The Effort Margin of Monopsony*

Ignacio Amaral†
Central Bank of Uruguay

Xincheng Qiu‡
Peking University

August 27, 2026

Abstract

This paper revisits the theory and measurement of monopsony when worker effort responds to wages. We derive a markdown formula incorporating the effort margin and show that ignoring it overstates labor market power. Using estimates of the effort-wage elasticity, we find that implied wage-to-marginal-product gaps in the US fall from roughly 9–48 percent to 3–13 percent. The production approach, however, remains valid with endogenous effort. In a shirking model with imperfect monitoring, stronger monitoring widens markdowns under standard preferences. Consistent with this prediction, firms with more intensive monitoring set wider markdowns in the data.

Keywords: monopsony, efficiency wages, markdowns, monitoring, management
JEL Codes: J42, J31, J41, M54

*We thank Yongsung Chang, Jan Eeckhout, Sergio Salgado, and seminar participants at Seoul National University.

†Email: iamaral@bcu.gub.uy.
‡Email: xincheng.qiu@gsm.pku.edu.cn.

1 Introduction

How much power do firms have to set wages below the marginal revenue product of labor? Modern empirical work infers this power from responsiveness of firm employment to wages: firms can sustain wider markdowns when employment responds weakly. Estimates based on this logic point to substantial and pervasive labor market power in the United States (Manning, 2011; Webber, 2015; Bassier et al., 2022; Berger et al., 2022). These findings have shaped both academic debates on wage determination and policy discussions on minimum wages and labor-market antitrust.

This paper argues that the standard treatment of monopsony omits an important margin of worker response. Workers respond to wages not only by deciding whether to work for a firm, but also by adjusting how much effort they supply once employed. A long efficiency-wage literature provides both theoretical and empirical grounding for such a response: higher wages reduce shirking, improve morale, and induce reciprocity. If a wage cut causes some workers to leave and those who stay to work less hard, then the relevant input for the firm is effective labor, rather than headcount alone. Measuring monopsony power from headcount alone therefore attributes too much wage-setting power to the firm.

We formalize this idea in a simple monopsony model in which both employment and effort respond to the wage. The optimal markdown depends on two elasticities: the wage elasticity of employment, ε_n, and the wage elasticity of effort, ε_e. The standard monopsony formula is the special case $\varepsilon_e = 0$. When $\varepsilon_e > 0$, however, the labor supply elasticity captures only part of the response of effective labor to wages, and the standard formula overstates the size of the markdown. When the headcount margin becomes arbitrarily inelastic, the firm extracts essentially all worker surplus in the standard model. With an effort margin, however, workers who cannot readily leave can still respond to lower wages by supplying less effort, which places an upper bound on the firm's wage-setting power.

Our first contribution is to show that existing markdown estimates can be corrected using only the effort-wage elasticity. Using the mean effort-wage elasticity of 0.72 across the studies surveyed by Fongoni (2024), we find that the exploitation rates implied by existing U.S. estimates falls from 0.09–0.48 to approximately 0.03–0.13. Even conservative effort elasticities imply economically meaningful reductions. Accounting for the effort margin therefore moves measured labor market power considerably closer to the competitive benchmark.

Our second contribution concerns the production approach to measuring markdowns. Brooks et al. (2021) and Yeh et al. (2022), adapting the production-based measurement of markup (De Loecker and Warzynski, 2012; De Loecker et al., 2020), infer markdowns from output elasticities and factor expenditure shares. We show that the identifying relationship underlying this approach is robust to endogenous effort. The empirical implementation, however, becomes more demanding. Usual instruments that address the standard endogeneity of labor with respect to productivity may still be correlated with unobserved effort. In a tractable benchmark, we show that this bias causes the production approach to recover the same misspecified markdown as the elasticity approach. Recovering the true markdown instead requires either controlling for the determinants of effort or isolating variation in employment that is orthogonal to effort.

We also provide a simpler strategy for applications interested in how markdowns vary across firms rather than in their levels. The production approach implies that the log labor-to-materials expenditure ratio equals the log markdown up to a term involving the corresponding output elasticities. Under the standard assumption that these elasticities are common across firms within sufficiently narrow industries, that term is absorbed by industry fixed effects. The labor-to-materials expenditure ratio can therefore be used to study cross-firm differences in markdowns without estimating the output elasticity of labor and without directly confronting the omitted-effort problem.

Our third contribution is to show that the effort margin creates a new link between management practices and labor market power. We model effort using a Shapiro and Stiglitz (1984) monitoring framework in which workers can shirk and firms detect shirking only imperfectly. More intensive monitoring allows the firm to elicit greater effort at a given wage, but it also changes the wage the firm optimally chooses. The effect on markdowns therefore depends on how these two forces interact. We derive an exact condition, governed by the curvature of utility and effort costs, under which monitoring widens the markdown. Under standard CRRA preferences and isoelastic convex effort costs, this condition is always satisfied, so more intensive monitoring unambiguously increases labor market power.

We bring this prediction to the management-practice data of Bloom and Van Reenen (2007). Using the labor-to-materials expenditure ratio to capture within-industry variation in markdowns, we find that firms with more intensive performance monitoring pay workers a smaller share of their marginal revenue product. A one-standard-deviation increase in monitoring is associated with a decline of roughly 0.15–0.19 log points in markdown. This evidence points to a role for management practices beyond their well-studied effects on productivity. By shaping workers' effort and firms' wage-setting incentives, management practices can also influence how the surplus is divided between workers and employers.

Related literature. First, this paper contributes to the growing literature measuring labor market power. A large strand of this literature infers monopsony power from the elasticity of labor supply to the firm (Staiger et al., 2010; Falch, 2010; Ransom and Sims, 2010; Webber, 2015; Bassier et al., 2022); Sokolova and Sorensen (2021) provide a meta-analysis of these estimates. Related work embeds finite firm-specific labor supply elasticities in richer structural models to quantify wage-setting power and welfare (Berger et al., 2022; Lamadon et al., 2022). A complementary approach measures markdowns using production data. Brooks et al. (2021) and Yeh et al. (2022), adapting production-based methods for measuring markups, recover the markdown from output elasticities and factor expenditure shares. We show how the effort margin changes the implementation of both approaches to measuring monopsony power.

Second, the paper builds on the efficiency-wage literature, which has long emphasized that worker effort responds to compensation. Classic theories generate this relationship through discipline and the threat of job loss (Shapiro and Stiglitz, 1984), gift exchange (Akerlof, 1982), and concerns about fair wages (Akerlof and Yellen, 1990); see Katz (1986) for an early survey. A related macroeconomic literature incorporates endogenous effort to study aggregate fluctuations (Danthine and Donaldson, 1990; Burnside et al., 1993; Basu and Kimball, 1997; Bils and Chang, 2003; Bils et al., 2022; Galí and Van Rens, 2021). Whereas this literature has primarily studied how the effort margin affects unemployment, productivity, and business cycles, our paper brings the insight to the study of labor market power.

Third, we relate to the literature on management practices and the organization of work. A series of papers by Bloom, Van Reenen, and coauthors develops systematic survey measures of management practices and documents that these practices vary widely across firms and countries and are strongly associated with productivity and firm performance (Bloom and Van Reenen, 2007, 2010; Bloom et al., 2012, 2014, 2019). A complementary body of work establishes causal evidence that better management raises productivity (Bloom et al., 2013; Bruhn et al., 2018; Giorcelli, 2019). Whereas this literature has focused primarily on how management practices affect productivity, we study how monitoring shapes workers' effort incentives and, through them, the firm's wage-setting power.

Roadmap. Section 2 reviews the mechanisms through which wages affect worker effort, develops the baseline model, and derives the markdown formula. Section 3 studies the implications of the effort margin for the elasticity and production approaches to measuring markdowns and quantifies the resulting correction to existing estimates. Section 4 microfounds effort function through an explicit monitoring technology, characterizes how monitoring affects wages and markdowns, and examines this relationship using management-practice data. Section 5 concludes. Appendix IV provides the proofs.

2 Theory

Our baseline analysis takes the effort function $e(w)$ as a primitive. Efficiency-wage theories imply that wages raise the effort of employed workers through discipline, gift exchange, and fair-wage concerns. In the shirking model (Shapiro and Stiglitz, 1984), higher wages increase the loss from dismissal and so makes effort more attractive. In sociological models that emphasize reciprocity (Akerlof, 1982; Akerlof and Yellen, 1990) and effort withheld when pay falls below a reference wage. All imply that effort increases in wage. We ask how this response affects a firm choosing wages along an upward-sloping labor supply curve. The main results are agnostic about the microfoundation. Section 4 adopts the shirking model to derive comparative statics for monitoring intensity.

2.1 Baseline Model

We set up a baseline framework to study labor market power where worker effort responds to the wage. A single firm employs a continuum of workers in a static, partial-equilibrium setting. The firm faces an upward-sloping labor supply curve $n(w)$ and an effort function $e(w)$, both increasing in the wage, so total effective labor is $n(w) \cdot e(w)$. A revenue function $R(\cdot)$ maps effective labor into output, and the firm chooses the wage to maximize profit:

$$\max_{w} \; R\big(n(w)e(w)\big) - wn(w).$$

We assume $R(\cdot)$ is increasing and strictly concave, that $n(\cdot)$ and $e(\cdot)$ are positive, increasing, and continuously differentiable, and that profit is strictly quasiconcave in w with an interior maximizer w^*. These conditions guarantee that the first-order condition characterizes the optimum.

Define the markdown ν as the ratio of the effective wage to the marginal revenue product of effective labor:

$$\nu \equiv \frac{w/e(w)}{R'\Big(n(w)e(w)\Big)}.$$

This coincides with the standard definition of the markdown as the wage divided by the marginal revenue product of a worker, $w/(\mathrm{d}R/\,\mathrm{d}n)$.[1] Values of ν below one are wage markdowns, with lower ν a wider markdown.

Labor supply responds to the wage along two margins: a *headcount margin*, the number of workers willing to work for the firm, and an *effort margin*, the level of effort each supplies. Write the wage elasticity of headcount as $\varepsilon_n = \frac{wn'(w)}{n(w)}$ and the wage elasticity of effort as $\varepsilon_e = \frac{we'(w)}{e(w)}$.

[1]Since revenue depends on total effective labor ne, $\mathrm{d}R/\,\mathrm{d}n = R'(ne) \cdot e$, so $w/(\mathrm{d}R/\,\mathrm{d}n) = w/[R'(ne) \cdot e] = \nu$.

Proposition 1. *The optimal markdown satisfies*

$$\nu = \frac{\varepsilon_n + \varepsilon_e}{1 + \varepsilon_n},$$

where all elasticities are evaluated at w^*.

The formula admits a transparent reading. Let $\ell \equiv n(w)e(w)$ denote the *effective* labor supply, and $\varepsilon_\ell = \varepsilon_n + \varepsilon_e$ the elasticity of effective labor supply. Then

$$\nu = \frac{\varepsilon_\ell}{1 + \varepsilon_n},$$

the ratio of the elasticity of effective labor supply to the elasticity of the wage bill $w\, n(w)$ with respect to the wage. Effort enters the numerator as a higher wage raises output per worker, but not the denominator, since effort does not add to the wage bill directly. This asymmetry drives the three results we now draw out.

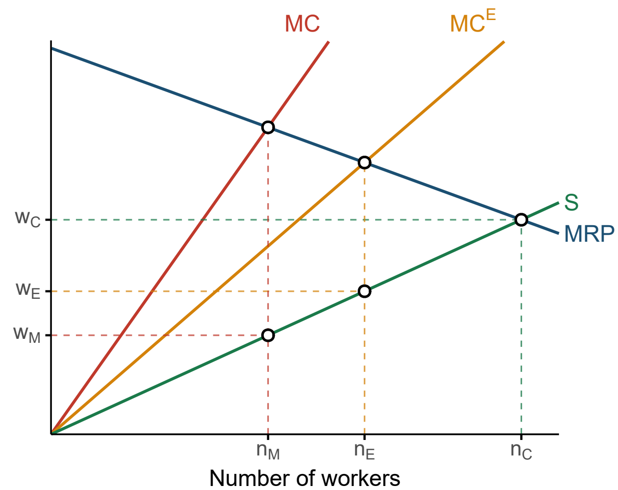
Effort responses attenuate labor market power. Decomposing,

$$\nu = \frac{\varepsilon_n}{1 + \varepsilon_n} + \frac{\varepsilon_e}{1 + \varepsilon_n} = \tilde{\nu} + \frac{\varepsilon_e}{1 + \varepsilon_n},$$

where $\tilde{\nu} \equiv \varepsilon_n/(1 + \varepsilon_n)$ is the standard monopsony markdown that obtains when effort is unresponsive to the wage ($\varepsilon_e = 0$). Whenever $\varepsilon_e > 0$ the true markdown ν strictly exceeds $\tilde{\nu}$, so the wedge between the wage and the marginal revenue product is narrower than the standard formula implies. A researcher who applies the standard formula recovers $\tilde{\nu}$ and thereby overstates the firm's labor market power by $\varepsilon_e/(1 + \varepsilon_n)$, the bias we quantify and correct in Section 3. The mechanism is that a wage cut costs the firm not only workers who leave but also effort from those who stay, raising the effective cost of wage suppression. Figure 1 illustrates. In the standard monopsony model the firm faces a downward-sloping marginal revenue product curve (MRP) and an upward-sloping curve S giving the number of workers supplied at each wage. Because hiring an additional worker requires raising the wage for all inframarginal workers, the marginal cost curve (MC) lies above S. Equating MRP with MC yields the standard monopsony outcome at (n_M, w_M) below the competitive benchmark (n_C, w_C). With effort responses, a wage increase also raises output per incumbent worker, so the effort-adjusted marginal cost MC^E lies between MC and S, and the optimum shifts to (n_E, w_E) with $n_M < n_E < n_C$ and $w_M < w_E < w_C$.

Unit-elastic effort eliminates labor market power. As $\varepsilon_e \to 1$, the markdown converges to one regardless of the headcount elasticity: $(\varepsilon_n + 1)/(1 + \varepsilon_n) = 1$. The intuition is that when

Figure 1: Monopsony with Effort Responses



Notes: S is the number of workers supplied at each wage and MRP the marginal revenue product. MC is the marginal cost curve in the standard monopsony model; MC^E is the effort-adjusted marginal cost curve.

effort rises proportionally with the wage, the effective cost per worker $w/e(w)$ is constant, so the firm faces perfectly elastic supply of effective labor and prices competitively. In Figure 1, MC^E rotates down toward S as $\varepsilon_e \to 1$, collapsing the monopsony wedge even for a workforce that is highly inelastic on the headcount margin. The same formula shows that effort responses can in principle more than offset monopsony power: for $\varepsilon_e > 1$ the firm sets $\nu > 1$, paying labor above its effective marginal revenue product because the wage's role in eliciting effort dominates its role in clearing the headcount margin.

Effort bounds the markdown even with an inelastic headcount. The effort margin limits how wide the markdown can become. As the headcount margin becomes completely inelastic ($\varepsilon_n \to 0$), the markdown converges to ε_e rather than to zero. This is because even when no rival can lure a worker away, each retains the ability to supply or withhold effort in response to the wage. The elasticity of effective labor supply, $\varepsilon_n + \varepsilon_e$, is therefore bounded below by $\varepsilon_e > 0$, so the firm cannot drive the effective wage arbitrarily far below the marginal revenue product.

3 Measuring Markdowns

Two broad approaches dominate the empirical measurement of markdowns. The *elasticity approach* infers markdowns from estimates of firm-specific labor supply elasticities. The *production approach* recovers the markdown from firms' optimality condition and an estimate of the production function. We show that the response of effort to wages makes the elasticity formula misspecified but leaves the production identity valid.

3.1 The Elasticity Approach

The elasticity approach estimates firm-specific labor supply elasticities (Staiger et al., 2010; Falch, 2010; Ransom and Sims, 2010)[2] and plugs them into the standard markdown formula

$$\tilde{\nu} = \left[\frac{1+\varepsilon_n}{\varepsilon_n}\right]^{-1} = \frac{\varepsilon_n}{1+\varepsilon_n}.$$

This is the firm's first-order condition under the assumption that effort does not respond to wages, and it is the special case of Proposition 1 with $\varepsilon_e = 0$. We call $\tilde{\nu}$ the misspecified markdown.

Proposition 2. *The true markdown satisfies*

$$1 - \nu = (1 - \varepsilon_e)(1 - \tilde{\nu}).$$

Whenever $\varepsilon_e > 0$, the true markdown ν strictly exceeds $\tilde{\nu}$. The complement of the markdown $1 - \nu$, which we refer to as the exploitation rate, is scaled down by $(1 - \varepsilon_e)$.

Correction. Elasticity-approach studies for the United States imply misspecified exploitation rates $1 - \tilde{\nu}$ ranging from roughly 0.09 to 0.48.[3] Correcting an elasticity-approach estimate requires only one input, i.e., the effort–wage elasticity ε_e. Fongoni (2024) surveys empirical estimates of the effort–wage elasticity across three approaches: production-function methods using firm- or industry-level data yield a mean of 0.51 (5 studies), field experiments a mean of 0.45 (6 studies), and laboratory experiments a mean of 1.21 (6 studies). Averaging across all 17 studies yields an overall mean elasticity of 0.72.

The three approaches recover the effort–wage elasticity with differing fidelity. Field experiments that randomize pay measure the causal response of effort to an own-wage change, the closest empirical counterpart to the model object. Production-function estimates relate output per worker to wages in observational data and may conflate the effort response with shifts in workforce composition or selection. Laboratory estimates are the largest; we treat them as

[2] These papers appear in the April 2010 issue of *Journal of Labor Economics*, "Modern Models of Monopsony in Labor Markets: Tests and Estimates," which helped establish the modern empirical monopsony literature.

[3] Equivalently, these studies report inverse-markdown estimates from about 1.10 to 1.93 under the older convention. We restrict attention to elasticity-approach markdown estimates for the United States, anchored by Webber (2015), Bassier et al. (2022), Berger et al. (2022), Lamadon et al. (2022), Kroft et al. (2025), and Roussille and Scuderi (2025); Webber (2015) sets the upper bound through an average firm-level labor supply elasticity of 1.08. Some studies for other countries find wider markdowns; see, e.g., Hirsch et al. (2010).

an upper bound, since the absence of an outside labor market and the salience of the wage manipulation are likely to amplify measured reciprocity.

Table 1 reports the corrected exploitation rates at both ends of the misspecified range $[0.09, 0.48]$ under each set of estimates. Applying the cross-study mean $\varepsilon_e = 0.72$ compresses this range to $[0.03, 0.13]$, substantially closer to competitive pricing than the misspecified estimates suggest. Even the most conservative correction, using field-experiment estimates alone ($\varepsilon_e = 0.45$), roughly halves the upper bound from $1 - \tilde{\nu} = 0.48$ to $1 - \nu = 0.26$. The laboratory column illustrates the boundary case of Section 2: with $\varepsilon_e = 1.21 > 1$ the correction returns $1 - \nu < 0$, a wage markup on effective labor.

Table 1: Estimated Values of $1 - \nu$ Under Alternative Effort–Wage Elasticities

	No adjustment	Adjusted ε_e estimates:			
	($\varepsilon_e = 0$)	All studies ($\varepsilon_e = 0.72$)	Prod. function ($\varepsilon_e = 0.51$)	Field exp. ($\varepsilon_e = 0.45$)	Lab exp. ($\varepsilon_e = 1.21$)
$1 - \nu$	[0.09,0.48]	[0.03,0.13]	[0.04,0.24]	[0.05,0.26]	[-0.10,-0.02]

Notes: Misspecified exploitation rates $1 - \tilde{\nu} \in [0.09, 0.48]$ are from U.S. elasticity-approach estimates. Corrected values follow from Proposition 2 as $1 - \nu = (1 - \varepsilon_e)(1 - \tilde{\nu})$. Effort-wage elasticity estimates are from Fongoni (2024): overall mean 0.72 (17 studies), production-function 0.51 (5 studies), field experiments 0.45 (6 studies), laboratory experiments 1.21 (6 studies).

3.2 The Production Approach

We now turn to the production approach. Developed in Brooks et al. (2021) and Yeh et al. (2022) building on De Loecker and Warzynski (2012) and De Loecker et al. (2020), this approach recovers the labor markdown by combining optimality conditions implied by the firm's cost minimization with an estimate of the output elasticity with respect to labor. We extend this approach to allow worker effort to respond to wages and find that the production-based markdown formula remains valid in the presence of an effort margin.

Suppose the firm produces according to a production function $Y = F(k, m, n\, e(w))$, where k denotes capital, m material inputs, n headcount and $e(w)$ effort per worker so that $n\, e(w)$ is the effective labor input. Material inputs are flexible and purchased competitively at a price the firm takes as given. Capital may be quasi-fixed and subject to adjustment costs; the production approach requires only that the firm uses at least one competitively priced flexible input. Labor, by contrast, is supplied to the firm along an upward-sloping curve, giving the firm monopsony power. On the output side, the firm faces a potentially downward-sloping demand curve and charges a product-market markup μ. The competitively priced flexible input identifies this markup because its first-order condition equates its output elasticity to its revenue share scaled by μ, and contains no input-market wedge of its own.

The firm chooses its inputs and wages, internalizing that the wage affects both the number of workers it employs and the effort supplied by each worker. Let $\theta_n \equiv \partial \ln F(k, m, ne)/\partial \ln n$ denote the output elasticity with respect to headcount and α_n the revenue share of labor. The following proposition shows that the standard production-based markdown formula continues to recover the true markdown ν even when effort responds to wages.

Proposition 3. *The markdown ν satisfies*

$$\nu \;=\; \mu \,\cdot\, \alpha_n \,\cdot\, \frac{1}{\theta_n}\,,$$

regardless of whether effort responds to wages.

Proof. Since $Y = F(k, m, ne)$ and effective labor is $\ell \equiv ne$, the marginal product of an additional worker on the headcount margin is

$$\frac{\partial Y}{\partial n} \;=\; F_\ell(k, m,\, ne) \,\cdot\, e,$$

where F_ℓ denotes the marginal product of effective labor. Taking the inverse markdown and multiplying and dividing by $n/(PY)$:

$$\nu^{-1} \;=\; \frac{R_Y \cdot \frac{\partial Y}{\partial n}}{w} \;=\; \frac{R_Y}{P} \,\cdot\, \frac{\frac{\partial Y}{\partial n} \cdot n}{Y} \,\cdot\, \frac{PY}{wn}.$$

We evaluate the three factors in turn. Since the firm's first-order condition for output equates marginal revenue to marginal cost, $R_Y = \mathrm{MC}$, we have $R_Y/P = \mathrm{MC}/P = \mu^{-1}$, the inverse of the product-market markup. The second factor is output elasticity with respect to headcount θ_n, by definition. The third factor is the inverse revenue share of labor, $PY/(wn) = \alpha_n^{-1}$. Combining these three factors yields $\nu^{-1} = \mu^{-1}\,\theta_n\,\alpha_n^{-1}$, or equivalently $\nu = \mu\,\alpha_n\,\theta_n^{-1}$. $\square$

Endogenous effort changes the optimal markdown but not the accounting identity on which the production approach rests. The intuition is as follows. The production approach applied to effective labor gives $\nu = \mu\,\alpha_\ell\,\theta_\ell^{-1}$, where α_ℓ is the revenue share of effective labor ℓ and θ_ℓ the output elasticity with respect to ℓ. Two facts make this coincide with the headcount version. First, the revenue shares agree ($\alpha_\ell = \alpha_n$) because total wage payment to effective labor and to workers are both wn. Second, the output elasticities agree ($\theta_\ell = \theta_n$) because headcount and effort enter production multiplicatively and hence a one-percent increase in headcount holding effort per worker fixed raises effective labor by one percent. Effort affects the level of effective labor without changing the identity.

Empirical implementation. The practical challenge is to estimate the partial elasticity θ_n. Suppose the production function is Cobb–Douglas. If effort responds to wages but is unobserved by the econometrician, the estimating equation is

$$\ln Y_{it} = \theta_k \ln k_{it} + \theta_m \ln m_{it} + \theta_n \ln n_{it} + \underbrace{\theta_n \ln e_{it} + \omega_{it}}_{\text{composite residual}},$$

where ω_{it} denotes true productivity shocks. Because wages and employment are jointly determined, the omitted effort component $\theta_n \ln e_{it}$ will generally be correlated with headcount. An estimator that addresses only the correlation between headcount and productivity is therefore insufficient. Consistent estimation requires the variation identifying θ_n to be orthogonal to the entire composite residual, including both ω_{it} and $\theta_n \ln e_{it}$. This additional orthogonality condition is the central empirical difficulty introduced by the effort margin.

Suppose one does not confront this additional empirical difficulty. To characterize the resulting bias, consider a tractable benchmark in which headcount and effort are isoelastic in the wage, $\ln n_{it} = \varepsilon_n \ln w_{it}$ and $\ln e_{it} = \varepsilon_e \ln w_{it}$, each up to an additive constant. Eliminating the wage gives $\ln e_{it} = (\varepsilon_e/\varepsilon_n) \ln n_{it}$, so the estimating equation reduces to

$$\ln Y_{it} \;=\; \theta_k \ln k_{it} + \theta_m \ln m_{it} + \tilde{\theta}_n \ln n_{it} + \omega_{it},$$

where

$$\tilde{\theta}_n \;\equiv\; \theta_n \cdot \frac{\varepsilon_n + \varepsilon_e}{\varepsilon_n}.$$

so the headcount coefficient is inflated by the factor $(\varepsilon_n + \varepsilon_e)/\varepsilon_n \geq 1$ and collapses to θ_n only when effort is unresponsive to the wage ($\varepsilon_e = 0$). Substituting $\tilde{\theta}_n$ for θ_n in the markdown formula in Proposition 3 yields

$$\mu\, \alpha_n\, \tilde{\theta}_n^{-1} \;=\; \left(\mu\, \alpha_n\, \theta_n^{-1}\right) \frac{\varepsilon_n}{\varepsilon_n + \varepsilon_e} \;=\; \frac{\varepsilon_n + \varepsilon_e}{1 + \varepsilon_n} \cdot \frac{\varepsilon_n}{\varepsilon_n + \varepsilon_e} \;=\; \frac{\varepsilon_n}{1 + \varepsilon_n} \;=\; \tilde{\nu}.$$

This is precisely the misspecified markdown obtained from the elasticity approach that ignores the effort margin. Thus, although the production approach is theoretically robust to endogenous effort, an implementation based on an output elasticity that ignores the effort margin reproduces the same bias as the elasticity approach.

There are two routes to identifying θ_n. The first is to isolate variation in headcount that is orthogonal to both productivity and effort. Beyond the exclusion restriction ordinarily imposed in production-function estimation that the instrument be uncorrelated with the productivity shock, a valid instrument here must also leave worker effort unaffected. This is more demanding, because effort responds to the wage and, through the worker's outside option, to conditions in the wider labor market. What is required instead is variation that moves the number of workers a firm hires through its recruiting technology while leaving the wage and the outside option unchanged, for instance, a shock to the firm's recruiting reach. The second route is to control directly for the determinants of effort. Suppose effort can be written as $e_{it} = e(w_{it}, x_{it})$, where x_{it} includes other observed determinants of effort, such as monitoring intensity or workers' outside options. Including a sufficiently flexible function of (w_{it}, x_{it}) absorbs the effort component of the residual. The remaining endogeneity of headcount with respect to productivity can then be addressed using standard production-function methods. In Section 4 we offer an alternative that sidesteps the estimation of θ_n altogether.

The implementation in Yeh et al. (2022) instruments current inputs with lagged inputs, including lagged headcount. Such instruments need not satisfy the additional exclusion restriction created by endogenous effort. If wage-setting is persistent, lagged headcount may predict current headcount partly through persistent wages and may therefore remain correlated with current effort. In that case, the estimated coefficient on headcount can incorporate the effort response and identify $\tilde{\theta}_n$ rather than the partial elasticity θ_n. Because $\tilde{\theta}_n > \theta_n$ when $\varepsilon_e > 0$, the estimated markdown $\mu\alpha_n/\tilde{\theta}_n$ is biased downward. Yeh et al. (2022) estimate an average ratio of marginal revenue product to wages of 1.53, implying an average markdown of $1/1.53 = 0.65$. The unobserved-effort margin implies that the implied markdown ν may be biased downward so that the labor market power may be overstated.

4 Monitoring and Labor Market Power

This section extends the baseline model by microfounding the effort function through an explicit monitoring technology in the spirit of Shapiro and Stiglitz (1984). The framework allows us to study how wages and markdowns respond to monitoring when workers choose effort optimally.

4.1 Model Setup

Worker's Problem. The firm offers a contract specifying a wage w and a required effort level e. A worker who complies with the contract receives utility

$$u(w) - h(e),$$

where $u(\cdot)$ is increasing and concave, and $h(\cdot)$ is increasing and convex in effort, with $h(0) = 0$.

Effort is not costlessly observed. Instead, the firm monitors workers through a stochastic monitoring technology. A worker who supplies less than the contractual effort requirement is detected with probability q and undetected with probability $1-q$. A worker caught shirking is dismissed and receives the outside utility $u(b)$, where $b > 0$ is the consumption-equivalent value of the worker's outside option. This outside option may represent home production, unemployment benefits, or employment at another firm. The parameter q indexes the intensity of the monitoring technology and is the object whose effect on the markdown we characterize below. Because every effort level below e leads to the same punishment when detected, a worker who shirks optimally supplies zero effort. Her expected utility from shirking is therefore

$$q\,u(b) + (1-q)\,u(w).$$

The worker complies if and only if:

$$u(w) - h(e) \;\geq\; q\,u(b) + (1-q)\,u(w).$$

Equivalently, the incentive constraint (IC) requires $h(e) \leq q\big[u(w) - u(b)\big]$.

Firm's Problem. Since the firm's revenue is increasing in effort, the firm therefore implements the highest effort level consistent with the worker's incentive constraint at a given (w, q). The binding IC yields the induced effort function:

$$e(w,q) = h^{-1}\left(q[u(w) - u(b)]\right). \tag{1}$$

A positive effort level requires $u(w) > u(b)$. Implicit differentiation of (1) gives

$$\frac{\partial e}{\partial w} = \frac{q\,u'(w)}{h'(e)} > 0, \qquad \frac{\partial e}{\partial q} = \frac{u(w) - u(b)}{h'(e)} > 0.$$

Higher wages increase the payoff lost upon dismissal and therefore raise effort through an efficiency-wage channel. More intensive monitoring also raises implementable effort by increasing the probability that shirking is detected and punished.

Let $n(w,q)$ denote the number of workers willing to accept employment at the firm under the contract induced by (w,q), with $\partial n(w,q)/\partial w > 0$. The firm takes the induced effort function $e(w,q)$ and the extensive-margin labor-supply function $n(w,q)$ as given and chooses its wage w to maximize profits:

$$\max_{w}\; R\Big(n(w,q)\,e(w,q)\Big) - w\,n(w,q), \tag{2}$$

where $R(\cdot)$ denotes revenue as a function of effective labor, with other inputs suppressed. The firm's wage choice internalizes the effect of wages on both the headcount margin and the effort margin.

4.2 Comparative Statics

Monitoring affects the markdown directly through effort and indirectly through the optimal wage. At a given wage, stronger monitoring raises effort and widens the markdown. The wage effect is, however, ambiguous. Monitoring may substitute for wages by making detection more likely, or complement wages by increasing the penalty from dismissal. The markdown widens only when wages rise by less than effort.

To isolate the effort channel, we impose two simplifying assumptions. First, revenue is linear in effective labor: $R(ne) = z\,ne$, where $z > 0$ denotes productivity.[4] Second, extensive-margin labor supply depends only on the wage, $n = n(w)$ with $n'(w) > 0$. Monitoring therefore affects the firm through worker effort but not directly through the number of workers willing to join it. Under these assumptions, the firm's problem becomes

$$\max_{w}\,(ze(w,q) - w)\,n(w),$$

and the markdown at the optimum is $\nu = w/(ze)$.

Let $w(q)$ denote the firm's optimal wage, and evaluate all effort derivatives at $(w(q), q)$. Totally differentiating the markdown with respect to monitoring gives

$$\frac{d\nu}{dq} = \underbrace{-\frac{\nu}{e}\frac{\partial e}{\partial q}}_{\text{direct effort effect}} + \underbrace{\frac{\nu}{w}\left(1 - \varepsilon_e\right) w'(q)}_{\text{wage effect}}. \tag{3}$$

where $\varepsilon_e \equiv \partial \log e/\partial \log w$ is the wage elasticity of effort. The direct effort effect is always negative: holding the wage fixed, stronger monitoring raises effort and therefore lowers ν. The wage effect captures both the change in pay and the effort response induced by that change. At an interior optimum with positive profits, $\nu < 1$, which is equivalent to $\varepsilon_e < 1$.[5] Consequently, an increase in the optimal wage raises ν and offsets the direct effort effect, whereas a decrease in the optimal wage reinforces it. If $w'(q) \leq 0$, monitoring unambiguously widens the markdown. If $w'(q) > 0$, the sign of $d\nu/dq$ depends on whether the wage increase is large enough to offset the direct increase in effort.

[4] This linear revenue specification arises when the firm is a price taker in the product market and operates a constant-returns-to-scale technology in which all inputs can be adjusted frictionlessly. The firm can then scale effective labor and the other inputs proportionally while preserving the optimal input mix, causing output, and hence revenue, to scale proportionally with effective labor.

[5] Under constant returns, positive profits require $ze > w$, or $\nu < 1$. By Proposition 1, this condition is equivalent to $\varepsilon_e < 1$. The constant-returns case therefore rules out $\varepsilon_e \geq 1$, which remains possible under the strictly concave revenue function in Section 2.

The direction and magnitude of the wage response depend on the shapes of the worker's utility and effort-cost functions. To state the relevant condition, define the employment utility surplus $s(w) \equiv u(w) - u(b)$, the firm's profit $\Pi(w, q) \equiv (ze(w, q) - w)n(w)$, and the following elasticities:

$$\varepsilon_h(e) \equiv \frac{eh'(e)}{h(e)}, \qquad \rho_h(e) \equiv \frac{eh''(e)}{h'(e)}, \qquad \mathcal{R}_u(w) \equiv -\frac{wu''(w)}{u'(w)}, \qquad \varepsilon_s(w) \equiv \frac{ws'(w)}{s(w)}.$$

Here, ε_h is the elasticity of total effort cost and ρ_h is the elasticity of marginal effort cost. On the utility side, $\mathcal{R}_u$ is the coefficient of relative risk aversion and ε_s is the wage elasticity of the worker's utility surplus from employment.

Proposition 4. *Suppose the firm-specific labor supply is isoelastic and $q \in (0, 1)$. Let $w^* > 0$ denote an optimal wage satisfying $\Pi_{ww}(w^*, q) < 0$, and let $e^* = e(w^*, q)$ denote the induced effort level. Monitoring strictly widens the markdown, $d(1 - \nu)/dq > 0$ (equivalently $d\nu/dq < 0$), if and only if*

$$\mathcal{R}_u(w^*) + \varepsilon_s(w^*) \; > \; \varepsilon_h(e^*) - \rho_h(e^*).$$

The proposition combines the direct and wage effects into a comparison between the curvature of utility and the curvature of effort costs. A larger $\mathcal{R}_u$ raises the left-hand side and therefore favors a widening of the markdown. On the cost side, what matters is the curvature of marginal effort cost relative to the elasticity of total effort cost: a larger ρ_h relative to ε_h lowers the right-hand side and also favors widening. Under isoelastic extensive-margin labor supply, its elasticity affects the level of the markdown but drops out of the condition determining the sign of its response to monitoring.[6]

A useful benchmark is the isoelastic effort-cost function $h(e) = e^\gamma/\gamma$ with $\gamma \geq 1$. In this case, $\varepsilon_h = \gamma$ and $\rho_h = \gamma - 1$, so that $\varepsilon_h - \rho_h = 1$ regardless of γ. The condition in Proposition 4 therefore reduces to $\mathcal{R}_u + \varepsilon_s > 1$. This inequality holds for CRRA preferences whenever the outside option is strictly positive and below the wage. We formalize this result in the following corollary.

Corollary 1. *Suppose the conditions of Proposition 4 hold. Suppose further that utility is CRRA with coefficient $\sigma > 0$,*

$$u(w) = \begin{cases} \dfrac{w^{1-\sigma} - 1}{1 - \sigma}, & \sigma \neq 1, \\ \log w, & \sigma = 1, \end{cases}$$

and that the effort cost is isoelastic and convex, $h(e) = e^\gamma/\gamma$ with $\gamma \geq 1$. Then monitoring strictly widens the markdown: $d(1 - \nu)/dq > 0$.

[6] With non-isoelastic extensive-margin supply, an additional term appears with the same sign as $\varepsilon_n'(w^*)$. If the labor supply elasticity is non-increasing in the wage, this term also pushes toward a wider markdown, so the condition in the proposition remains sufficient.

Proof. Since the isoelastic effort-cost function implies $\varepsilon_h - \rho_h = 1$, it suffices to show that the utility side of the condition always exceeds one. CRRA utility implies $\mathcal{R}_u = \sigma$ and, for $\sigma \neq 1$,

$$\varepsilon_s(w) = \frac{w\,u'(w)}{u(w) - u(b)} = \frac{1-\sigma}{1-x}, \qquad x \equiv (b/w)^{1-\sigma},$$

we have

$$\mathcal{R}_u + \varepsilon_s - 1 = (1-\sigma)\left(\frac{1}{1-x} - 1\right) = (1-\sigma)\,\frac{x}{1-x}.$$

When $\sigma < 1$, we have $1-\sigma > 0$ and $x \in (0,1)$, so $x/(1-x) > 0$; when $\sigma > 1$, we have $1-\sigma < 0$ and $x > 1$, so $x/(1-x) < 0$. The product is strictly positive in both cases. For $\sigma = 1$, $\mathcal{R}_u + \varepsilon_s = 1 + 1/\ln(w/b) > 1$ directly. Hence $\mathcal{R}_u(w) + \varepsilon_s(w) > 1$ for all $0 < b < w$. The condition in Proposition 4 is therefore satisfied, and monitoring strictly widens the markdown. □

This indicates that the condition in Proposition 4 is only weakly restrictive, making the proposition's prediction correspondingly strong. For monitoring instead to narrow the markdown, preferences would need to be sufficiently close to risk neutrality and the elasticity of effort costs would need to decline sufficiently rapidly with effort, a configuration not generated by standard parametric families.

4.3 Empirical Evidence

The monitoring model predicts that monitoring intensity affects the markdown a firm sets. Under standard parametric specifications, more intensive monitoring widens the markdown. We take this prediction to the data and ask whether more intensive monitoring is associated with wider markdowns. We use the management-practice survey of Bloom and Van Reenen (2007). Our measure of markdowns is based on the production approach, but the exercise does not require us to estimate the output elasticity with respect to headcount. Instead, we exploit within-industry variation in markdowns rather than their levels. Under the standard assumption in production-function estimation that the output elasticities are common to firms within an industry, these industry-level output elasticities are differenced out by industry fixed effects. This sidesteps the estimation challenge that unobserved effort poses for the production approach in Section 3.2. To our knowledge, this provides the first empirical evidence on how monitoring intensity affects the markdown a firm sets on labor.

Data. The Bloom and Van Reenen (2007) survey measures management practices among 732 medium-sized manufacturing firms in the United States, the United Kingdom, France, and Germany. The management scores were collected in 2004. Each firm receives separate scores for several dimensions of management, including performance monitoring, incentives, and target setting. Appendix V describes the construction of these scores. The monitoring score is the closest empirical counterpart to the monitoring technology q in the model: it measures how systematically the firm tracks, records, and reviews worker performance. We standardize the score so that a one unit change corresponds to a one standard deviation change. These management scores are merged with accounting data from Compustat and Amadeus, covering the period 1994–2004. We keep firm-year observations for which the variables needed to implement the production-function approach are available, in particular wages, employment, and materials. The resulting panel contains 2,774 firm-year observations for 433 firms.

Specification. The goal is to estimate

$$\ln \nu_{it} = \beta \, \text{Monitoring}_i + \gamma' X_{it} + \delta_j + \lambda_t + \kappa_c + u_{it},$$

where Monitoring_i is the standardized monitoring score, δ_j, λ_t, and κ_c denote industry, year, and country fixed effects, and X_{it} is a vector of controls that varies across specifications. In all specifications, X_{it} includes the standardized incentives and target-setting scores. We then add firm-age fixed effects, log employment, and a measure of product-market competition. The coefficient β, our object of interest, is the conditional semi-elasticity of the markdown with respect to monitoring.

The main empirical challenge is that the markdown ν_{it} is not directly observed. Under the assumptions of Section 3.2, Proposition 3 and the first-order condition for materials m_{it} imply

$$\ln \nu_{it} = \ln \frac{w_{it} n_{it}}{m_{it}} + \ln \frac{\theta_{m,it}}{\theta_{n,it}}.$$

Substituting this expression into the estimating equation above gives

$$\ln \frac{w_{it} n_{it}}{m_{it}} = \beta \, \text{Monitoring}_i + \gamma' X_{it} + \delta_j + \lambda_t + \kappa_c - \ln \frac{\theta_{m,it}}{\theta_{n,it}} + u_{it}.$$

We assume that the output-elasticity ratio can be decomposed into industry, year, and country components. The term $\ln(\theta_{m,it}/\theta_{n,it})$ is then absorbed by the industry, year, and country fixed effects. This follows the standard production-function assumption that firms within an estimation sample share common technology coefficients, but allows the relevant elasticity ratio to vary systematically across industries, years, and countries (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015). We can therefore use the log ratio of the wage bill to material expenditure, $\ln(w_{it} n_{it}/m_{it})$, as the dependent variable. After residualizing by these fixed effects, variation in this cost-share ratio coincides one-for-one with variation in the log markdown. A lower value therefore corresponds to a lower ν and hence a wider markdown. This approach does not recover the level of any firm's markdown. Its advantage is that it also does not require estimating $\theta_{n,it}$ from production data and the unobserved-effort problem discussed in Section 3.2 is therefore avoided.

Results. Table 2 reports the estimates. The coefficient on monitoring is negative and statistically significant in every specification. A one-standard-deviation increase in monitoring is associated with a decline of 0.15 to 0.19 log points in the labor-to-materials cost ratio. Under the maintained assumption that firms within an industry share common output elasticities, this translates one-for-one into wider markdowns. Firms with more intensive monitoring therefore pay workers a smaller fraction of their marginal revenue product, consistent with the model's prediction that monitoring widens the markdown under standard specifications. The association is not simply a generic management-quality relationship: Appendix VI shows that the coefficient on incentives is close to zero, while targets enter with the opposite sign.

Table 2: Markdowns and Monitoring Technology

	(1)	(2)	(3)	(4)
Monitoring score	-0.15^{**}	-0.19^{**}	-0.19^{**}	-0.17^{*}
	(0.07)	(0.09)	(0.09)	(0.09)
Industry, year, and country FE	✓	✓	✓	✓
Firm age FE		✓	✓	✓
Log employees			✓	✓
Competition FE				✓
N	2,774	2,770	2,770	2,770

Notes: The dependent variable is the log ratio of the wage bill to materials expenditure. The monitoring score is standardized to have unit standard deviation. All specifications control for the incentives and targets management scores. Standard errors clustered at the firm level are reported in parentheses. Data are from Bloom and Van Reenen (2007). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5 Conclusion

Workers respond to wages through both employment and effort. This effort margin is largely absent from standard models of monopsony power that focus on the employment margin. Conventional markdown measures based on firm-specific labor supply elasticities alone overstate labor market power. Our correction incorporates the effort–wage elasticity and substantially narrows existing markdown estimates.

The identity underlying the production approach remains valid with endogenous effort, but unobserved effort complicates estimation of the output elasticity of labor. Conventional instruments used in production-function estimation may be correlated with effort, causing the resulting markdown estimate to inherit this bias. Researchers must therefore either control for the determinants of effort or isolate variation in employment that is orthogonal to effort. For applications concerned with variation in markdowns rather than their levels, we provide a simpler strategy that controls for industry fixed effects to sidestep the production-function estimation.

Finally, the monitoring model links management practices to labor market power. Monitoring limits workers' ability to shirk and changes the firm's optimal tradeoff between wages and effort. Under standard preferences, stronger monitoring widens markdowns. Consistent with this result, firms with more intensive monitoring in Bloom and Van Reenen (2007) pay workers a smaller fraction of their marginal revenue product. The result points to a broader role for management practices in shaping the division of surplus, beyond the productivity effects emphasized in the existing literature.

Several directions for future research follow naturally. First, this paper uses monitoring as one concrete organizational force shaping the effort margin, but the effort-wage elasticity varies across firms for many other reasons, such as workplace culture. Second, our framework is deliberately stylized. Incorporating the effort channel into a quantitative general-equilibrium model of labor market power would allow its aggregate and distributional consequences to be reassessed, including implications for wages, employment, welfare, and inequality. Finally, because the effort-wage elasticity is the key sufficient statistic in our correction, improving its measurement is itself an important agenda. More credible estimates across workers, firms, and markets would sharpen both the assessment of monopsony power and the quantitative importance of the effort margin.

References

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 2015, *83* (6), 2411–2451.

Akerlof, George A, "Labor Contracts as Partial Gift Exchange," *The Quarterly Journal of Economics*, 1982, *97* (4), 543–569.

Akerlof, George A. and Janet L. Yellen, "The Fair Wage-Effort Hypothesis and Unemployment," *The Quarterly Journal of Economics*, 1990, *105* (2), 255–283.

Bassier, Ihsaan, Arindrajit Dube, and Suresh Naidu, "Monopsony in Movers," *Journal of Human Resources*, 2022, *57* (S), S50–s86.

Basu, Susanto and Miles S Kimball, "Cyclical Productivity with Unobserved Input Variation," 1997.

Berger, David, Kyle Herkenhoff, and Simon Mongey, "Labor Market Power," *American Economic Review*, April 2022, *112* (4), 1147–93.

Bewley, Truman F, *Why Wages Don't Fall during a Recession*, Harvard university press, 1999.

Bils, Mark and Yongsung Chang, "Welfare Costs of Sticky Wages When Effort can Respond," *Journal of Monetary Economics*, 2003, *50* (2), 311–330.

_ , _ , and Sun-Bin Kim, "How Sticky Wages in Existing Jobs Can Affect Hiring," *American Economic Journal: Macroeconomics*, 2022, *14* (1), 1–37.

Bloom, Nicholas and John Van Reenen, "Measuring and Explaining Management Practices Across Firms and Countries," *Quarterly Journal of Economics*, 2007, *122* (4), 1351–1408.

_ and _ , "Why Do Management Practices Differ Across Firms and Countries?," *Journal of Economic Perspectives*, 2010, *24* (1), 203–224.

_ , Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does Management Matter? Evidence from India," *The Quarterly Journal of Economics*, 2013, *128* (1), 1–51.

_ , Erik Brynjolfsson, Lucia Foster, Ron Jarmin, Megha Patnaik, Itay Saporta-Eksten, and John Van Reenen, "What Drives Differences in Management Practices?," *American Economic Review*, 2019, *109* (5), 1648–1683.

_ , Raffaella Sadun, and John Van Reenen, "The Organization of Firms Across Countries," *The Quarterly Journal of Economics*, 2012, *127* (4), 1663–1705.

— , **Renata Lemos, Raffaella Sadun, Daniela Scur, and John Van Reenen**, "JEEA-FBBVA Lecture 2013: The New Empirical Economics of Management," *Journal of the European Economic Association*, 2014, *12* (4), 835–876.

Brooks, Wyatt J, Joseph P Kaboski, Yao Amber Li, and Wei Qian, "Exploitation of Labor? Classical Monopsony Power and Labor's Share," *Journal of Development Economics*, 2021, *150*, 102627.

Bruhn, Miriam, Dean Karlan, and Antoinette Schoar, "The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico," *Journal of Political Economy*, 2018, *126* (2), 635–687.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo, "Labor Hoarding and the Business Cycle," *Journal of Political Economy*, 1993, *101* (2), 245–273.

Danthine, Jean-Pierre and John B Donaldson, "Efficiency Wages and the Business Cycle Puzzle," *European Economic Review*, 1990, *34* (7), 1275–1301.

De Loecker, Jan and Frederic Warzynski, "Markups and Firm-Level Export Status," *American Economic Review*, 2012, *102* (6), 2437–2471.

— , **Jan Eeckhout, and Gabriel Unger**, "The Rise of Market Power and the Macroeconomic Implications," *The Quarterly Journal of Economics*, 2020, *135* (2), 561–644.

Falch, Torberg, "The Elasticity of Labor Supply at the Establishment Level," *Journal of Labor Economics*, 2010, *28* (2), 237–266.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl, "Does Fairness Prevent Market Clearing? An Experimental Investigation," *The Quarterly Journal of Economics*, 1993, *108* (2), 437–459.

Fongoni, Marco, "Asymmetric Reciprocity and the Cyclical Behavior of Wages, Effort, and Job Creation," *American Economic Journal: Macroeconomics*, July 2024, *16* (3), 52–89.

Galí, Jordi and Thijs Van Rens, "The Vanishing Procyclicality of Labour Productivity," *The Economic Journal*, 2021, *131* (633), 302–326.

Giorcelli, Michela, "The Long-Term Effects of Management and Technology Transfers," *American Economic Review*, 2019, *109* (1), 121–152.

Hirsch, Boris, Thorsten Schank, and Claus Schnabel, "Differences in Labor Supply to Monopsonistic Firms and the Gender Pay Gap: An Empirical Analysis Using Linked Employer-Employee Data from Germany," *Journal of Labor Economics*, 2010, *28* (2), 291–330.

Katz, Lawrence F, "Efficiency Wage Theories: A Partial Evaluation," *NBER Macroeconomics Annual*, 1986, *1*, 235–276.

Kroft, Kory, Yao Luo, Magne Mogstad, and Bradley Setzler, "Imperfect Competition and Rents in Labor and Product Markets: The Case of the Construction Industry," *American Economic Review*, September 2025, *115* (9), 2926–2969.

Lamadon, Thibaut, Magne Mogstad, and Bradley Setzler, "Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market," *American Economic Review*, January 2022, *112* (1), 169–212.

Leibenstein, Harvey, *Economic Backwardness and Economic Growth: Studies in the Theory of Economic Development*, Wiley, New York, 1957.

Levinsohn, James and Amil Petrin, "Estimating Production Functions Using Inputs to Control for Unobservables," *Review of Economic Studies*, 2003, *70* (2), 317–341.

Manning, Alan, "Imperfect Competition in the Labor Market," in "Handbook of Labor Economics," Vol. 4, Elsevier, 2011, pp. 973–1041.

Olley, G. Steven and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 1996, *64* (6), 1263–1297.

Ransom, Michael R and David P Sims, "Estimating the Firm's Labor Supply Curve in a "New Monopsony" Framework: Schoolteachers in Missouri," *Journal of Labor Economics*, 2010, *28* (2), 331–355.

Roussille, Nina and Benjamin Scuderi, "Bidding for Talent: A Test of Conduct in a High-Wage Labor Market," NBER Working Paper 33848, National Bureau of Economic Research May 2025.

Salop, Steven C, "A Model of the Natural Rate of Unemployment," *The American Economic Review*, 1979, *69* (1), 117–125.

Seegmiller, Bryan, "Valuing Labor Market Power: The Role of Productivity Advantages," Working Paper, Kellogg School of Management 2025.

Shapiro, Carl and Joseph E Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," *The American Economic Review*, 1984, *74* (3), 433–444.

Sokolova, Anna and Todd Sorensen, "Monopsony in Labor Markets: A Meta-Analysis," *ILR Review*, 2021, *74* (1), 27–55.

Staiger, Douglas O, Joanne Spetz, and Ciaran S Phibbs, "Is There Monopsony in the Labor Market? Evidence from a Natural Experiment," *Journal of Labor Economics*, 2010, *28* (2), 211–236.

Stiglitz, Joseph E., "The Efficiency Wage Hypothesis, Surplus Labour, and the Distribution of Income in L.D.C.s," *Oxford Economic Papers*, 1976, *28* (2), 185–207.

Webber, Douglas A., "Firm Market Power and the Earnings Distribution," *Labour Economics*, 2015, *35*, 123–134.

Weiss, Andrew, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political economy*, 1980, *88* (3), 526–538.

Yeh, Chen, Claudia Macaluso, and Brad Hershbein, "Monopsony in the US Labor Market," *American Economic Review*, July 2022, *112* (7), 2099–2138.

Yellen, Janet L, "Efficiency Wage Models of Unemployment," *The American Economic Review*, 1984, *74* (2), 200–205.

I Background on Efficiency Wage

Our baseline analysis takes the effort function $e(w)$ as a primitive of the firm's problem. This subsection situates that choice within the efficiency-wage literature, which has advanced several distinct mechanisms through which wages raise worker effort. The efficiency-wage idea originates in models of low-income economies, where higher wages improve nutrition and health and thereby raise productive capacity (Leibenstein, 1957; Stiglitz, 1976). The modern literature, surveyed by Yellen (1984) and Katz (1986), identifies channels relevant to industrialized labor markets. We focus on the two that act on the effort of employed workers and therefore map directly onto our reduced-form $e(w)$: the incentive channel and the sociological channel.[7]

Incentive channel. The most influential mechanism is the shirking model of Shapiro and Stiglitz (1984), in which firms cannot continuously observe effort and detected shirking is punished only by dismissal. Workers weigh the utility of supplying effort against the expected utility of shirking and being caught; a higher wage raises the cost of dismissal and so makes compliance with the firm's required effort level more attractive. The labor-turnover model of Salop (1979) is formally analogous: when separations are costly to the firm, a higher wage reduces quits and conserves hiring and training expenditure, preserving firm-specific human capital.

Sociological channel. A second line of work links wages to effort through social and relational considerations. Akerlof (1982) models the employment relationship as a partial gift exchange in which workers reciprocate wages paid above their reservation level with effort above the contractual minimum, while firms anticipate this reciprocity when setting wages. Akerlof and Yellen (1990) formalize a closely related "fair wage-effort hypothesis" under which workers withhold effort in proportion to the gap between their actual wage and a reference fair wage. Survey evidence on workplace morale is consistent with firms internalizing such norms (Bewley, 1999). Fehr et al. (1993) provide experimental support, documenting systematic effort responses to wage offers in bilateral exchange games.

These mechanisms share a common reduced-form implication: effort, defined as output per worker conditional on hours, is increasing in the own wage. The foundational models establish

[7] We set aside the adverse-selection channel of Weiss (1980), in which a higher posted wage draws applicants of higher unobserved ability and so raises average labor quality. Although it too implies that output per worker rises with the wage, it operates on the composition of hires rather than on the effort of an employed worker.

this relationship for competitive labor markets, in which firms face perfectly elastic labor supply and pay above market-clearing to elicit effort rather than to exploit any market power. Our contribution is to ask what the same positive response implies once the firm sets wages along an upward-sloping labor supply curve, and in particular how it biases conventional estimates of labor-market power.

II Productivity and Markdowns

This appendix studies how changes in a firm's productivity affect its optimal markdown. We introduce a multiplicative productivity shifter $z > 0$ into the revenue function. Technically, z is the productivity of the only firm in the economy. However, we hold the workers' outside option b and the extensive-margin labor supply function $n(\cdot)$ fixed, so the exercise is best interpreted as a single firm experiencing an idiosyncratic productivity improvement while the rest of the economy remains unaffected. The firm's problem (2) becomes

$$\max_{w} \; zR\Big(n(w,q)\, e(w,q)\Big) - w\, n(w,q). \tag{A1}$$

Since the markdown $\nu = (\varepsilon_n + \varepsilon_e)/(1+\varepsilon_n)$ depends on z only through the optimal wage $w^*(z)$—the elasticities ε_n and ε_e are themselves functions of w—productivity affects markdowns entirely through the wage it induces. We therefore begin by characterizing the direction of that wage response.

Lemma 1. *An increase in productivity z raises the optimal wage: $dw^*/dz > 0$.*

The proof is in Appendix IV. The intuition is straightforward: a more productive firm obtains a higher return from each unit of effective labor, making it more valuable to attract workers and induce effort, which raises the wage. To state when this positive wage response translates into wider markdowns, recall the coefficient of relative risk aversion $\mathcal{R}_u(w) = -wu''(w)/u'(w)$ and the effort-cost elasticity $\varepsilon_h(e) = eh'(e)/h(e)$ defined in Section 4.

Proposition 5. *Let the utility function satisfy $\mathcal{R}_u(w) > 1$ over the relevant wage range, the effort cost function exhibit non-decreasing effort-cost elasticity ε_h, and the extensive-margin labor supply function exhibit non-increasing wage elasticity. Then, at any optimum with $\nu \leq 1$, productivity strictly widens markdowns: $d(1-\nu)/dz > 0$, or equivalently $d\nu/dz < 0$.*

The proof is in Appendix IV. By Lemma 1, higher productivity raises the optimal wage, and that wage increase reshapes markdowns through both the extensive- and intensive-margin elasticities. Under the proposition's assumptions, the two channels point in the same direction.

A non-increasing ε_n already widens markdowns as wages rise, a direction consistent with the empirical pattern of wider markdowns at more productive firms documented in Seegmiller (2025). A common assumption in the literature is to take the extensive-margin labor supply to be isoelastic; in that case the extensive-margin contribution drops out entirely, and the markdown's response to productivity is driven exclusively by the intensive margin.

Under the proposition's assumptions, the effort-wage elasticity ε_e is strictly decreasing in the wage. Since $\nu = (\varepsilon_n + \varepsilon_e)/(1 + \varepsilon_n)$ is increasing in ε_e, a fall in ε_e lowers ν and widens the markdown. To understand why ε_e falls with the wage, we differentiate the binding incentive constraint with respect to w and rearrange, which yields the decomposition

$$\varepsilon_e(w) = \frac{\varepsilon_s(w)}{\varepsilon_h(e(w))},$$

where $\varepsilon_s(w) = wu'(w)/[u(w) - u(b)]$ is the surplus elasticity and ε_h the effort-cost elasticity, both defined in Section 4. The proposition's conditions on $\mathcal{R}_u$ and ε_h each discipline one piece of this ratio: the relative-risk-aversion condition makes ε_s strictly decreasing in the wage, while non-decreasing ε_h keeps the denominator of ε_e from falling as effort rises with the wage. Economically, diminishing marginal utility of income means each additional dollar of wages purchases proportionally less utility surplus, while a non-decreasing elasticity of effort costs makes translating that surplus into effort increasingly expensive in proportional terms.

The relative-risk-aversion condition is stronger than needed. A more general sufficient condition for ε_s to be decreasing in the wage is $-wu''(w)/u'(w) > 1 - wu'(w)/[u(w) - u(b)]$, i.e., $\mathcal{R}_u(w) + \varepsilon_s(w) > 1$, which is satisfied by any CRRA utility with a positive outside option, even when the coefficient of relative risk aversion is below one. The result therefore extends well beyond the parametric class implied by $\mathcal{R}_u > 1$.

In this isoelastic benchmark, our model delivers $1 - \nu = [1 - \varepsilon_e(w)]/(1 + \bar{\varepsilon})$, strictly increasing in z because ε_e falls as w rises. A model that abstracts from the effort margin sets $\varepsilon_e = 0$ and yields $1 - \nu = 1/(1 + \bar{\varepsilon})$, a markdown entirely invariant to productivity. The intensive margin is therefore essential for generating a positive relationship between productivity and wider markdowns in this benchmark case.

These results deliver a clear economic message: a more productive firm does not only attract more workers by offering higher wages, growing on the extensive margin. It also extracts greater rents from each worker it employs, as higher wages weaken the effort-wage incentive and allow the firm to widen the wedge between productivity and pay. Productivity thus amplifies labor market power through both margins simultaneously, a dual effect entirely absent from the standard isoelastic benchmark where markdowns are invariant to z and productivity affects only employment, not rent extraction.

III Monitoring and Profits

This appendix records how a firm's profits respond to its monitoring technology, under the two assumptions maintained in Section 4.2: constant returns, $R(ne) = z\,ne$, and an extensive margin that depends on the wage alone, $n = n(w)$. The second assumption is what makes the exercise clean. Because monitoring cannot alter the number of workers willing to join the firm, it reaches profits only by raising the effort of those already employed, an unambiguously beneficial change. The more realistic case, in which more intensive monitoring also deters some workers from joining, would set this effort gain against a loss of headcount; we leave that richer problem aside and report only the effort-channel benchmark.

Proposition 6. *Firm profits are strictly increasing in monitoring technology q.*

The proof is in Appendix IV and applies the envelope theorem: better monitoring raises effort at any given wage, generating a direct profit gain, while the firm's optimal wage adjustment in response has only a second-order effect and does not offset it.

IV Proofs

Proposition 1. *The optimal markdown satisfies*

$$\nu = \frac{\varepsilon_n + \varepsilon_e}{1 + \varepsilon_n},$$

where all elasticities are evaluated at w^*.

Proof. For ease of exposition, we omit the arguments of all functions. The first-order condition with respect to w is:

$$\frac{d}{dw}\left[R(ne) - wn\right] = 0.$$

Applying the product rule:

$$R'(n'e + ne') - n - wn' = 0.$$

Dividing through by n:

$$R'\left(e\frac{n'}{n} + e'\right) - 1 - w\frac{n'}{n} = 0.$$

Let $\varepsilon_n = w\frac{n'}{n}$ (the wage elasticity of headcount) and $\varepsilon_e = w\frac{e'}{e}$ (the wage elasticity of effort). Then $\frac{n'}{n} = \frac{\varepsilon_n}{w}$ and $e' = \frac{\varepsilon_e e}{w}$.

Substituting:

$$R'\left(e\frac{\varepsilon_n}{w} + \frac{\varepsilon_e e}{w}\right) - 1 - \varepsilon_n = 0.$$

Simplifying:

$$R'e\frac{\varepsilon_n + \varepsilon_e}{w} = 1 + \varepsilon_n.$$

Solving for $R'e$:

$$R'e = \frac{w(1 + \varepsilon_n)}{\varepsilon_n + \varepsilon_e}.$$

The markdown is defined as $\nu = \frac{w/e(w)}{R'}$, which can be rewritten as $\nu = \frac{w}{R'e}$. Substituting the expression for $R'e$:

$$\nu = \frac{w}{\frac{w(1+\varepsilon_n)}{\varepsilon_n+\varepsilon_e}} = \frac{\varepsilon_n + \varepsilon_e}{1 + \varepsilon_n}.$$

□

Proposition 2. *The true markdown satisfies*

$$1 - \nu = (1 - \varepsilon_e)(1 - \tilde{\nu}).$$

Proof. The misspecified markdown $\tilde{\nu}$ is obtained by setting $\varepsilon_e = 0$ in the optimal markdown formula of Proposition 1:

$$\tilde{\nu} = \frac{\varepsilon_n}{1 + \varepsilon_n}.$$

Solving for ε_n gives $\varepsilon_n = \tilde{\nu}/(1 - \tilde{\nu})$. Substituting into $\nu = (\varepsilon_n + \varepsilon_e)/(1 + \varepsilon_n)$:

$$\nu = \frac{\dfrac{\tilde{\nu}}{1-\tilde{\nu}} + \varepsilon_e}{1 + \dfrac{\tilde{\nu}}{1-\tilde{\nu}}} = \tilde{\nu} + \varepsilon_e(1 - \tilde{\nu}).$$

Rearranging, $1 - \nu = (1 - \varepsilon_e)(1 - \tilde{\nu})$. □

Lemma 1. *An increase in productivity z raises the optimal wage: $dw^*/dz > 0$.*

Proof. The first-order condition for (A1) is

$$F(w, z) \equiv zR'\big(n(w, q)\, e(w, q)\big)\big(n_w(w, q)\, e(w, q) + n(w, q)\, e_w(w, q)\big) - n(w, q) - w\, n_w(w, q) = 0,$$

where $n_w \equiv \partial n/\partial w$. Differentiating with respect to z:

$$F_z = R'(ne)\big(n_w e + n e_w\big) > 0,$$

since $R' > 0$, $n_w > 0$, $e > 0$, $n > 0$, and $e_w > 0$. The second-order condition requires $F_w < 0$. By the implicit function theorem, $dw^*/dz = -F_z/F_w > 0$. □

Define $\mathcal{R}_u(w) \equiv -wu''(w)/u'(w)$, $\varepsilon_h(e) \equiv eh'(e)/h(e)$, $\rho_h(e) \equiv eh''(e)/h'(e)$, and $\varepsilon_s(w) \equiv$

$wu'(w)/[u(w) - u(b)]$.

Proposition 5. *Let the utility function satisfy $\mathcal{R}_u(w) > 1$ over the relevant wage range, the effort cost function exhibit non-decreasing effort-cost elasticity ε_h, and the extensive-margin labor supply function exhibit non-increasing wage elasticity. Then, at any optimum with $\nu \leq 1$, productivity strictly widens markdowns: $d(1-\nu)/dz > 0$, or equivalently $d\nu/dz < 0$.*

Proof. Since $dw^*/dz > 0$ by Lemma 1, we have $\text{sign}(d\nu/dz) = \text{sign}(d\nu/dw)$, where

$$\frac{d\nu}{dw} = \frac{\varepsilon_n'(1-\varepsilon_e) + (1+\varepsilon_n)\varepsilon_e'}{(1+\varepsilon_n)^2},$$

with $\varepsilon_n' \equiv d\varepsilon_n/dw$ and $\varepsilon_e' \equiv d\varepsilon_e/dw$.

We first establish that $\varepsilon_e'(w) < 0$. Differentiating the binding IC (1) with respect to w and rearranging yields

$$\varepsilon_e(w) = \frac{\varepsilon_s(w)}{\varepsilon_h(e(w))},$$

where $\varepsilon_s(w) \equiv wu'(w)/[u(w) - u(b)]$ is the elasticity of the net utility surplus and $\varepsilon_h(e) \equiv eh'(e)/h(e)$ is the elasticity of the effort cost function. The effort-cost-elasticity assumption implies $\varepsilon_h(e(w))$ is non-decreasing in w, since $e(w)$ is increasing. It remains to show that ε_s is strictly decreasing. Its log derivative satisfies

$$\frac{d\log\varepsilon_s(w)}{d\log w} = 1 + \frac{wu''(w)}{u'(w)} - \frac{wu'(w)}{u(w)-u(b)} = 1 - \mathcal{R}_u(w) - \varepsilon_s(w),$$

where $\mathcal{R}_u(w) \equiv -wu''(w)/u'(w)$ is relative risk aversion. Since $\mathcal{R}_u(w) > 1$ and $\varepsilon_s(w) > 0$, this derivative is strictly negative. Thus ε_s is strictly decreasing in w, while $\varepsilon_h(e(w))$ is non-decreasing, so $\varepsilon_e'(w) < 0$.

With $\varepsilon_e' < 0$, the second term satisfies $(1+\varepsilon_n)\varepsilon_e' < 0$. At any optimum with $\nu \leq 1$, the markdown formula implies $\varepsilon_e \leq 1$, so $1 - \varepsilon_e \geq 0$ and the first term satisfies $\varepsilon_n'(1-\varepsilon_e) \leq 0$ when $\varepsilon_n' \leq 0$. Both terms are then non-positive and the second is strictly negative, so $d\nu/dw < 0$. Since $dw^*/dz > 0$, it follows that $d\nu/dz < 0$ and $d(1-\nu)/dz > 0$. □

Proposition 6. *Firm profits are strictly increasing in monitoring technology q.*

Proof. Let $\Pi^*(q) = (ze(w^*(q), q) - w^*(q))n(w^*(q))$ denote maximized profits. By the envelope theorem:

$$\frac{d\Pi^*}{dq} = \frac{\partial}{\partial q}\left[(ze(w,q) - w)n(w)\right]\bigg|_{w=w^*(q)} = z\frac{\partial e}{\partial q}n(w^*(q)) > 0,$$

since $\partial e/\partial q = (u(w) - u(b))/h'(e) > 0$ (workers prefer employment to the outside option) and $n > 0$. □

Proof of Proposition 4

Proposition 4. *Suppose the firm-specific labor supply is isoelastic and $q \in (0,1)$. Let $w^* > 0$ denote an optimal wage satisfying $\Pi_{ww}(w^*, q) < 0$, and let $e^* = e(w^*, q)$ denote the induced effort level. Monitoring strictly widens the markdown, $d(1-\nu)/dq > 0$ (equivalently $d\nu/dq < 0$), if and only if*

$$\mathcal{R}_u(w^*) + \varepsilon_s(w^*) \;>\; \varepsilon_h(e^*) - \rho_h(e^*).$$

The proof proceeds in three lemmas. Lemma 2 derives the elasticities of the effort function with respect to the wage and to monitoring from the binding incentive constraint. Lemma 3 computes how the surplus elasticity ε_s and the effort-cost elasticity ε_h respond to their arguments, and with them how the effort–wage elasticity ε_e moves with the wage and with monitoring. Lemma 4 applies the implicit function theorem to the firm's first-order condition to obtain the response of the optimal wage to monitoring. The proof of the proposition then combines the three lemmas in the identity $d\ln\nu/d\ln q = (1-\varepsilon_e)\,\eta - 1/\varepsilon_h$, where η is the elasticity of the optimal wage with respect to monitoring, and signs the result.

Throughout, the revenue function exhibits constant returns, $R(ne) = z\,ne$, and the extensive-margin labor supply is isoelastic, $n(w) = \bar{n}\,w^{\varepsilon_n}$, so profits are $\Pi(w,q) = (ze(w,q) - w)\,n(w)$ with $e(w,q)$ given by the binding incentive constraint (1). All elasticities are evaluated at the optimum, and we suppress function arguments where convenient. Multiplying the first-order condition $\Pi_w = 0$ by w/n puts it in elasticity form:

$$F(w,q) \equiv ze(w,q)\,[\varepsilon_e(w,q) + \varepsilon_n] - w(1+\varepsilon_n) = 0, \tag{A2}$$

where $\varepsilon_e = we_w/e$.

Lemma 2. *The effort function defined by the binding incentive constraint* (1) *satisfies*

$$\frac{\partial \ln e}{\partial \ln w} = \varepsilon_e = \frac{\varepsilon_s(w)}{\varepsilon_h(e)}, \qquad \frac{\partial \ln e}{\partial \ln q} = \frac{1}{\varepsilon_h(e)}.$$

Proof. Differentiating $h(e) = q\,[u(w) - u(b)]$ with respect to w and q gives

$$h'(e)\,e_w = qu'(w), \qquad h'(e)\,e_q = u(w) - u(b).$$

For the first claim, solve the first equation for e_w, multiply by w/e, and then multiply and divide by $u(w) - u(b)$:

$$\frac{\partial \ln e}{\partial \ln w} = \frac{we_w}{e} = \frac{wqu'(w)}{e\,h'(e)} = \underbrace{\frac{wu'(w)}{u(w)-u(b)}}_{=\varepsilon_s(w)} \cdot \frac{q\,[u(w)-u(b)]}{e\,h'(e)} = \varepsilon_s(w) \cdot \frac{h(e)}{e\,h'(e)} = \frac{\varepsilon_s(w)}{\varepsilon_h(e)},$$

where the fourth equality substitutes the incentive constraint $q\,[u(w)-u(b)] = h(e)$ and the last uses the definition $\varepsilon_h = eh'/h$. For the second claim, solve the second equation for e_q, multiply by q/e, and substitute the incentive constraint again:

$$\frac{\partial \ln e}{\partial \ln q} = \frac{qe_q}{e} = \frac{q\,[u(w)-u(b)]}{e\,h'(e)} = \frac{h(e)}{e\,h'(e)} = \frac{1}{\varepsilon_h(e)}. \qquad \square$$

Lemma 3. *Define $\sigma_h(e) \equiv d\ln\varepsilon_h(e)/d\ln e$, the elasticity of the effort-cost elasticity. Then*

$$\frac{d\ln\varepsilon_s(w)}{d\ln w} = 1 - \mathcal{R}_u(w) - \varepsilon_s(w), \qquad \sigma_h(e) = 1 + \rho_h(e) - \varepsilon_h(e),$$

and the effort–wage elasticity $\varepsilon_e = \varepsilon_s(w)/\varepsilon_h(e(w,q))$ satisfies

$$\frac{\partial \ln\varepsilon_e}{\partial \ln w} = (1 - \mathcal{R}_u - \varepsilon_s) - \sigma_h\,\varepsilon_e, \qquad \frac{\partial \ln\varepsilon_e}{\partial \ln q} = -\frac{\sigma_h}{\varepsilon_h}.$$

Proof. Taking logs of the definition of the surplus elasticity, $\ln\varepsilon_s(w) = \ln w + \ln u'(w) - \ln\,[u(w)-u(b)]$, and differentiating with respect to $\ln w$ term by term:

$$\frac{d\ln\varepsilon_s}{d\ln w} = 1 + \frac{wu''(w)}{u'(w)} - \frac{wu'(w)}{u(w)-u(b)} = 1 - \mathcal{R}_u(w) - \varepsilon_s(w).$$

Likewise, $\ln\varepsilon_h(e) = \ln e + \ln h'(e) - \ln h(e)$, so

$$\sigma_h(e) = \frac{d\ln\varepsilon_h}{d\ln e} = 1 + \frac{eh''(e)}{h'(e)} - \frac{eh'(e)}{h(e)} = 1 + \rho_h(e) - \varepsilon_h(e).$$

For the effort–wage elasticity, Lemma 2 gives $\ln\varepsilon_e = \ln\varepsilon_s(w) - \ln\varepsilon_h(e(w,q))$, where the second term depends on w and q only through $e(w,q)$. Differentiating with respect to $\ln w$, the chain rule and Lemma 2 give

$$\frac{\partial \ln\varepsilon_e}{\partial \ln w} = \frac{d\ln\varepsilon_s}{d\ln w} - \sigma_h\,\frac{\partial \ln e}{\partial \ln w} = (1 - \mathcal{R}_u - \varepsilon_s) - \sigma_h\,\varepsilon_e.$$

Since $\varepsilon_s(w)$ does not depend on q, only the effort-cost term responds to monitoring:

$$\frac{\partial \ln\varepsilon_e}{\partial \ln q} = -\sigma_h\,\frac{\partial \ln e}{\partial \ln q} = -\frac{\sigma_h}{\varepsilon_h}. \qquad \square$$

Lemma 4. *Define*

$$D \equiv (\varepsilon_e + \varepsilon_n)(\varepsilon_e - 1) + \varepsilon_e\left(1 - \mathcal{R}_u - \varepsilon_s - \sigma_h\,\varepsilon_e\right).$$

Then $D < 0$, and the elasticity of the optimal wage with respect to monitoring is

$$\eta \equiv \frac{d \ln w^*}{d \ln q} = -\frac{\varepsilon_e + \varepsilon_n - \sigma_h \, \varepsilon_e}{\varepsilon_h \, D}.$$

Proof. Write $\hat{F}_x \equiv \partial F / \partial \ln x$ for $x \in \{w, q\}$, with F defined in (A2).

The monitoring derivative. In F, only e and ε_e depend on q. Using $\partial e / \partial \ln q = e \cdot \partial \ln e / \partial \ln q = e / \varepsilon_h$ from Lemma 2 and $\partial \varepsilon_e / \partial \ln q = \varepsilon_e \cdot \partial \ln \varepsilon_e / \partial \ln q = -\sigma_h \, \varepsilon_e / \varepsilon_h$ from Lemma 3,

$$\hat{F}_q = z \, \frac{\partial e}{\partial \ln q} \, (\varepsilon_e + \varepsilon_n) + ze \, \frac{\partial \varepsilon_e}{\partial \ln q} = ze \left[\frac{\varepsilon_e + \varepsilon_n}{\varepsilon_h} - \frac{\sigma_h \, \varepsilon_e}{\varepsilon_h} \right] = \frac{ze}{\varepsilon_h} \, (\varepsilon_e + \varepsilon_n - \sigma_h \, \varepsilon_e) \, .$$

The wage derivative. Differentiating F with respect to $\ln w$ term by term,

$$\hat{F}_w = z \, \frac{\partial e}{\partial \ln w} \, (\varepsilon_e + \varepsilon_n) + ze \, \frac{\partial \varepsilon_e}{\partial \ln w} - w(1 + \varepsilon_n).$$

The first term equals $ze \, \varepsilon_e (\varepsilon_e + \varepsilon_n)$ by Lemma 2; the second equals $ze \, \varepsilon_e \, (1 - \mathcal{R}_u - \varepsilon_s - \sigma_h \, \varepsilon_e)$ by Lemma 3; and in the third, the first-order condition (A2) gives $w(1 + \varepsilon_n) = ze(\varepsilon_e + \varepsilon_n)$. Hence

$$\hat{F}_w = ze \left[\varepsilon_e (\varepsilon_e + \varepsilon_n) + \varepsilon_e \, (1 - \mathcal{R}_u - \varepsilon_s - \sigma_h \, \varepsilon_e) - (\varepsilon_e + \varepsilon_n) \right] = ze \, D,$$

where the last equality groups the first and third terms inside the brackets as $(\varepsilon_e + \varepsilon_n)(\varepsilon_e - 1)$.

The sign of D. Since $F = (w/n) \Pi_w$ and $\Pi_w = 0$ at the optimum, $\partial F / \partial w = (w/n) \Pi_{ww} < 0$ there by the second-order condition. Hence $\hat{F}_w = w \, \partial F / \partial w < 0$, and since $ze > 0$, it follows that $D < 0$.

The wage response. Applying the implicit function theorem to $F(w^*(q), q) = 0$,

$$\eta = -\frac{\hat{F}_q}{\hat{F}_w} = -\frac{(ze / \varepsilon_h) \, (\varepsilon_e + \varepsilon_n - \sigma_h \, \varepsilon_e)}{ze \, D} = -\frac{\varepsilon_e + \varepsilon_n - \sigma_h \, \varepsilon_e}{\varepsilon_h \, D}. \qquad \square$$

Proof of Proposition 4. Along the optimal path, $\ln \nu = \ln w^* - \ln z - \ln e(w^*, q)$, so by Lemma 2

$$\frac{d \ln \nu}{d \ln q} = \eta - \left(\varepsilon_e \, \eta + \frac{1}{\varepsilon_h} \right) = (1 - \varepsilon_e) \, \eta - \frac{1}{\varepsilon_h},$$

which is the elasticity form of (3). Substituting η from Lemma 4 and collecting terms over the common denominator $\varepsilon_h D$,

$$\frac{d \ln \nu}{d \ln q} = -\frac{N}{\varepsilon_h \, D}, \qquad N \equiv (1 - \varepsilon_e) \, (\varepsilon_e + \varepsilon_n - \sigma_h \, \varepsilon_e) + D.$$

Expanding N using the definition of D in Lemma 4,

$$N = \underbrace{(1-\varepsilon_e)(\varepsilon_e+\varepsilon_n)+(\varepsilon_e+\varepsilon_n)(\varepsilon_e-1)}_{=0} - (1-\varepsilon_e)\,\sigma_h\,\varepsilon_e \;+\; \varepsilon_e\,(1-\mathcal{R}_u-\varepsilon_s) - \sigma_h\,\varepsilon_e^2.$$

The second term expands to $-\sigma_h\,\varepsilon_e+\sigma_h\,\varepsilon_e^2$, whose quadratic part cancels against the last term, leaving

$$N = \varepsilon_e\,(1-\sigma_h-\mathcal{R}_u-\varepsilon_s) = \varepsilon_e\,[(\varepsilon_h-\rho_h)-(\mathcal{R}_u+\varepsilon_s)]\,,$$

where the last equality uses $1-\sigma_h = \varepsilon_h-\rho_h$ from Lemma 3. Therefore

$$\frac{d\ln\nu}{d\ln q} = -\frac{\varepsilon_e}{\varepsilon_h\,D}\,[(\varepsilon_h-\rho_h)-(\mathcal{R}_u+\varepsilon_s)]\,.$$

Since $\varepsilon_e > 0$ and $\varepsilon_h > 0$, and $D < 0$ by Lemma 4, the factor $-\varepsilon_e/(\varepsilon_h D)$ is strictly positive, so $d\nu/dq$ has the sign of $(\varepsilon_h-\rho_h)-(\mathcal{R}_u+\varepsilon_s)$. In particular, $d\nu/dq < 0$ if and only if $\mathcal{R}_u(w^*)+\varepsilon_s(w^*) > \varepsilon_h(e^*)-\rho_h(e^*)$. □

V Construction of the Management Scores

The monitoring, targets, and incentives scores are taken from the management-practice survey of Bloom and Van Reenen (2007). The survey scores eighteen management practices, each from one (worst practice) to five (best practice), grouped into four areas: operations, monitoring, targets, and incentives. We use the latter three. Scores are assigned through open-ended, double-blind telephone interviews with plant managers: managers are not told that they are being scored, and interviewers are not told the firm's performance, so each score reflects the firm's actual practices rather than its aspirations. Each practice is standardized across firms, and our monitoring, targets, and incentives scores are the averages of the standardized practices in their respective areas.

Monitoring. Five practices capturing how the firm tracks performance and acts on it:

- Performance tracking: whether key performance indicators are collected, how frequently, and who sees them.
- Performance review: how those indicators are reviewed and communicated.
- Performance dialogue: the structure and quality of review meetings.
- Consequence management: what follows when agreed results or actions are not delivered.

- Performance clarity and comparability: whether workers know their own targets and can compare their performance with that of others.

Targets. Five practices capturing whether goals are well designed:

- Target breadth: whether goals are purely financial or also operational and broader.
- Target interconnection: how goals cascade and link across levels of the organization.
- Target time horizon: the balance between short- and long-run goals.
- Target stretch: how demanding, yet still attainable, the targets are.
- Developing talent: whether senior managers actively attract and develop talented people.

Incentives. Five practices capturing how the firm rewards and manages people:

- Rewarding high performance: how appraisals, bonuses, and nonfinancial rewards operate.
- Removing poor performers: how underperformance is addressed and how long it is tolerated.
- Promoting high performers: whether promotion rests on performance or on tenure.
- Attracting talent: what makes the firm distinctive to work for.
- Retaining talent: what the firm does to keep its star performers.

Scoring. Each practice is scored from one to five on the basis of open questions about actual practices and concrete examples, with the discussion continuing until the interviewer can make an accurate assessment. Performance tracking, for instance, scores one when tracking is ad hoc and does not reveal whether business objectives are being met, three when most indicators are tracked formally and overseen by senior management, and five when performance is tracked continuously and communicated to all staff. The eighteen practices are then standardized and averaged within each area to form the scores used in the text.

VI Additional Empirical Results

This appendix reports the coefficients on all three management-practice scores in the markdown regression of Table 2. Table A-1 includes the monitoring, incentives, and targets scores jointly in every specification, while adding the same controls across columns as in the main table. The pattern differs across management practices: monitoring enters negatively, consistent with wider markdowns; incentives is close to zero throughout; and targets enters positively, indicating a narrower markdown.

Table A-1: Markdowns and Other Management Practices

	(1)	(2)	(3)	(4)
Monitoring score	-0.15^{**}	-0.19^{**}	-0.19^{**}	-0.17^{*}
	(0.07)	(0.09)	(0.09)	(0.09)
Incentives score	0.02	-0.05	-0.02	0.01
	(0.11)	(0.12)	(0.12)	(0.12)
Targets score	0.13	0.20^{**}	0.19^{**}	0.16^{*}
	(0.08)	(0.08)	(0.08)	(0.09)
Industry, year, and country FE	✓	✓	✓	✓
Firm age FE		✓	✓	✓
Log employees			✓	✓
Competition FE				✓
N	2,774	2,770	2,770	2,770

Notes: The dependent variable is the log labor-to-materials cost ratio, ln(wage bill) − ln(materials), as in Table 2; a negative coefficient indicates a wider markdown. Management scores are standardized. Standard errors clustered at the firm level are reported in parentheses. Data are from Bloom and Van Reenen (2007). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.