (0.024) |
| CZ 8 | 0.355 | 0.354 | CZ 17 | 0.353 | 0.358 |
| | (0.046) | (0.075) | | (0.024) | (0.040) |
| CZ 9 | 0.357 | 0.371 | CZ 19 | 0.352 | 0.359 |
| | (0.032) | (0.053) | | (0.021) | (0.033) |
| CZ 10 | 0.320 | 0.326 | CZ 20 | 0.417 | 0.424 |
| | (0.044) | (0.073) | | (0.032) | (0.052) |

| Linear wage index | | Quadratic wage index | |
|---|---|---|---|
| Mean implied $\sigma_{cz}$ | 1.555 | Mean implied $\sigma_{cz}$ | 1.570 |
| Range of implied $\sigma_{cz}$ | 1.272–2.079 | Range of implied $\sigma_{cz}$ | 1.280–2.105 |
| Year effects | Yes | Year effects | Yes |
| Industry effects | Yes | Industry effects | Yes |
| Observations | 222,000 | Observations | 222,000 |

Notes: The table reports estimates of the nested-logit labor-supply equation described in Section 6.1. The dependent variable is the four-year change in the log firm share relative to the outside option. Shares are measured in units of ability-adjusted labor. Let $w_{jt} \equiv \log \overline{W}_{jt}$ denote the log expected wage per efficiency unit. Column (1) specifies a linear wage index, while column (2) allows the wage index to be quadratic. Firm and workforce characteristics enter in four-year changes.

Panel C reports the commuting-zone-specific coefficients on the four-year change in log within-market share. Under the nested-logit model, these coefficients equal $1-\sigma_{cz}^{-1}$, so that the implied nesting parameter is $\sigma_{cz} = [1-(1-\sigma_{cz}^{-1})]^{-1}$. The mean and range of the implied $\sigma_{cz}$ are calculated across the commuting zones reported in the estimation. Standard errors are in parentheses. Wage and within-market shares are instrumented using the firm-side productivity shocks and rival-firm characteristics described in Section 6.1. Commuting zones 5 and 18 are excluded due to insufficient observations. Observation counts are rounded to the nearest thousand.

Table G.6: LABOR SUPPLY ELASTICITIES AND IMPLIED MONOPSONY WEDGES

| | Linear wage index | | Quadratic wage index | |
|---|---|---|---|---|
| Statistic | $\epsilon^L_{Wjt}$ | $\mu^{\epsilon}_{jt}$ | $\epsilon^L_{Wjt}$ | $\mu^{\epsilon}_{jt}$ |
| Mean | 2.662 | 0.724 | 2.869 | 0.738 |
| Std. Dev. | 0.395 | 0.029 | 0.470 | 0.031 |
| p1 | 1.971 | 0.663 | 1.943 | 0.660 |
| p10 | 2.099 | 0.677 | 2.266 | 0.694 |
| p50 | 2.639 | 0.725 | 2.840 | 0.740 |
| p90 | 3.041 | 0.753 | 3.446 | 0.775 |
| p99 | 3.889 | 0.796 | 4.242 | 0.809 |
| Observations | 374,000 | | 374,000 | |

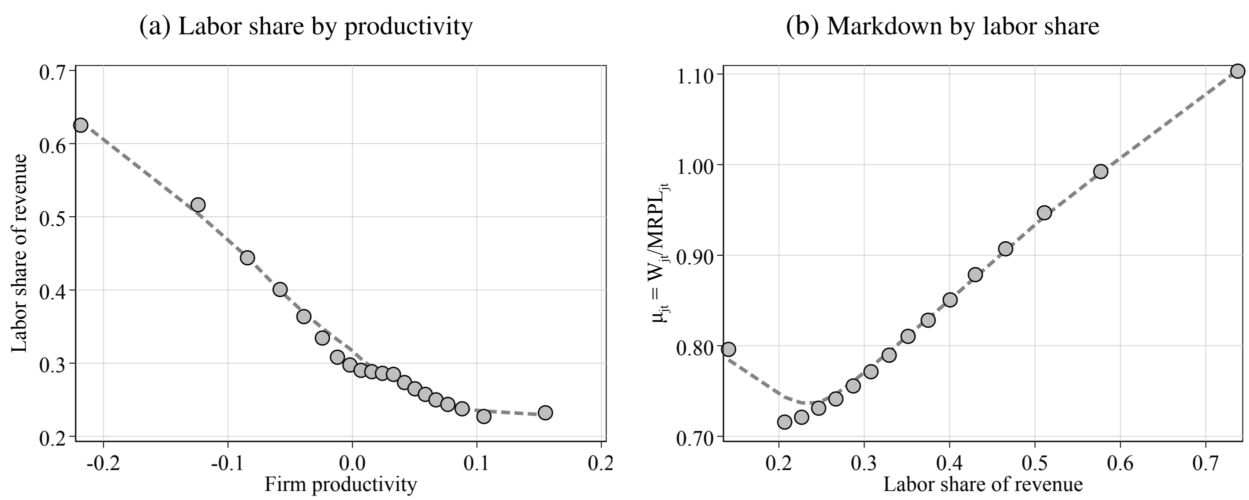
Notes: The table reports moments of the firm-level labor-supply elasticity and the implied static monopsony wedge under the linear and quadratic wage-index specifications. For the quadratic specification, the wage-index slope is $\partial f^u(W_{jt})/\partial \log W_{jt} = \gamma_1 + 2\gamma_2 \log \bar{W}_{jt}$; for the linear specification, $\gamma_2 = 0$. Firm-level elasticities are constructed as

$$\epsilon^L_{Wjt} = \frac{\partial f^u(W_{jt})}{\partial \log W_{jt}} \left[\sigma_{cz} + (1-\sigma_{cz})s_{jgt} - s_{jt}\right],$$

and the implied monopsony wedge is $\mu^{\epsilon}_{jt} = \epsilon^L_{Wjt}/(1+\epsilon^L_{Wjt})$. The quadratic wage-index specification is the baseline specification used in the main text. Distributional moments are evaluated for the full firm-year estimation sample using the parameters estimated from the four-year-difference regressions in Table G.5. Observation counts are rounded to the nearest thousand. Percentiles are calculated as the mean of adjacent milliles.

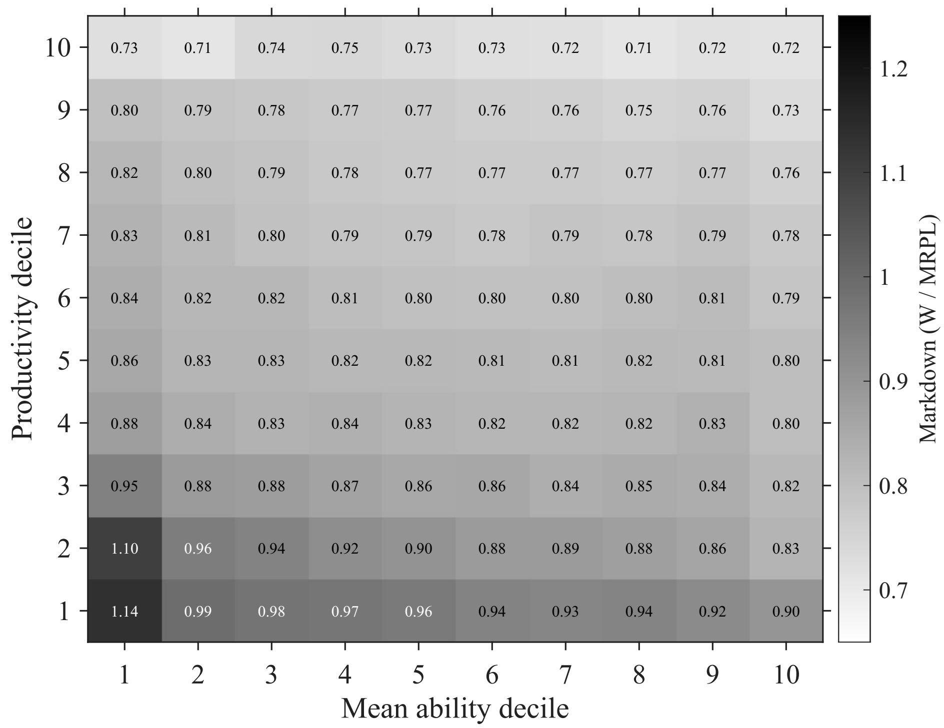
## G.2 Figures

Figure G.1: MARKDOWNS AND LABOR SHARES



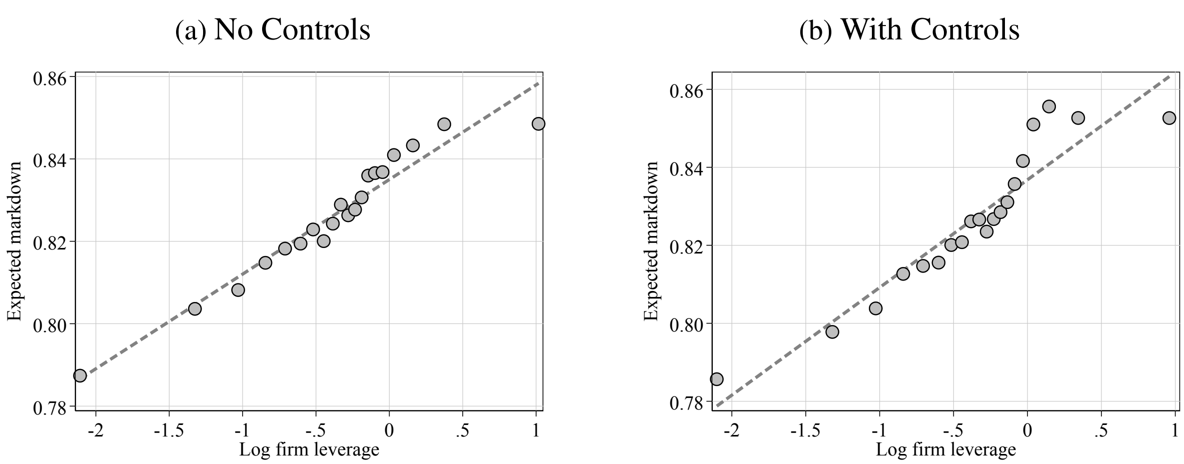

*Notes:* Panel (a) shows a binscatter plot of the firm-level labor share of revenue by firm productivity ($\omega_{jt}$). Panel (b) shows a binscatter plot of the firm-level markdown ($\mu_{jt} = W_{jt}/\text{MRPL}_{jt}$) by the labor share of revenue. These results are based on a sample of approximately 374,000 firm-year observations.

Figure G.2: Markdown Heatmap by Productivity and Mean Worker Ability



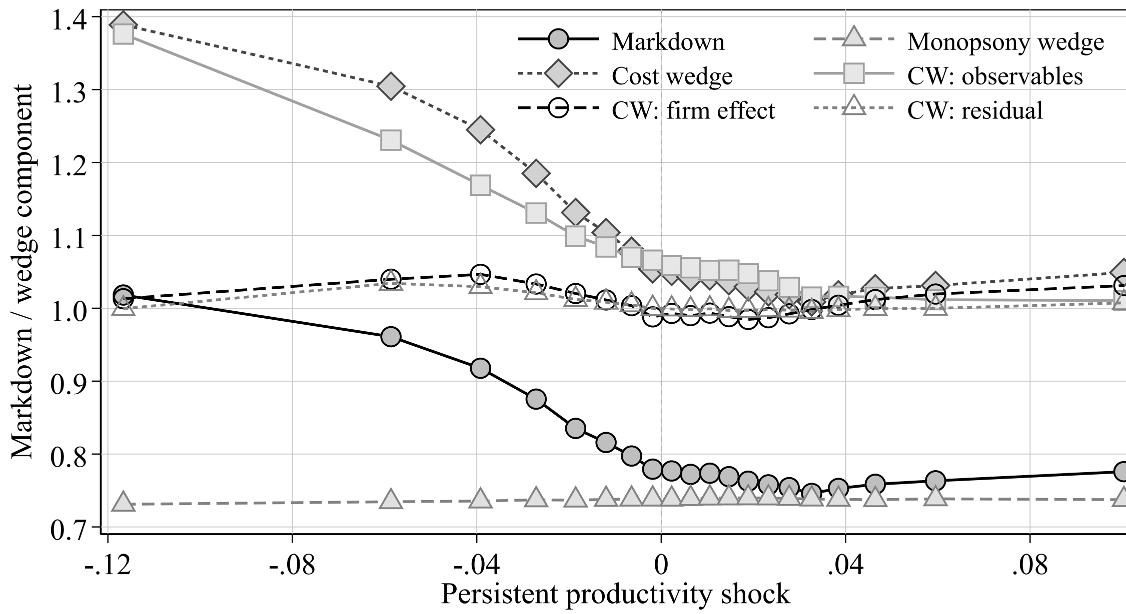

*Notes:* Figure G.2 shows the average markdown $\mu_{jt} = W_{jt}/\text{MRPL}_{jt}$ across bins of firm productivity and mean worker ability. Lighter cells indicate lower markdowns where workers are paid less of their marginal product. Cell values report the bin-level average markdown.

Figure G.3: Markdown by Leverage



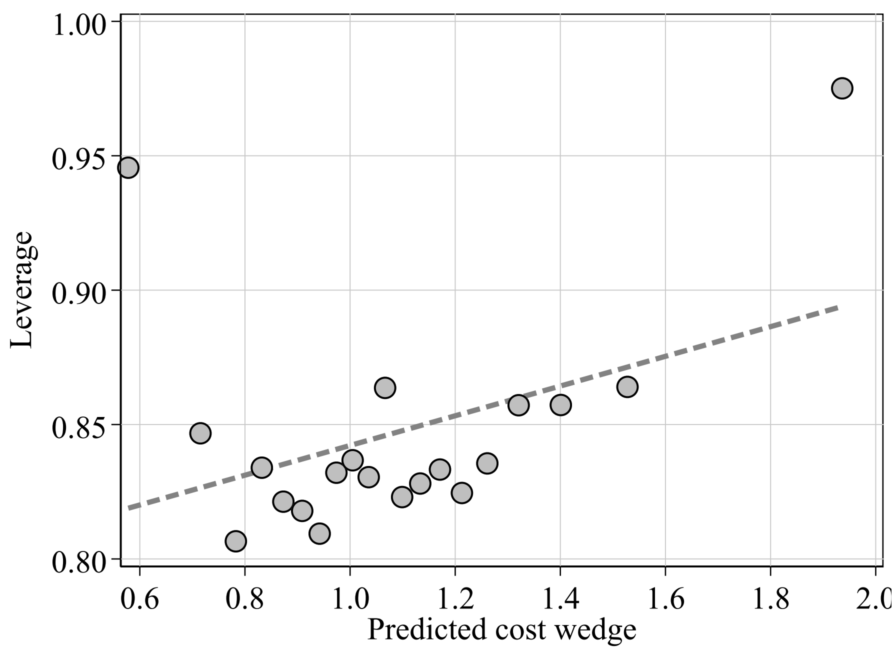

*Notes:* Figure G.3 shows binscatter plots of markdowns on firm leverage. The x-axis in both panels represents log firm leverage, and the y-axis represents the average within-bin expected markdown ($\overline{\mu}_{jt}$). Panel (a) shows the raw relationship, while Panel (b) controls for lagged log employment, lagged log revenue, and year and industry effects.

Figure G.4: TFP SHOCKS AND COST WEDGE DECOMPOSITION



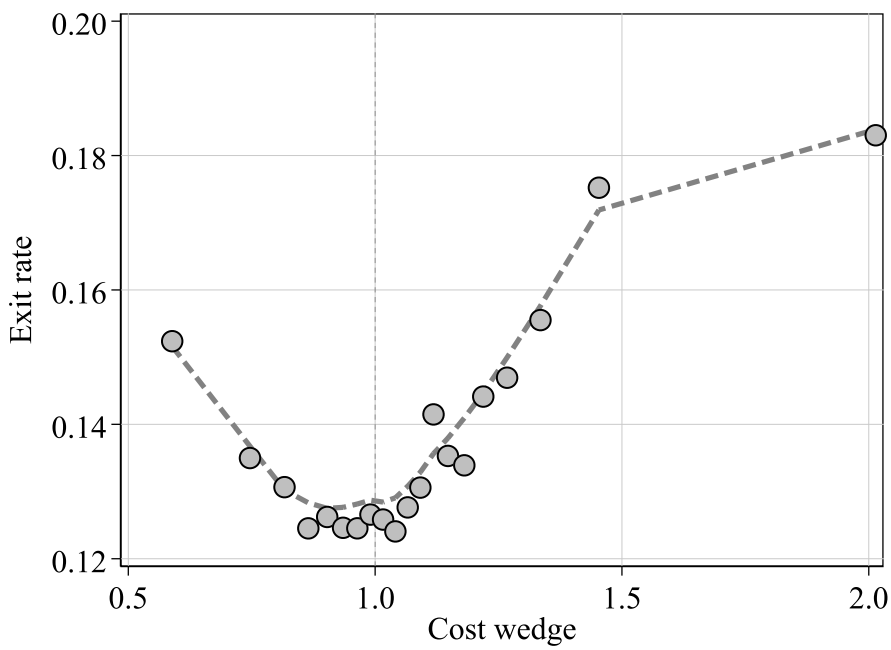

*Notes:* Figure G.4 plots means of the markdown and its structural components across bins of the persistent productivity shock, $\eta_{jt}$. The markdown is $\overline{\mu}_{jt} = \overline{W}_{jt}/\overline{MRPL}_{jt}$. CW denotes the components in the cost wedge. The monopsony wedge, $\mu^{\varepsilon}_{jt}$, is implied by the estimated firm-level labor supply elasticity, and the cost wedge, $\mu^{\phi}_{jt}$, is recovered from $\mu_{jt} = \mu^{\varepsilon}_{jt}\mu^{\phi}_{jt}$. The remaining series decompose the estimated cost wedge into the component explained by observed firm states, the firm fixed-effect component, and the residual component. The vertical dashed line marks $\eta_{jt} = 0$.

Figure G.5: CORRELATION BETWEEN LEVERAGE AND PREDICTED COST WEDGE




*Notes:* Figure G.5 plots bin means of firm leverage against the predicted cost wedge, $\widehat{\mu}^{\phi}_{jt}$. The predicted cost wedge is the fitted component from the estimated marginal-cost specification, combining observed firm states and persistent firm heterogeneity. Leverage is measured as debt divided by assets.

Figure G.6: EXIT RATES BY COST WEDGE




*Notes:* Figure G.6 plots exit rates by bins of the cost wedge distribution, controlling for firm productivity, employment, capital, industry, and year. See Section 6.2 for further discussion.