3-4-2022

Monopsony in the U.S. Labor Market

Chen Yeh
Federal Reserve Bank of Richmond, chen.yeh@rich.frb.org

Claudia Macaluso
Federal Reserve Bank of Richmond, claudia.macaluso@rich.frb.org

Brad J. Hershbein
W.E Upjohn Institute for Employment Research, hershbein@upjohn.org

Upjohn Institute working paper ; 22-364

Citation

Yeh, Chen, Claudia Macaluso, and Brad J. Hershbein. 2022. "Monopsony in the U.S. Labor Market." Upjohn Institute Working Paper 22-364. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. https://doi.org/10.17848/wp22-364

Monopsony in the U.S. Labor Market

Authors

Chen Yeh, *Federal Reserve Bank of Richmond*
Claudia Macaluso, *Federal Reserve Bank of Richmond*
Brad J. Hershbein, *W.E Upjohn Institute for Employment Research*

Upjohn Author(s) ORCID Identifier

https://orcid.org/0000-0002-2534-8164

Published Version
In American Economic Review 112(7): 2099-2138

W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH
Researching the causes and consequences of unemployment

Monopsony in the U.S. Labor Market

Upjohn Institute Working Paper 22-364

Chen Yeh
FRB Richmond
chen.yeh@rich.frb.org

Claudia Macaluso
FRB Richmond
claudia.macaluso@rich.frb.org

Brad Hershbein
W.E. Upjohn Institute
hershbein@upjohn.org

March 2022

ABSTRACT

This paper quantifies the extent to which the U.S. manufacturing labor market is characterized by employer market power and how such market power has changed over time. We find that the vast majority of U.S. manufacturing plants operate in a monopsonistic environment and, at least since the early 2000s, the labor market in U.S. manufacturing has become more monopsonistic. To reach this conclusion, we exploit rich administrative data for U.S. manufacturers and estimate plant-level markdowns—the ratio between a plant's marginal revenue product of labor and its wage. In a competitive labor market, markdowns would be equal to unity. Instead, we find substantial deviations from perfect competition, as markdowns average 1.53. This result implies that a worker employed at the average manufacturing plant earns 65 cents on each dollar generated on the margin. To investigate long-term trends in employer market power, we propose a novel measure for the aggregate markdown that is consistent with aggregate wedges and also incorporates the local nature of labor markets. We find that the aggregate markdown decreased between the late 1970s and the early 2000s, but has been sharply increasing since.

JEL Classification Codes: E2, J2, J3, J42

Key Words: Monopsony, labor market power, markdowns, secular trends

Acknowledgments: We thank Adrien Auclert, Christian Bayer, Steven J. Davis, Jim Eeckhout, Grey Gordon, Lisa Kahn, Loukas Karabarbounis, Ayşegül Şahin, Pierre-Daniel Sarte, Felipe Schwartzman, Chad Syverson, Nicholas Trachter, and four anonymous referees, in addition to the audiences at many seminars and conferences, including SITE "Micro and Macro of Labor markets", NBER 2019 SI CRIW and MP groups, Barcelona Summer Forum EGF 2019, FRB Richmond "Market Structure and the Macroeconomy", University of Oslo-CEPR "Micro-consistent Macro", and many others. This research was conducted while Claudia Macaluso and Chen Yeh were Special Sworn Status researchers at the U.S. Census Bureau. We also thank the hospitality of the University of Minnesota, and the Chicago, Minnesota, and Federal Reserve Board Federal Research Centers, where part of the research was conducted. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the Federal Reserve Bank of Richmond, or the Federal Reserve System. All results have been reviewed to ensure that no confidential information is disclosed. First version: May 15, 2018.

1 Introduction

Is the U.S. labor market perfectly competitive? In perfectly competitive labor markets, marginal revenue products of labor are equal to workers' wages, meaning that every dollar generated on the margin is paid to workers. Although it's a convenient modeling assumption, does this benchmark accurately describe the U.S. labor market? Wedges between marginal revenue products of labor and wages may constitute evidence of monopsony and suggest a departure from allocative efficiency. In this paper, we provide estimates of these wedges—"markdowns"— across U.S. manufacturing plants between 1976 and 2014. Specifically, we show that (i) the U.S. manufacturing labor market is characterized by significant markdowns, consistent with employer market power, and (ii) the degree of this market power decreased between the late 1970s and the early 2000s but increased sharply afterwards.

Quantifying employers' market power and understanding that market power's dynamics across employers and over time is fundamental to devising appropriate policy responses. Reliable evidence on employer market power is particularly relevant when evaluating policies that directly affect workers' compensation and mobility, such as changes in the minimum wage. Similarly, when assessing regulatory limits on the growth of large firms, it is helpful to consider the extent to which such firms are able to compensate labor below their marginal revenue products. Policymakers have recently considered enacting these policies to mitigate a perceived increase in employers' market power (cfr. FTC, 2018).[1] While this rise in employer market power can be plausibly connected to several labor market trends, measures of employer market power that directly compare the wedge between the marginal revenue product of labor and the wage are not available to inform the current policy debate.[2]

Our paper responds precisely to this gap. We estimate plant-level markdowns for the whole

[1] See the Federal Trade Commission Hearing #3: Multi-Sided Platforms, Labor Markets and Potential Competition, on October 15–17, 2018.

[2] Several complementary measures have been proposed, including those related to labor market concentration (Azar, Marinescu and Steinbaum, 2020a; Azar et al., 2020b; Benmelech, Bergman and Kim, 2020; Rinz, 2020; Schubert et al., 2021), as well as fully structural approaches (Posner, Weyl and Naidu, 2018; Azar, Berry and Marinescu, 2019b; Jarosch, Nimczik and Sorkin, 2021; Berger, Herkenhoff and Mongey, Forthcoming). The measure we develop is based on the production function approach and is unique in that it quantifies, with minimal assumptions, plant-level wedges between the marginal revenue product and the wage.

U.S. manufacturing sector and study their relationship with employer size, age, and productivity. As well, we look at the evolution of aggregate markdowns over time.

Our analysis of labor market monopsony starts with estimating and characterizing the distribution of plant-level markdowns. In our baseline framework, firms internalize a finitely elastic labor supply curve and thus operate in a monopsonistic environment. Without imposing further restrictions on the labor supply curve, we interpret gaps between the output elasticity of labor and labor's revenue share as market power, jointly in output (product markups) and input (labor markdowns). Under the assumption that at least one other observable input is flexible—that is, free of adjustment costs and monopsony power—we show that markups and markdowns can be identified and estimated separately. The key insight is that the wedge for the flexible input reflects only product markups, and so the *ratio* of the labor wedge and the wedge for the flexible input permits identification of both markups and labor markdowns. To implement this insight empirically, we adapt the production function approach from the industrial organization (IO) literature.

This approach has several advantages. First, we can remain agnostic about the sources of employer market power. In fact, we show that our approach is consistent with a broad range of monopsony models. Second, although we do need to impose a functional form for a firm's production, we can be highly flexible by specifying a translog function—a second-order approximation to *any* arbitrary, differentiable production function. A third benefit is that the production function approach remains valid regardless of the assumptions made on other inputs besides labor and materials (and thus can accommodate capital adjustment costs). Finally, the approach readily permits several extensions and modifications, including heterogeneous labor within plants, labor adjustment costs, ex-ante specified returns to scale, and alternative measures of labor compensation, such as inclusion of benefits.

Estimating such production functions and markdowns with comprehensive administrative data for the U.S. manufacturing sector (using the Census and Annual Surveys of Manufactures), we find that that labor markets in U.S. manufacturing are far from perfectly competitive. The average plant's marginal revenue product of labor is 53 percent higher than its wage, implying that a worker employed there receives about 65 cents on the marginal dollar. Furthermore, we document a substantial amount of dispersion across plants even within 3-digit NAICS industries, with an average within-industry interquartile range of 61.6 percent. Investigating the sources of heterogeneity in markdowns, we find a robust

positive association between markdowns and size, whether measured as an establishment's (or firm's) relative share of employment or in terms of industrial and geographical scope. This result supports the hypothesis that employer size matters when assessing the welfare implications of labor market power. On the other hand, we find that plant-level dispersion in markups and in productivity account for little of the heterogeneity in markdowns. We conclude that the typical U.S. manufacturing plant operates in a monopsonistic environment, and that a significant degree of variation persists within narrowly defined industries.

We next use our estimates of micro-level markdowns to describe trends in macro-level markdowns since 1977. This is not straightforward, as there is no uncontested framework that delivers a clear aggregation rule for markdowns. We propose a novel "aggregate markdown" measure that satisfies two requirements. First, aggregate markdowns and markups reflect *aggregate* wedges, the gaps that a fictional representative firm would face. This interpretation has the advantage that no specific market structure for labor or output needs to be imposed for aggregation. Second, aggregate markdowns need to account for the local nature of labor markets, consistent with evidence on the cost of distance during job search. In the end, we show that aggregation occurs through sales-weighted harmonic averages, where weights are adjusted for heterogeneity in output elasticities. This measure of the aggregate markdown displays a U-shaped evolution over time, decreasing between 1977 and 2002, and sharply increasing afterwards. We thus find support for the hypothesis that monopsony in the U.S. manufacturing labor market has increased since the early 2000s.

Finally, we relate our aggregate markdown estimates to measures of labor market concentration that have been commonly used in recent studies, as described below, on account of their simplicity of construction. Despite our finding that *plant-level* markdowns increase with employment size, we find little correlation between concentration and markdowns at the aggregate *market* level, a consequence of accounting for heterogeneity in output elasticities across firms within a market and our aggregation rule. Furthermore, although aggregate local concentration and our aggregate markdowns show qualitatively similar declines in the late 20th century, the concentration measure does not show the sharp reversal of markdowns since the early 2000s. We conclude that cross-sectional and time variation in local employment concentration do not necessarily reflect variation in employer market power as measured by markdowns—at least within manufacturing.

CONTRIBUTION TO THE LITERATURE. Our paper contributes to a recently reinvigorated research agenda on the prevalence and evolution of labor market monopsony in the U.S. economy. This interest in the exercise of market power by firms, and especially large firms, has been motivated heavily by the secular decline in labor's share of income (Elsby, Hobijn and Şahin, 2013; Karabarbounis and Neiman, 2013), which has in turn been linked to changes in industry-level sales concentration, with "superstar" firms potentially having higher product markups and lower labor shares(Autor et al., 2020).[3]

Our contribution is twofold. First, we use comprehensive administrative data for U.S. manufacturers and provide direct estimates of the wedge between an employer's marginal revenue product of labor and its wage. In so doing, we document substantial dispersion in plant-level markdowns and document how this heterogeneity varies with employer characteristics, such as size, age, productivity, and geographic scope. Second, we develop a new, theory-grounded way to characterize aggregate markdowns and document their evolution over the past four decades.

Our markdown estimation procedure relies on the "production approach" (De Loecker, 2011), which combines insights from Hall (1988) with production function estimation techniques from the IO literature (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; De Loecker and Warzynski, 2012; Ackerberg, Caves and Frazer, 2015). In our estimation procedure, we explicitly identify markups and markdowns separately. As a result, we do not confound these two sources of market power. Most previous studies tend to focus on only one source of market power instead, and thus possibly overstate the extent of that source's market power. Exceptions to this practice, however, include Dobbelaere and Mairesse (2013) and Morlacco (2020), who also exploit the flexibility of material inputs to study monopsony in, respectively, non-U.S. labor markets and the market for foreign intermediate inputs. Brooks et al. (2021b) also use techniques analogous to those in this paper to estimate markdowns in China and India, and conclude that "in the context of developing economies, markdowns substantially lower the labor share."[4]

A related literature has documented a contemporaneous increase in markups, arguing that

[3]With the exception of the lowest-productivity establishments, which we discuss further below, we find a positive relationship between plant-level productivity and markdowns. This finding is consistent with the thesis in Autor et al. (2020), as long as labor shares fall faster than product markups rise.

[4]In a companion paper, Brooks et al. (2021a) further show that highway construction in India offset markdowns and increased the labor share among nearby firms.

the latter could be a unifying explanation behind many observed secular trends in the U.S. economy, including the decrease in the labor share (Eggertsson, Robbins and Wold, 2021; De Loecker, Eeckhout and Unger, 2020). Our paper contributes related evidence on the dynamics of the labor share and wages at the micro-level. Specifically, we document substantial variation in plant-level markdowns for the manufacturing sector, both across and within narrowly defined industries, and illustrate a tight positive relationship between markdowns and size.

In our baseline measure, we assume that firms take monopsony forces into account by internalizing a finitely elastic labor supply curve, thus reflecting the assumption of an upward-sloping labor supply curve common in many of the current models of monopsony. This includes frameworks based on Burdett and Mortensen (1998), as in Bontemps, Robin and Van den Berg (2001), Manning (2003), Mortensen (2003), Manning (2011), and Webber (2015). It also encompasses the class of additive random utility models as characterized in Chan, Kroft and Mourifie (2019), which include Card et al. (2018) and Lamadon, Mogstad and Setzler (2022), and environments based on monopsonistic competition, as in Bhaskar and To (1999), Staiger, Spetz and Phibbs (2010), and Berger, Herkenhoff and Mongey (Forthcoming). Our paper contributes to this literature by proposing a strategy to estimate markdowns that, while compatible with many of the frameworks studied previously, is not tightly linked to a specific micro-foundation but instead is quite general.[5]

Finally, our paper relates to the burgeoning literature on labor market concentration, as we compare concentration indices to markdowns. Interest in concentration indices stems from their ease and breadth of use in both academic research and the practice of antitrust in the U.S. economy. These have been calculated at the national (Autor et al., 2020) and local levels (Rossi-Hansberg, Sarte and Trachter, 2020), and show diverging long-run trends.[6] Recent work by Azar, Marinescu and Steinbaum (2020a) and Azar et al. (2020b) shows the negative association between concentration and wages using vacancy data from online

[5]Standard arguments dating back to Robinson (1933) imply that markdowns are one-to-one with labor supply elasticities. As a result, our markdown estimates also speak to the literature evaluating the elasticity of labor supply. Our implied average elasticity estimate of 1.88 is only slightly above the median elasticity estimate from more than 800 research papers covered in the meta-study by Sokolova and Sorensen (2020).

[6]A recent paper by Benmelech, Bergman and Kim (2020) also computes employment concentration and relates it to average wages in U.S. manufacturing. Lipsius (2018) and Rinz (2018) both provide estimates of concentration in firm-level employment from the Longitudinal Business Database and conclude that, though local concentration reduces earnings and increases inequality, observed changes in concentration are unable to explain the rise in income inequality observed in the U.S. economy.

sources and argues for extending antitrust considerations to mergers that affect labor market concentration. Despite this popular usage, however, it is unclear from a theoretical standpoint whether a market's labor concentration is necessarily positively correlated with its level of competitiveness in the markdown sense (Syverson, 2019). Our paper contributes to this conversation by documenting that the correlation between markdowns and employment concentration is quite modest, both cross-sectionally (across local labor markets) and in the aggregate over time. We view this result as highlighting the challenges posed by aggregation when comparing micro-founded measures of employer market power, such as markdowns, to reduced-form indices, such as employment concentration.

OVERVIEW OF THIS PAPER. Section 2 lays out our estimation procedure and describes the data. Section 3 illustrates our markdown estimates and discusses heterogeneity. Section 4 proposes a novel measure for aggregate markdowns and shows that the time trend in aggregate markdowns is U-shaped, with a minimum in the early 2000s. It concludes with documenting a weak relationship between our estimated aggregate markdown and an index of local employment concentration. In section 5, we discuss the robustness of our baseline results. Section 6 summarizes the evidence and concludes. We provide several additional results, derivations, and robustness tests in the Appendix and Online Appendix.

2 Markdown estimation

Our analysis of monopsony in the U.S. labor market is based on markdowns, the percentage gap between a plant's marginal revenue product of labor (MRPL) and the wage it pays its workers. This is a direct measure of employer market power that is easy to compare to the benchmark of perfect competition. In a perfectly competitive labor market, markdowns would be equal to unity. When markdowns are larger than unity, however, the employer compensates workers less than dollar-for-dollar for every unit of revenue generated at the margin.

In this section, we describe our basic framework. We begin by using the optimality conditions from a firm's profit maximization problem to show a one-to-one relationship between markdowns and firm-level labor supply elasticity. We then use the firm's dual problem (through cost minimization) to derive an estimation strategy for markdowns in the spirit of Hall (1988) and De Loecker and Warzynski (2012). Using this strategy and detailed

administrative data on plants' output and inputs, we retrieve micro-level markdowns in the U.S. manufacturing sector. Our approach simultaneously allows for positive product markups.

2.1 Obtaining markdowns through duality

2.1.1 Profit maximization

Our notion of an individual employer's monopsony power is rooted in the idea that a monopsonistic employer can compensate its workers below their marginal revenue product of labor; a definition of monopsony power popularized by Manning (2003). We refer to this percentage gap as a firm's markdown. In the following, we will show that a firm's markdown has a one-to-one relationship with its (perceived) labor supply elasticity.[7] To see this, consider a firm's profit maximization problem:

$$\max_{\ell \geq 0} R(\ell) - w(\ell)\ell$$

where $R(\ell) \equiv \text{rev}(\ell; \mathbf{X}^*_{-\ell}(\ell))$ is shorthand notation for revenues in which all inputs are evaluated at their optimum with the exception of labor ℓ. For ease of notation, we drop the firm's index for the moment. Given this structure and assuming that the revenue function and wage schedule are differentiable, a firm's optimality condition can be rearranged as:

$$\begin{aligned} R'(\ell^*) &= \left[\frac{w'(\ell^*)\ell^*}{w(\ell^*)} + 1\right] w(\ell^*) \\ &= \left[\varepsilon_S^{-1} + 1\right] \cdot w(\ell^*) \end{aligned} \tag{1}$$

where the firm's perceived (inverse) elasticity of labor supply is defined as: $\varepsilon_S^{-1} \equiv \frac{w'(\ell)\ell}{w(\ell)}\big|_{\ell=\ell^*}$. Therefore, it is sufficient to characterize a firm's labor supply elasticity in order to retrieve its markdown. Hence, we get:

$$\nu \equiv \frac{R'(\ell^*)}{w(\ell^*)} = \varepsilon_S^{-1} + 1, \tag{2}$$

so that the markdown, ν, is expressed as the ratio of the MRPL to the wage, or the inverse

[7] This parallels the intuition behind the Lerner index formula, which relates residual demand elasticities with price-cost markups.

labor supply elasticity plus one.[8] In this conceptual framework, we do not take a specific stance on the sources of monopsony power; the only requirement is finite, firm-specific labor supply elasticities. In Online Appendix O.7, we show that our setup is quite general and nests a variety of monopsony frameworks, including wage-posting models (e.g., Burdett and Mortensen, 1998), additive random utility models (e.g., Card et al., 2018; Chan, Kroft and Mourifie, 2019; and Lamadon, Mogstad and Setzler, 2022), and monopsonistic competition models (e.g., Bhaskar and To, 1999; Staiger, Spetz and Phibbs, 2010; and Berger, Herkenhoff and Mongey, Forthcoming).

2.1.2 Cost minimization

A complication, however, is that estimating a firm's perceived elasticity of labor supply in a general setting is challenging, in part because of the potential for firm market power over both inputs (monopsony) and output (monopoly). In this section, we propose a "production approach" to retrieve markdowns for U.S. manufacturers in a general setting, building on insights from Hall (1988), De Loecker (2011), and De Loecker and Warzynski (2012). The key insight is that wedges between output elasticities and revenue shares can reflect market power in both input and output markets. Intuitively, the output elasticity of labor captures the gain from an additional unit of labor, whereas labor's share of revenue reflects its cost (normalized by a firm's total revenue). If this wedge is larger than unity, the marginal gain is larger than its costs, and the firm must be capturing margins through either markups on its output or markdowns on its inputs.

The production approach starts with a firm's optimal input choices. Suppose there are $K > 1$ inputs, denoted by $\mathbf{X}_{it} = \left(X_{it}^1, \ldots, X_{it}^K\right)'$. These inputs have pricing schedules $\{V_{it}^k\}_{k=1}^K$, and adjustment costs for some input k are captured by the function $\Phi_t^k(X_{it}^k, X_{it-1}^k)$. Also, we denote a firm i's productivity level at time t by ω_{it}. Then, to derive markdowns, we adopt the following set of assumptions:

ASSUMPTION I. A firm engages in cost minimization.

There exists at least one input k' that satisfies the following:

ASSUMPTION II. There are no adjustment costs for input k', i.e. $\Phi_t^{k'}(\cdot, \cdot) = 0$.

ASSUMPTION III. Input k' is not subject to monopsony forces, i.e. $V_{it}^{k'}(X_{it}^{k'}) = V_{it}^{k'}$.

[8] Some studies define the markdown as the inverse of our measure, since it reflects the extent to which wages are marked down. Under this convention, markdowns below unity reflect labor market power, whereas in our measure, markdowns *above* unity reflect labor monopsony.

ASSUMPTION IV. Input k' is chosen statically.

ASSUMPTION V. Production $F(\cdot;\omega_{it})$ is twice continuously differentiable in $X_{it}^{k'}$ and it satisfies:

$$\lim_{X_{it}^{k'}\to 0}\frac{\partial F(\mathbf{X}_{it};\omega_{it})}{\partial X_{it}^{k'}}=+\infty \text{ and } \lim_{X_{it}^{k'}\to +\infty}\frac{\partial F(\mathbf{X}_{it};\omega_{it})}{\partial X_{it}^{k'}}=0.$$

for any $\omega_{it}\in\mathbb{R}_+$. Furthermore, the demand schedule $P_{it}(\cdot)$ is continuously differentiable and strictly decreasing.

ASSUMPTION VI. Input k' is used for the production of output only.

Any input k' that satisfies assumptions **II**–**VI** simultaneously, will be referred to as a **flexible** input. Assumptions **I** and **V** are regularity assumptions and ensure that (a subset of) inputs can be characterized through their first-order conditions alone. Assumption **VI** is relatively weak and requires the flexible input to be used solely for production purposes.[9] Assumption **IV** implies that an input k' cannot directly affect a firm's future outcomes, ruling out certain dynamic narratives.[10]

Hence, the challenge becomes to find an input that simultaneously satisfies assumptions **II** and **III**: the existence of a static input k' free of adjustment costs and for which firms are price-takers. If such an input k' exists, we can establish the following result:

PROPOSITION 1. Let assumption **I** hold. If assumptions **II**–**VI** hold for some input k' other than labor, we can characterize a firm's product markup with the gap of its flexible input k': $\mu_{it}=\frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}}$, where $\theta_{it}^{k'}$ and $\alpha_{it}^{k'}$ denote a firm's output elasticity of input k' and its share of revenue, respectively. If assumptions **II** and **IV**–**VI** *also* apply to labor ℓ, and firm i faces a differentiable, finitely elastic wage schedule, then its markdown, ν_{it}, satisfies:

$$\nu_{it}=\frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}}\cdot\mu_{it}^{-1} \tag{3}$$

where θ_{it}^{ℓ} and α_{it}^{ℓ} denote a firm's output elasticity of labor and its labor share of revenue,

[9] This rules out, for example, inputs designed purely to increase the demand for output, such as marketing. Because our data are at the establishment level and allow us to separate production from nonproduction labor, this assumption is relatively innocuous, and we discuss it in detail in Section 5.

[10] Our setup allows for (dynamic) capital adjustment costs as long as there is a flexible input other than capital. We do rule out, however, mechanisms in which current output (which depends on current inputs) affects a firm's future demand, such as by affecting the customer base (see, for example, Foster, Haltiwanger and Syverson, 2016).

respectively.

Proof. See Appendix A.1. □

Proposition 1 implies that the ratio between the output elasticity of labor and labor's share of revenue equals the product of the markdown and the markup. In the remainder of this paper, we follow the IO literature and assume that conditions **II**–**VI** hold for material inputs (therefore referred to as "the flexible input" and indexed by M). The availability of a flexible input is the key factor that allows us to distinguish between markdowns and markups, isolating our measure for labor market power from market power for outputs.

Thus, a key question for our identification is whether materials lack adjustment costs and monopsony power in our context. Several pieces of evidence suggest these are reasonable assumptions. First, the Census Bureau's definitions of material inputs includes largely generic, primary goods, as well as contract services, which tend to be traded on open, often global markets.[11] Second, Atalay (2014) does not find that prices for material inputs vary with quantities (as required by Assumption **III**).[12] Third, even if material inputs were subject to monopsony forces, our estimates would reflect markdowns for labor *relative* to markdowns for material inputs, implying our estimates would be lower bounds for labor market power.[13]

For labor markdown estimation, Proposition 1 also requires that Assumptions **II** and **IV**–**VI** apply to *labor*. For Assumptions **II** and **IV** to hold, we need to rule out labor adjustment costs, including non-spot-market contracts.[14] We address the sensitivity of our results to possible labor adjustment costs in Appendix C and find that these matter relatively little, for both convex and nonconvex adjustment costs.

Finally, our setup also excludes labor being used for any purpose other than the production of output (Assumption **VI**). This could occur, for example, when labor is used for market-

[11] See Table XVI in Online Appendix O.5 for a complete list.

[12] Atalay (2014) also finds that some U.S. manufacturing plants pay "low materials prices because their suppliers are exceptionally productive." Hence, the variation in material input prices across plants can be partially explained by variation in the marginal costs of suppliers, rather than plants exploiting their monopsonistic power.

[13] We discuss these issues in more detail in Section 5.

[14] Adjustment costs in labor would occur, for instance, in the presence of hiring and firing costs, such as a corporate tax schedule that varies with firm size, or legal requirements that limit employment at will. However, these provisions are not especially binding in the U.S. labor market.

ing or hiring. This assumption may seem strong, but it is less restrictive in our context of U.S. manufacturing plants, where most workers are indeed production workers. Furthermore, we show in Section 5 that our results are not affected when we explicitly distinguish production from nonproduction labor.

2.2 Production function estimation

A distinct advantage of the production approach is its generality. We do not need to make any assumptions on the sources of market power in order to quantify markdowns. In particular, we do not take a stance on the market structure for labor or the form of labor supply curves that firms face—a distinguishing feature of this paper from the fully developed structural models of Card et al. (2018), Chan, Kroft and Mourifie (2019), and Lamadon, Mogstad and Setzler (2022), and Berger, Herkenhoff and Mongey (Forthcoming). Furthermore, we need make no additional assumptions for inputs besides materials and labor: our approach is valid as long as firms are subject to some finitely elastic labor supply curve and material inputs are flexible. Finally, we can explicitly distinguish between market power in output and labor markets. Our result in Proposition 1 implies that observing output elasticities and revenue shares is sufficient for constructing markdowns. Revenue shares are directly observable in administrative data on U.S. manufacturing plants, but output elasticities need to be estimated.

To do so, we estimate production functions through "proxy variable" methods (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; De Loecker and Warzynski, 2012; Ackerberg, Caves and Frazer, 2015). We adopt standard assumptions from the proxy-variable literature, particularly Ackerberg, Caves and Frazer (2015), which we discuss below.

ASSUMPTION 1. A firm i's information set at time t, $\mathcal{F}_{it}$, is generated by $\{\omega_{i\tau}\}_{\tau=0}^{t}$. The transitory shock ε_{it} is not observed by the firm and satisfies $\mathbb{E}\left[\varepsilon_{it}|\mathcal{F}_{it}\right] \equiv \mathbb{E}_t(\varepsilon_{it}) = 0$.

ASSUMPTION 2. A firm i's state variables at time t are given by the pair (k_{it}, ω_{it}). Furthermore, its stock of capital accumulates as a function of lagged capital k_{it-1} and investment ι_{it-1}:

$$k_{it} = \kappa(k_{it-1}, \iota_{it-1}) \tag{4}$$

ASSUMPTION 3. The technology parameters $\boldsymbol{\beta}$ are constant across time and common

within an industry group. Productivity evolves according to a first-order Markov process:

$$p(\omega_{it+1}|\mathcal{F}_{it}) = p(\omega_{it+1}|\omega_{it}) \tag{5}$$

where the distribution $p(\cdot|\mathcal{F}_{it})$ is known to firms and is stochastically increasing in ω_{it}.

ASSUMPTION 4. For at least one flexible input, the only unobservable factor (from the econometrician's point of view) in a firm's input demand function is productivity ω_{it}, which is a scalar and is Hicks-neutral.

ASSUMPTION 5. For any flexible input satisfying assumption 4, a firm's input demand function is strictly monotone in ω_{it}.

Assumptions 1 and 2 are standard in the literature and encompass a rich set of frameworks of firm behavior. Because the econometrician does not observe firm-level productivity, a least squares regression of output on inputs would lead to biased estimates (the "transmission bias" in Griliches and Mairesse, 1998). To address this problem, Assumption 3 places some general structure on idiosyncratic productivity by having it follow a first-order Markov process. [15]

Assumption 4, also known as the "scalar unobservable" assumption, requires that idiosyncratic productivity be the only input demand factor unobserved by the econometrician. Bond et al. (2021) argue that this assumption is not consistent within the context of market power, and that other estimators that do not rely on it should be used instead (e.g., Blundell and Bond, 2000). However, in Online Appendix O.5, we run a set of Monte Carlo simulations and demonstrate that our empirical approach, relying on the proxy variable estimator of De Loecker and Warzynski (2012), still produces estimates with less bias than do estimators that do not rely on the scalar unobservable assumption. Consequently, we do not view Assumption 4 as restrictive in practice.

Finally, we require that flexible input demand functions are invertible in productivity. This assumption allows us to obtain a nonparametric estimate of output without observing productivity. While implicitly ruling out some production functions, this assumption still allows for a general specification, such as the translog production function that we adopt in

[15] Although not without loss of generality, this structure is the most general in the production function estimation literature (Ackerberg, 2020) and allows for considerably greater flexibility than other common persistent stochastic processes, such as an AR(1).

our baseline estimates.[16] Its flexibility rests on its interpretability as a second-order approximation to *any* arbitrary, differentiable production function (see De Loecker and Warzynski, 2012). Hence, a translog specification nests and is substantially more general than, for example, a Cobb-Douglas specification.[17]

In the following, we briefly describe the mechanics of the proxy-variable methodology and how we obtain output elasticities. We refer the reader interested in a more detailed treatment of our estimation procedure to Appendix A.2. In section 5, we further discuss possible challenges to the proxy-variable methodology and how they relate to our results.

We denote y_{it} as log output and $\mathbf{x}_{it}$ as the vector of log inputs. This vector of inputs contains the first-, cross-, and second-order terms of the vector $\tilde{\mathbf{x}}_{it} = (k_{it}, \ell_{it}, m_{it}, e_{it})'$, consisting of capital, labor, materials, and energy. Because of the unobserved productivity parameter, we require instruments for the input vector to recover consistent production parameters. Let $\mathbf{z}_{it}$ be the vector instrumenting for the set of endogenous inputs $\mathbf{x}_{it}$. Following De Loecker and Warzynski (2012), we construct $\mathbf{z}_{it}$ by taking the lag of each input in $\mathbf{x}_{it}$ with the exception of capital. Last, let $f(\mathbf{x}_{it}; \boldsymbol{\beta})$ denote the log transformation of the production function. In the end, our goal is to estimate production function parameters $\boldsymbol{\beta}$ in the following setting:

$$y_{it} = f(\mathbf{x}_{it}; \boldsymbol{\beta}) + \omega_{it} + \varepsilon_{it} \tag{6}$$

where ε_{it} reflects measurement error.[18] The loglinearity of output in productivity comes from the second part of Assumption 4. Given Assumptions 1–5, we estimate production function parameters $\boldsymbol{\beta} \in \mathbb{R}^Z$ for each industry-specific production function in a three-step process:

[16] This is important because the production approach does require a functional form on a firm's production function. For our estimates to have some external validity, it is therefore desirable to adopt a production structure that is as general as possible. We believe this is achieved through a translog specification.

[17] Assumption 2 allows production parameters to vary across detailed industry groups (i.e., three-digit NAICS) but imposes that they are constant over time. However, in our context with translog production, this does not imply that *output elasticities* are constant over time. Indeed, under a translog specification for gross output, output elasticities are allowed to vary across plants with the (time-varying) level of each firm's inputs. Furthermore, explicitly allowing time-varying production parameters does not greatly alter our conclusions. As a result, this part of Assumption 2 is without much loss of generality.

[18] Other interpretations for ε_{it} are possible (Bond et al., 2021). We follow De Loecker and Warzynski (2012) and interpret ε_{it} as measurement error. Its exact interpretation is not important for our results as long as ε_{it} is unobserved by the firm and the econometrician.

1. Run a third-order polynomial regression of y_{it} on $\tilde{\mathbf{x}}_{it}$ and a set of controls.[19] Obtain nonparametric estimates of log output φ_{it} free of measurement error.

2. Construct an estimate of productivity as $\omega_{it}(\tilde{\boldsymbol{\beta}}) = \varphi_{it} - f(\mathbf{x}_{it}; \tilde{\boldsymbol{\beta}})$ and run a third-order polynomial regression of $\omega_{it}(\tilde{\boldsymbol{\beta}})$ on $\omega_{it-1}(\tilde{\boldsymbol{\beta}})$ to obtain estimates of productivity shocks $\xi_{it}(\tilde{\boldsymbol{\beta}})$.

3. Obtain estimates $\hat{\boldsymbol{\beta}}$ of the production function parameters $\boldsymbol{\beta}$ through the GMM system induced by the moment conditions $\mathbb{E}\left(\xi_{it}(\tilde{\boldsymbol{\beta}}) \cdot \mathbf{z}_{it}\right) = \mathbf{0}_{Z\times 1}$.

Once estimates of $\boldsymbol{\beta}$ are obtained, it is a straightforward matter to calculate output elasticities. Under a Cobb-Douglas specification, for example, the parameters $\boldsymbol{\beta}$ are equal to output elasticities. However, under our translog setup, output elasticities are a linear function of the inputs in $\tilde{\mathbf{x}}_{it}$, with coefficients that depend on $\boldsymbol{\beta}$. A complete description on the construction of output elasticities under translog production is found in Appendix A.2.

A crucial part of the proxy variable methodology is to obtain transitory shocks to firm-level productivity. As a result, we need to separately identify productivity ω_{it} and measurement error ε_{it}. Under Assumptions 4 and 5, productivity can be written as a function of observables only. This allows us to identify ε_{it} in the first step.[20] Our translog structure then allows us to obtain firm-level productivity in step two. Finally, we are able to identify transitory shocks ξ_{it} to productivity through the Markov property in Assumption 3. These shocks are the key behind our moment conditions: current inputs are orthogonal to future shocks in productivity through Assumption 2.

INTUITION BEHIND IDENTIFICATION. The econometric literature on production function estimation has not provided formal arguments on whether proxy variable estimators produce consistent estimates. However, informal arguments for identification can be given through the logic of an IV estimator. As demonstrated in step three of our estimation procedure, we construct our moment conditions through the instrument vector $\mathbf{z}_{it}$. Therefore,

[19] Our baseline estimates include a set of year fixed effects, but our results do not change by much when other controls, such as size and age, are included.

[20] By Assumptions 1, 2, and 4, we can write $m_{it} = m_t(\omega_{it}; k_{it})$. Whenever Assumption 5 also holds, m_t is invertible in productivity and there exists some function $\omega_{it} = h_t(m_{it}; k_{it})$. Thus, we have $y_{it} = f(\mathbf{x}_{it}; \boldsymbol{\beta}) + h_t(m_{it}; k_{it}) + \varepsilon_{it} \equiv \phi(\mathbf{x}_{it}; \boldsymbol{\gamma}) + \varepsilon_{it}$. Hence, we can obtain estimates for output net of measurement error by running a nonparametric regression (e.g., a high-order polynomial) in only observables.

we are "instrumenting" the endogenous input vector $\mathbf{x}_{it}$ with $\mathbf{z}_{it}$. To understand why the above system can retrieve valid estimates of $\boldsymbol{\beta}$, we can verify a set of exogeneity and rank conditions.

Exogeneity implies that $\mathbf{z}_{it}$ is orthogonal to the innovations to productivity in period t. By Assumption 2, firms choose inputs k_{it}, ℓ_{it-1}, m_{it-1}, and e_{it-1} without knowing what the productivity shock ξ_{it} will be. As a result, exogeneity holds through our timing assumptions. However, how do we ensure that the "instruments" are valid, that those moment conditions associated with ℓ_{it-1}, m_{it-1}, and e_{it-1} are actually informative for the production function coefficients on inputs ℓ_{it}, m_{it}, and e_{it}? For this to follow, we need factor-input demand to evolve relatively smoothly, ruling out, for example, production functions that reflect perfect substitutes or perfect complements. We also need price schedules to similarly evolve smoothly and be somewhat—but not perfectly—persistent over time. We believe these are plausible requirements; Atalay (2014), for instance, finds empirical support for the partial persistence of materials prices.

For consistent estimates, we further require reasonably long panel data (Pesaran and Smith, 1995; Gandhi, Navarro and Rivers, 2020; Bond et al., 2021). In particular, Gandhi, Navarro and Rivers (2020) formalize that time-series variation in the prices for material inputs—our flexible input—is critical for identification, as it is the *only* residual source of variation that can identify $\boldsymbol{\beta}$ under the proxy variable methodology.[21] As our data, described below, span almost 40 years, we believe there is sufficient variation in materials prices, whether in aggregate or industry-specific, to identify production-function parameters.[22]

[21] As Flynn, Gandhi and Traina (2019) point out, it is important to note that this input price variation should be orthogonal to productivity and output prices. As a result, some forms of unobserved heterogeneity in inputs can be problematic. For example, it cannot reflect differences in the quality of purchased inputs. While it is difficult to verify these assumptions explicitly in the absence of input quantity data, we did verify that the overwhelming majority of the variation in material input deflators from the NBER-CES Manufacturing Database comes from the time series.

[22] We should note that our measures of output and inputs are based on deflated expenditures. While we show that markdowns can be obtained even if only (deflated) revenue elasticities can be estimated (see Section 5), the absence of input prices does technically violate the scalar unobservable assumption (see Hu, Huang and Sasaki, 2020; Bond et al., 2021). However, this issue does not seem to be a major concern in practice, as we summarize in the "Scalar unobservable assumption" subsection of Section 5. Using Monte Carlo methods, we show more fully in Online Appendix O.5.3 that our preferred proxy variable estimator appears more robust than other estimators that do not rely on the scalar unobservable assumption.

2.3 Data: Censuses and Annual Surveys of Manufactures

We use two administrative data sets for the estimation of markdowns: the Census of Manufactures (CM) and the Annual Survey of Manufactures (ASM), both from the U.S. Census Bureau. The Census of Manufactures is a quinquennial survey that covers the universe of manufacturing establishments in years ending in "2" and "7." Crucially, the CM contains establishment-level data on revenues and inputs, the two necessary ingredients for production-function estimation. We construct our measures of output (revenues) and inputs (capital, labor, materials, and energy) using deflators from the NBER-CES Manufacturing Database, following the standard procedures described in Syverson (2004a) and Kehrig (2015).

To construct markdowns for non-census years, we use the Annual Survey of Manufactures (ASM). The ASM contains a representative, rotating sample of manufacturing plants. While large plants are sampled with near certainty, small plants are sampled less frequently based on their size.[23] We use provided sampling weights to ensure that our estimates are representative of the whole manufacturing sector. Our main results are thus based on a nonbalanced panel for manufacturing plants in years 1976–2014. To avoid artificial spikes in census years, we keep only those plants that are in the rotating sample of the ASM in these years.

3 Markdowns in U.S. manufacturing

3.1 Cross-sectional distribution

We present results of our estimation procedure in Table I. These paint a clear picture: markdowns are sizable and considerably larger than unity. The average establishment throughout the period charges a markdown of 1.53—that is, a plant's marginal revenue product of labor is, on average, 53 percent higher than the wage it pays its workers. Alternatively, taking the reciprocal, a markdown of 1.53 implies that a worker receives about 65 cents on the marginal dollar generated. Furthermore, we find that labor market power is widespread across manufacturing plants. Half charge a markdown of at least 1.364 (73 cents on the dollar), and the interquartile range in markdowns exceeds 0.6. Although these markdown estimates may seem large, they are largely in line with implied estimates from previous

[23] Plant size is determined by the U.S. Census Bureau in terms of revenues and/or employment.

studies (see Manning, 2003; Webber, 2015; Sokolova and Sorensen, 2020). When we compare our markdown estimates with the meta-analysis on labor supply elasticities by Sokolova and Sorensen (2020), our estimates fall below the median of this literature. We conclude that the data support the hypothesis that the average (or even median) manufacturing plant operates in a monopsonistic environment.

Table I: Estimated plant-level markdowns in U.S. manufacturing: markdowns are sizable and considerably larger than unity. The average manufacturing plant operates in a monopsonistic environment.[a]

INDUSTRY GROUP	Median	Mean	IQR_{75-25}	SD
Petroleum Refining	2.391	2.547	1.828	1.267
Computer and Electronics	2.296	2.558	1.227	1.075
Plastics and Rubber	1.812	1.906	0.582	0.584
Food and Kindred Products	1.761	1.913	0.872	0.823
Paper and Allied Products	1.695	1.795	0.573	0.625
Chemicals	1.623	1.817	0.941	0.870
Lumber	1.540	1.623	0.467	0.522
Primary Metals	1.450	1.503	0.506	0.479
Motor Vehicles	1.368	1.422	0.376	0.432
Printing and Publishing	1.345	1.495	0.454	0.632
Electrical Machinery	1.317	1.416	0.519	0.513
Fabricated Metal Products	1.257	1.313	0.339	0.360
Nonelectrical Machinery	1.246	1.317	0.532	0.454
Miscellaneous Manufacturing	1.208	1.254	0.348	0.358
Textile Mill Products	1.208	1.266	0.412	0.454
Furniture and Fixtures	1.150	1.167	0.320	0.358
Nonmetallic Minerals	1.139	1.217	0.372	0.522
Apparel and Leather	1.035	1.146	0.413	0.539
Whole sample	**1.364**	**1.530**	**0.618**	**0.708**
Sample size	$1.393 \cdot 10^6$			

[a] Markdowns are estimated under the assumption of a translog specification for gross output. Each industry group in manufacturing corresponds to the manufacturing categorization of the U.S. Bureau of Economic Analysis, which approximately follows a 3-digit NAICS specification. The distributional statistics are calculated using sampling weights provided in the data. Source: Authors' calculations from ASM/CM data in 1976–2014.

Moreover, there is considerable variation in markdowns across plants *within* the same industry. The average *within-industry* interquartile range (standard deviation) of markdowns is 61.6 (60.4) percent. This suggests that heterogeneity in markdowns likely relates to idiosyncratic factors, such as plant-level productivity differences or specific human capital,

rather than industry-wide characteristics, such as legacy structure, institutional agreements, or industry regulations.[24]

Recent studies have emphasized that the welfare cost of market power distortions can be considerable (Edmond, Midrigan and Xu, 2021; Baqaee and Farhi, 2020; Berger, Herkenhoff and Mongey, Forthcoming), and so we next turn to understanding the determinants of markdown variation.

3.2 Heterogeneity in markdowns

VARIANCE DECOMPOSITION. To investigate markdown heterogeneity, we first decompose markdowns into their components according to Equation (3). Micro-level markdowns are additively separable (in natural logs) according to:

$$\ln(\nu) = \ln(\theta_\ell) - \ln(\alpha_\ell) - \ln(\mu) \tag{7}$$

Recall that θ_ℓ is the elasticity of output with respect to labor, α_ℓ is labor's share of revenue, and μ is the product markup. We can then apply the following variance decomposition:

$$\begin{aligned} V(\ln(\nu)) = {} & V(\ln(\theta_\ell)) + V(\ln(\alpha_\ell)) + V(\ln(\mu)) \\ & - 2 \cdot [\text{cov}(\ln(\theta_\ell), \ln(\alpha_\ell)) - \text{cov}(\ln(\alpha_\ell), \ln(\mu)) + \text{cov}(\ln(\theta_\ell), \ln(\mu))] \end{aligned} \tag{8}$$

In Table II, we document the contribution of each component.

The variation in markdowns is largely accounted for by the variation in output elasticities θ_ℓ and labor shares α_ℓ, as well as their covariance. Variations in markups, on the other hand, play a quantitatively small role for markdown variation.[25] Our results consequently imply that the main determinants of markdown variation are different from those that drive variation in markups.

[24] This pattern accords with the dispersion in revenue-based total factor productivity documented by Syverson (2004b). In Online Appendix O.1, we explore whether *industry*-level characteristics (e.g., unionization) can explain some of the observed heterogeneity in markdowns. Qualitatively, we find a slight negative relationship between markdowns and unionization at the industry-state level.

[25] Although the focus of this paper is on the markdown estimation, we acknowledge that a more complete treatment of the relationship between plant-level markups and markdowns is worthy of future research.

Table II: Variance in plant-level markdowns is accounted for by the variances and covariances of output elasticities θ_ℓ and labor shares α_ℓ. Variance in markups is quantitatively small.[b]

		Variance	*Relative contribution*
Markdown	ν	0.1696	1.000
Elasticity	θ_ℓ	0.3149	1.857
Labor share	α_ℓ	0.3813	2.248
Markup	μ	0.0276	0.1627
		Covariance	*Relative contribution*
	θ_ℓ, α_ℓ	0.2804	-3.307
	θ_ℓ, μ	-0.00601	0.0709
	α_ℓ, μ	-0.00271	-0.0320

[b]Variance decomposition of plant-level markdowns as based on Equation (8). Source: Authors' calculations from ASM/CM data in 1976–2014.

SIZE, AGE, AND PRODUCTIVITY. We proceed to investigate the source(s) of markdown variation by focusing on idiosyncratic factors, especially those likely to be related to labor supply elasticities and labor shares. In particular, we look at the relationship between markdowns and establishment size. A recent literature has emphasized the welfare costs of markups and markdowns that vary through size alone (Edmond, Midrigan and Xu, 2021; Berger, Herkenhoff and Mongey, Forthcoming), and it is thus natural to ask whether size can account for a substantial amount of variation in our markdown estimates.

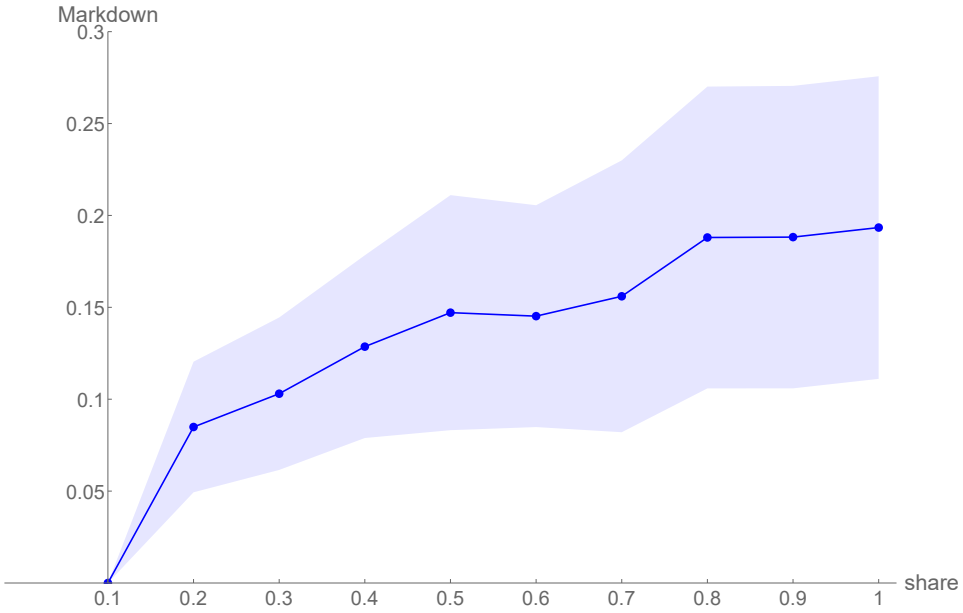
As mentioned by Haltiwanger, Jarmin and Miranda (2013), however, it is important to control for age while assessing size effects because the two are heavily correlated and could thus confound each other. We therefore run a set of nonparametric regressions to flexibly capture the heterogeneity of markdowns by size and age. These regressions are of the following form:

$$\nu_{it} = \beta_0 + \sum_{d=1}^{S} \beta_d^{\text{size}} \cdot \mathbf{1}_{s_{it} \in S_d} + \sum_{d=1}^{A} \beta_d^{\text{age}} \cdot \mathbf{1}_{\text{age}_{it} \in A_d} + \mathbf{X}_{it}'\gamma + \varepsilon_{it}, \tag{9}$$

where $\mathbf{X}_{it}$ contains a full set of industry, state, and year fixed effects. We create size dummies over a grid of $S = 10$ equally spaced bins of a plant's employment share of its

local labor market (NAICS3-county cell).[26] We categorize age into 8 groups.[27] The results, depicted in figure 1, display a clear picture: markdowns are monotonically increasing in size.

Figure 1: Average markdowns increase with establishment size.



Note: The figure shows point estimates and 95 percent confidence intervals of plant-specific markdowns on size (as measured by employment share) indicators, controlling for indicators for plant age and industry, as well as state and year fixed effects. The omitted group is the smallest size indicator, so coefficients reflect deviations relative to this baseline. The indicator labeled "0.1" is equal to unity for those plants with employment shares $s \in (0, 0.1]$. Other indicators are defined similarly. Standard errors are clustered at the industry level. Source: Authors' own calculations from ASM/CM data in 1976–2014.

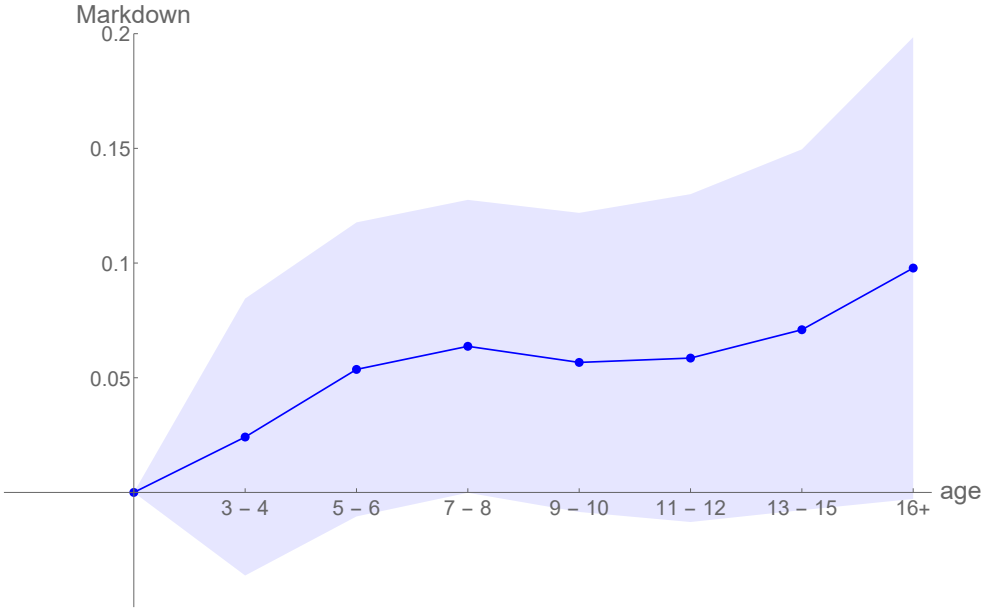
Conditional on plant age, industry, and other covariates, markdowns for plants with the highest shares of employment are, on average, roughly 20 percent higher than for the smallest plants.

The results for age are somewhat similar but less clear-cut. Without size controls, there is a statistically significant positive age gradient in markdowns. However, as shown in figure 2, this relationship is attenuated once size controls are included, and we cannot

[26]Following Haltiwanger, Jarmin and Miranda (2013), we apply employment weights. However, our results are little affected if we do not use employment weights.

[27]These age groups include: 0–2 years, 3–4, 5–6, 7–8, 9–10, 11–12, 13–15, and 16+ years. To minimize reporting bias in size and age, we take a plant's employment share and age from the Longitudinal Business Database (LBD), which contains the universe of employers, and merge them to ASM/CM at the establishment-year level.

Figure 2: Markdowns tend to increase with establishment age, but this result is relatively weak conditional on establishment size.



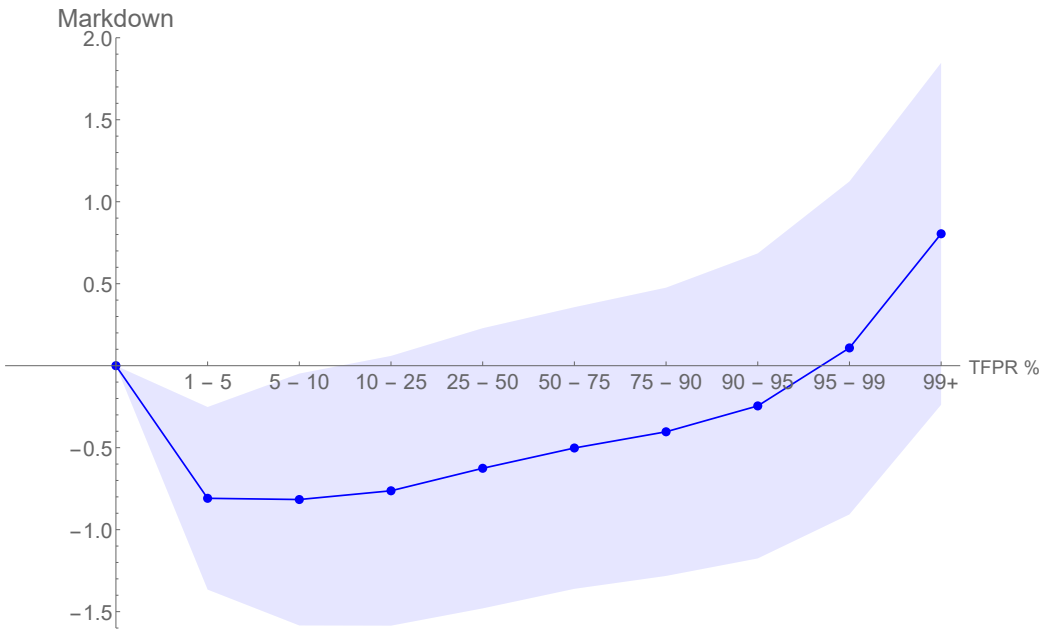
Note: The figure shows point estimates and 95 percent confidence intervals of plant-specific markdowns on age category indicators, controlling for indicators for plant size and industry, as well as state and year fixed effects. The omitted group is the smallest age category, less than three years, so coefficients reflect deviations relative to this baseline. Standard errors are clustered at the industry level. Source: Authors' own calculations from ASM/CM data in 1976–2014.

reject that average markdowns are similar across the plant age distribution. Consequently, the relationship between markdowns and plant age is not especially robust.

We also investigate the relationship between markdowns and plant-level productivity. Previous studies have identified a positive association between wages and profits (or sales) per worker. Christofides and Oswald (1992), for example, find a robust relationship between industry profits and firm-level wages, while Van Reenen (1996) documents that innovative firms tend to pay their workers higher wages. Similarly, strategies popularized by Abowd, Kramarz and Margolis (1999) have found that positive sorting and correlation between workers' bargaining power and firms' profitability measures partially explain the positive relationship between wages and productivity. More recently, Card, Devicienti and Maida (2014) estimate an elasticity of wages to (economic) rents of approximately 4 percent, and Seegmiller (2021), using a dynamic wage posting model, finds that public firms higher in the labor productivity distribution have greater markdowns. In a related vein, we correlate

plant-level productivity not simply to the average wage but rather to its markdown—the ratio between the marginal revenue product and the wage. Since we do not observe quantities, we proxy physical productivity (TFPQ) by revenue productivity (TFPR).[28]

Figure 3: There is a qualitative U-shaped relationship between establishment-level markdowns and TFPR.



Nonparametric regressions of markdowns on productivity (employment-weighted). To avoid collinearity issues, we follow Haltiwanger, Jarmin and Miranda (2013) and apply the normalization $\beta_1^{\text{TFPR}} = 0$ (lower percentile of the TFPR distribution). Hence, productivity coefficients should be interpreted as deviations relative from this baseline. Standard errors are clustered at the industry level. Source: Authors' own calculations from ASM/CM data in 1976–2014.

Unlike that for size and age, we find that the relationship between a plant's markdown and productivity is not monotonic. As shown in Figure 3, the data suggest more of a U-shaped association between markdowns and productivity. Markdowns are increasing in TFPR only after about the 10th percentile in the TFPR distribution. Though the variation in markdowns across the TFPR distribution is large, the coefficients on the productivity percentile categories are noisily estimated, and most estimates are not significantly different from zero at the 5 percent level.[29] Although there is suggestive evidence that the most productive estab-

[28] Although being able to observe TFPQ would be ideal, Foster, Haltiwanger and Syverson (2008) show that TFPQ and TFPR are highly correlated with each other in a subsample of manufacturing plants for which both measures of productivity can be constructed.

[29] The U-shape relationship between markdowns and TFPR does not appear to be driven by outliers in

lishments may have larger markdowns, given the relative imprecision in these estimates, we do not make strong inferences on the productivity-markdown relationship.

SCOPE AND HIGH-TECH STATUS. Another feasible dimension of heterogeneity is the extent to which markdowns vary for plants belonging to firms with multiple establishments or with wide industrial or geographical scope. Such plants likely belong to firms with greater capitalization and internal networks for resource reallocation, potentially increasing the scope for markdowns (Giroud and Mueller, 2019). We thus create binary variables that equal one when a plant is owned by a firm that has at least two active establishments ("multi-unit"), owned by a firm with establishments in two or more different 6-digit NAICS industries ("industry-scope"), or owned by a firm with establishments in two or more counties ("geography-scope").

Table III: Plants belonging to multi-unit firms or firms active in more than one sector/location have higher markdowns.[c]

Dependent variable: MARKDOWNS				
	MULTI-UNIT	INDUSTRIAL	GEOGRAPHICAL	HIGH-TECH
Premium	0.2514	0.2543	0.2558	−0.09054
	(0.04236)	(0.04173)	(0.04247)	(0.1081)
Observations (in millions)	1.393	1.393	1.393	1.393
R^2	0.2668	0.2696	0.2697	0.2511

[c] See note to Table I on markdown estimation. Industrial and geographical scope refer to a plant that is owned by a firm active in multiple 6-digit NAICS industries or 5-digit FIPS counties, respectively. A plant is considered "high-tech" based on its 4-digit NAICS code and the categorization of Decker et al. (2016). Standard errors, in parentheses, are clustered at the industry level. Source: Authors' calculations from ASM/CM data in 1976–2014.

Table III shows that plants owned by multi-unit firms charge markdowns more than 0.25 greater, on average, than stand-alone plants. We find quantitatively similar markdown premia for plants with greater industrial or geographical scope. These results continue to hold if we control for firm size.

We additionally investigate whether plants in the high-tech sector have higher markdowns.[30] High-tech firms play a disproportionate role in aggregate employment and productivity

TFPR or by entering and exiting plants. If we drop these observations from our sample, the U-shape flattens somewhat, but neither the qualitative pattern nor the statistical significance changes.

[30] We follow the definition for high-tech sectors of Decker et al. (2016). For manufacturing, these include the 4-digit NAICS industries 3254 (pharmaceuticals), 3341 (computers), 3342 (communications equip-

growth (see Decker et al., 2016), thus it is interesting to know whether they charge higher markdowns on their labor. We find, however, that average markdowns for high-tech plants are (weakly) *lower* than for other plants, which is consistent with the results on more innovative firms in Van Reenen (1996).

HETEROGENEOUS LABOR. Our baseline estimates allow for different types of labor across plants but implicitly assume that labor is homogeneous *within* plants. The ASM and CM break down the plant wage bill into components of production and nonproduction workers, allowing us to test for heterogeneity in markdowns across these two types of labor (treating them as separate inputs in the production function).[31] Table IV shows estimated markdowns by industry separately for each labor type. Allowing for labor heterogeneity does not greatly affect the pattern from our original estimates. Markdowns for nonproduction workers correspond closely to the baseline in Table I, while those for production workers, while more variable, are somewhat higher, on average, than the baseline. However, there is little evidence of any systematic difference between the two groups that would suggest markdowns are driven by only one type of labor.[32]

That markdowns are similar for production and nonproduction labor may seem surprising to the extent that these groups are presumed synonyms for low-skill and high-skill, respectively, and that low-skill workers should have an easier time finding a comparable outside employment option. However, upon reflection, the pattern we find should not be surprising. First, production and nonproduction workers are *not* synonyms for low- and high-skill workers; rather, the former group, in addition to "fabricating, processing, [and] assembling," also includes highly skilled craftspersons, inspectors, and product developers. Second, the summary results in Table IV subsume spatial heterogeneity. The portability of a worker's skills across jobs depends not only on that worker's type of skill but on the demand for that skill in the local labor market.[33] Thus, it is plausible that production workers are

ment), 3344 (semiconductors and electronic components), 3345 (precision and control instruments) and 3364 (aerospace).

[31]The Census Bureau defines production workers as those "engaged in fabricating, processing, assembling, inspecting, receiving, packing, warehousing, shipping (but not delivering), maintenance, repair, janitorial, guard services, product development, auxiliary production for [the] plant's own use, record keeping, and other closely associated services." This includes line supervisors but not managerial and administrative positions.

[32]It is also reassuring that, even under the strict interpretation of Assumption VI and Proposition 1, we continue to find markdowns for production workers.

[33]Marinescu and Rathelot (2018) show that 81 percent of job seekers apply within their metropolitan area of residence, while Macaluso (2019) finds that earnings of laid-off workers recover faster if their last job used skills common to many jobs in the workers' metropolitan area.

Table IV: Markdowns for both production and nonproduction workers exceed unity and are similar to baseline estimates in Table I. Markdowns for one group are not systematically higher than for the other.[d]

	Nonproduction		*Production*	
INDUSTRY GROUP	Mean	Median	Mean	Median
Food and Kindred Products	2.395	2.174	2.014	1.848
Textile Mill Products	1.924	1.736	1.460	1.403
Apparel and Leather	1.311	1.216	1.186	1.122
Lumber	1.660	1.553	1.707	1.620
Furniture and Fixtures	1.372	1.310	1.199	1.138
Paper and Allied Products	1.232	1.125	2.150	2.049
Printing and Publishing	2.021	1.896	1.243	1.142
Chemicals	1.599	1.400	2.473	2.146
Petroleum Refining	2.682	2.356	2.254	1.804
Plastics and Rubber	1.398	1.317	1.802	1.713
Nonmetallic Minerals	1.299	1.204	1.628	1.504
Primary Metals	1.824	1.760	1.416	1.339
Fabricated Metal Products	1.474	1.384	1.530	1.422
Nonelectrical Machinery	1.539	1.359	5.018	4.530
Electrical Machinery	1.383	1.311	1.667	1.526
Motor Vehicles	1.450	1.411	1.523	1.439
Computer and Electronics	2.620	2.436	3.383	2.954
Miscellaneous Manufacturing	1.532	1.456	1.344	1.258
Whole sample	1.682	1.488	1.963	1.527
Baseline	1.530	1.364		

[d]See note to Table I on markdown estimation. The summary statistics under "Nonproduction" ("Production") reflect markdowns applied to nonproduction (production) workers. Source: Authors' calculations from ASM/CM data in 1976–2014.

subject to lower markdowns in some locations (where the opportunities for alternative employment are plentiful) but not in others (where alternative employers are scarce), and this accords with their greater dispersion in markdowns. Third, it is not clear *a priori* whether production workers are more subject to labor market power than nonproduction workers. Outside employment options for both groups may be limited by noncompete employment contracts (Starr, Prescott and Bishara, 2021), which are quite prevalent in manufacturing (Colvin and Shierhold, 2019). Indeed, using a structural nested logit model, Azar, Berry and Marinescu (2019b) find little difference in labor market power between higher- and lower-paying occupations.

4 Secular trends in aggregate market power

4.1 Aggregation of markdowns

Thus far, we have focused on cross-sectional markdown dispersion, pooling across years, and have shown that (i) the average manufacturing plant operates in a monopsonistic environment, and (ii) plant-level markdowns vary substantially across and within industries but are positively associated with plant size. While an increase in labor market power is consistent with several observed secular trends in the U.S. economy, there is still little direct time-series evidence for widening gaps between marginal revenue product of labor and wages (Syverson, 2019). In this section, we investigate time trends in *aggregate* markdowns to gauge whether monopsony in U.S. manufacturing has increased over time.

Although we have estimates for markdowns at the plant level, aggregation is not straightforward. Previous studies on markups have relied on weighted averages based on sales (De Loecker, Eeckhout and Unger, 2020) or employment (Rossi-Hansberg, Sarte and Trachter, 2020), but it is unclear in which context and for which questions it is appropriate to use these particular weights for markdown aggregation.[34] We propose instead a flexible measure of aggregate markdowns that is 1) theoretically consistent with aggregate wedges, in the spirit of Edmond, Midrigan and Xu (2021), and 2) accounts for the local nature of labor markets.

We argue that a measure for aggregate markdowns needs to satisfy these two requirements. First, consistency with aggregate wedges is natural since micro-level markdowns are based on micro-level wedges.[35] Hsieh and Klenow (2009), and Itskhoki and Moll (2019) use similar approaches in defining aggregate productivity as a function of micro-level productivities. Importantly, we do not have to impose a specific structure for labor or output markets in order to achieve consistency with aggregate wedges. Consequently, our measure for the aggregate markdown is consistent with a variety of monopsony models.[36]

[34] Aggregation is more straightforward when one is willing to impose more structure. Berger, Herkenhoff and Mongey (Forthcoming) show that in their model, a labor market counterpart to Atkeson and Burstein (2008), Herfindahl indices of payroll are sufficient statistics to calculate aggregate labor market power, but this need not hold more generally.

[35] Aggregate wedges are consistent with gaps that a fictional representative firm would face. This is the interpretation adopted in, for example, Cole and Ohanian (2002), Gali et al. (2007), and Karabarbounis (2014). In particular, the aggregate wedge that defines the aggregate markdown in our setup is part of the gap between marginal product of labor and real wages in Karabarbounis (2014).

[36] We discuss these in Online Appendix O.7.

Second, several studies have shown that labor markets are "local" because workers find it costly to search for jobs in locations far from where they reside. For instance, Manning and Petrongolo (2017) estimate that the attractiveness of jobs decays sharply with distance, while Marinescu and Rathelot (2018) find that job seekers are 35 percent less likely to apply to a job 10 miles away from their zip code of residence. It is similarly costly to search in settings using different skills or performing different tasks (Kambourov and Manovskii, 2009). We thus characterize a *local* labor market as a sector-location pair, using 3-digit NAICS codes and counties, resulting in a total of more than 20 distinct sectors (within manufacturing) and over 3,000 locations. In what follows, we denote sectors by j and locations by l.[37]

We define the aggregate markup $\mathcal{M}_{jlt}$ in a labor market (j, l) as the wedge between the aggregate output elasticity of some flexible input and its revenue share.[38] We define the aggregate markdown $\mathcal{V}_{jlt}$ as the part of the wedge between the aggregate output elasticity of labor and the labor share that is not accounted for by markups. By construction, the following identities hold at the market level:

$$\frac{\theta^L_{jlt}}{\alpha^L_{jlt}} = \mathcal{M}_{jlt} \cdot \mathcal{V}_{jlt} \tag{10}$$

$$\frac{\theta^M_{jlt}}{\alpha^M_{jlt}} = \mathcal{M}_{jlt}, \tag{11}$$

where, with some abuse of notation, θ^L_{jlt} and α^L_{jlt} are, respectively in some market, the aggregate output elasticity of labor and the labor share. These objects are defined analogously for material inputs. We say that any measures for the aggregate markup $\mathcal{M}_{jlt}$ and markdown $\mathcal{V}_{jlt}$, that are based on micro-level markups and markdowns, are consistent with aggregate wedges whenever $\mathcal{M}_{jlt}$ and $\mathcal{V}_{jlt}$ satisfy equations (10) and (11).

Then, we can show the following:

PROPOSITION 2. Let Assumption **I** hold. Furthermore, let Assumptions **II**–**VI** hold for material inputs and assumptions **II** and **IV**–**VI** hold for labor. If firm-level wage schedules

[37] We thank Jan Eeckhout for his suggestion to explore aggregate markdowns while thinking of the local nature of labor markets.

[38] Edmond, Midrigan and Xu (2021) adhere to a similar definition but instead assume that labor is fully flexible.

are differentiable, then the **aggregate markdown** and **aggregate markup** for a local labor market (j,l) are consistent with aggregate wedges whenever they are equal to:

$$\mathcal{V}_{jlt} = \frac{\left(\sum_{i\in F_t(j,l)} s_{it} \cdot \frac{\theta^L_{it}}{\theta^L_{jlt}} \cdot (\nu_{it}\mu_{it})^{-1}\right)^{-1}}{\left(\sum_{i\in F_t(j,l)} s_{it} \cdot \frac{\theta^M_{it}}{\theta^M_{jlt}} \cdot \mu_{it}^{-1}\right)^{-1}} \tag{12}$$

$$\mathcal{M}_{jlt} = \left(\sum_{i\in F_t(j,l)} s_{it} \cdot \frac{\theta^M_{it}}{\theta^M_{jlt}} \cdot \mu_{it}^{-1}\right)^{-1}, \tag{13}$$

where s_{it} are sales weights (i.e., $s_{it} = \frac{p_{it}y_{it}}{P_{jlt}Y_{jlt}}$) and $F_t(j,l)$ denotes the set of firms in labor market (j,l).

Proof. See Appendix B.1. □

Whenever the market for material inputs is perfectly competitive, we can use an insight similar to the one used in Proposition 1. Recall that Proposition 1 states that firm-level markups are equal to the ratio between the output elasticity for materials and their revenue share. If we define the *aggregate* markup as being equal to the ratio between the *aggregate* output elasticity for materials and their *aggregate* revenue share, then the aggregate markup is a weighted harmonic average of firm-level markups, i.e., $\mathcal{M}_{jlt} = \left(\sum_{i\in F_t(j,l)} s_{it} \cdot \frac{\theta^M_{it}}{\theta^M_{jlt}} \cdot \mu_{it}^{-1}\right)^{-1}$, similar to Edmond, Midrigan and Xu (2021).

Using an analogous argument, we derive that the product of the aggregate markdown and markup is a weighted harmonic average of the product of firm-level markdowns and markups. We obtain:

$$\mathcal{V}_{jlt} \cdot \mathcal{M}_{jlt} = \left(\sum_{i\in F_t(j,l)} s_{it} \cdot \frac{\theta^L_{it}}{\theta^L_{jlt}} \cdot (\nu_{it}\mu_{it})^{-1}\right)^{-1}$$

Given that we have an expression for the aggregate markup, the expression for the aggregate markdown follows automatically. If output elasticities do not vary across firms within a given labor market—i.e., firms have Cobb-Douglas production technologies—then aggregation follows by taking sales-weighted harmonic averages. If production technologies are not Cobb-Douglas, we need only apply correction terms that deal with heterogeneity in

output elasticities—as can be seen from Equations (12) and (13) in proposition 2.

We then aggregate across labor markets through employment weights, defining the aggregate markdown as:

$$\mathcal{V}_t = \sum_{j \in J} \sum_{l \in L} \omega_{jlt} \mathcal{V}_{jlt}, \tag{14}$$

where $\mathcal{V}_{jlt}$ is as in equation (12), and ω_{jlt} denotes the employment share of labor market (j, l). Following the literature on markups (e.g., De Loecker, Eeckhout and Unger, 2020) and concentration (Autor et al., 2020; Rossi-Hansberg, Sarte and Trachter, 2020), we proceed by constructing markdowns at the firm, rather than plant, level using the CM.[39]

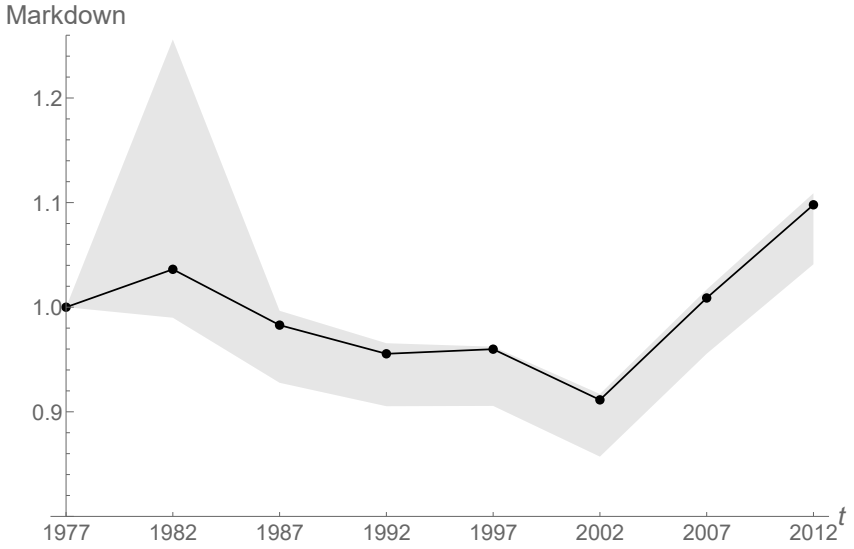
Figure 4 illustrates the resulting time trend of aggregate markdowns, $\mathcal{V}_t$. In contrast with previous trend estimates of markups (e.g., De Loecker, Eeckhout and Unger, 2020), the aggregate markdown $\mathcal{V}_t$ is not monotonic. Instead, $\mathcal{V}_t$ falls between the early 1980s and early 2000s, after which it begins to sharply increase. This pattern is inconsistent with the notion that increasing labor market power by firms is the primary cause of the decline in the labor share, which began well before the early 2000s. Yet the stark increase in the aggregate markdown since this time is interesting, as others have noted acceleration in the decline in U.S. business dynamism over the same horizon (see, e.g., Decker et al., 2016).[40]

Contrasting the time series for the aggregate markdown in equation (14) with two commonly used alternatives highlights the importance of using a local measure of aggregate markdowns that is also micro-founded. The first alternative we consider is a labor mar-

[39] By construction, the aggregate markdown is an employment-weighted average of markdowns at the market level. The latter is constructed using Equations (12) and (13). However, it is difficult to construct these objects with the previously used ASM sample, since our definition of a local labor market is rather narrow. Recall that the ASM is a representative sample and does not contain the universe of manufacturing plants. This is sufficient for use in a repeated cross-section, as in our earlier analyses of the distribution of plant-level markdowns, but not for employment-weighted aggregation. In particular, the number of observations available to construct $\mathcal{V}_{jlt}$ and $\mathcal{M}_{jlt}$ might be rather small for some labor markets (j, l) and induce measurement error biases in these objects. Thus, we instead utilize the CM, which contains the universe of manufacturing plants but only at a quinquennial frequency.

[40] To understand which groups of firms determine movements in the aggregate markdown, we have applied a decomposition in the spirit of Foster, Haltiwanger and Krizan (2001) to $\mathcal{V}_{jlt}$. This decomposition analyzes the role of changes within firms, across firms, and through firm entry and exit. As documented in Online Appendix O.2.4, no single component drives the trend. In Online Appendix B.2, we show that the trend in the aggregate markdown also cannot be explained by changes in the composition of local labor markets or by excluding health and pension benefits from the labor share.

Figure 4: Time evolution of the aggregate markdown across U.S. manufacturing plants from 1977 to 2012.



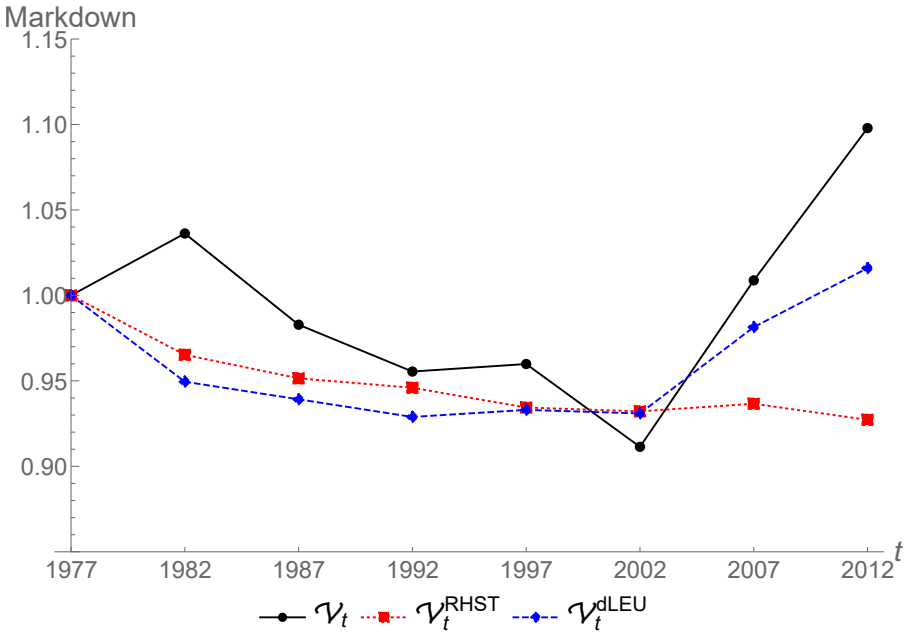
Markdowns are constructed under the assumption of translog production and aggregated according to expressions (12) and (14). The aggregate markdown is normalized relative to its initial value in 1977. Standard errors are obtained through a block bootstrap procedure and are percentile-based and thus not symmetrical; production function parameters enter firm-level markdowns in a highly nonlinear fashion, and firm-level markdowns also enter the aggregate markdown nonlinearly. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

ket equivalent of the aggregate markup measure used in De Loecker, Eeckhout and Unger (2020):

$$\begin{aligned}\mathcal{V}_t^{\text{dLEU}} &= \sum_{p\in P_t} \omega_{pt}\nu_{pt} \\ &= \sum_{f\in F_t} \omega_{ft}\left[\sum_{p\in P_t(f)} s_{pft}\nu_{pt}\right] \\ &\equiv \sum_{f\in F_t} \omega_{ft}\nu_{ft}, \end{aligned} \tag{15}$$

where P_t denotes the set of active plants in year t and s_{pft} the employment share of plant p in firm f. By construction, $\mathcal{V}_t^{\text{dLEU}}$ is an employment-weighted average of plant-level markdowns. This is identical to a firm-level average whenever firm-level markdowns are calculated as employment-weighted averages across a firm's plants.

Figure 5: Our micro-founded aggregate markdown measure $\mathcal{V}_t$ (solid black) decreases between 1977 and 2002 and increases afterwards. The employment-weighted aggregate markdown à la De Loecker, Eeckhout and Unger (2020) (dashed blue) shows a similar pattern qualitatively, while a local aggregate inspired by local concentration in Rossi-Hansberg, Sarte and Trachter (2020) (dotted red) is steadily decreasing.



Markdowns are constructed under the assumption of translog production and aggregated according to expressions equation (14), equation (15), and equation (16), respectively. All measures are normalized relative to their initial value in 1977. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

A second option is a measure for the aggregate markdown that mirrors the aggregate measure for local employment concentration, as in Rossi-Hansberg, Sarte and Trachter (2020). This approach still aggregates micro-level markdowns through employment weights, similar to Equation (15), but does so in two stages. First, micro-level markdowns are aggregated through their respective employment shares *within* each market, then markets are aggregated through employment weights to construct an aggregate measure. This leads to:

$$\mathcal{V}_t^{\text{RHST}} = \sum_{j \in J} \sum_{l \in L} \omega_{jlt} M_{jlt} \quad \text{with} \quad M_{jlt} = \sum_{f \in F_t(j,l)} \omega_{fjlt} \nu_{fjlt}. \tag{16}$$

Figure 5 illustrates that while our preferred measure $\mathcal{V}_t$ is decreasing until the early 2000s and sharply increasing afterwards, the alternatives display a different time evolution. While

$\mathcal{V}_t^{\text{dLEU}}$ follows our measure in a qualitative sense, $\mathcal{V}_t^{\text{RHST}}$ monotonically decreases over the whole period.[41]

4.2 Comparing markdowns with concentration indices

Several recent studies have used measures of concentration—either in output or input markets—as proxies for market power, both cross-sectionally and longitudinally. In this subsection, we discuss whether concentration is an accurate proxy for market power—at least within manufacturing—by comparing its cross-sectional *and* time-series properties with our estimated markdowns.

The Herfindahl-Hirschman Index (HHI) is a canonical way to summarize the level of concentration in output markets (Autor et al., 2020; Rossi-Hansberg et al., 2020) and has been increasingly popular in studies of labor markets as well (Rinz, 2018; Azar et al., 2020b; Azar, Marinescu and Steinbaum, 2020a; Benmelech, Bergman and Kim,2020; Dodini et al., 2020). Yet there is no *a priori* reason concentration and market power must be positively correlated. It may seem intuitive that large employers are able to exert more labor market power, but as Syverson (2019) points out for output markets, a negative correlation can arise naturally in the framework of Melitz and Ottaviano (2008) and has been empirically observed in several studies (Syverson, 2004a; Syverson, 2004b; Goldmanis et al., 2010). Despite these critiques, concentration indices have never been explicitly compared to direct, wedge-based measures of market power, at least not at a scale as wide as the whole manufacturing sector. This is precisely our aim in this section.

For our comparison between aggregate markdowns and measures of concentration, we adopt the HHI as our main measure of market-level concentration and define it in a standard fashion:

$$\text{HHI}_{mt} = \sum_{f \in F_t(m)} \left(\frac{x_{ft}}{X_{F(m)t}} \right)^2 \quad \text{s.t. } X_{F(m)t} = \sum_{f' \in F_t(m)} x_{f't}, \tag{17}$$

where m denotes a market, $F_t(m)$ the set of firms in market m during a year t, and x is

[41] These differing trends can be rationalized by how markdowns are aggregated at the market level. Markdowns are aggregated linearly under the measure $\mathcal{V}_{jlt}^{\text{RHST}}$, but $\mathcal{V}_{jlt}$ is constructed through the ratio of two harmonic weighted averages. Furthermore, each of these averages contain markups and reflect heterogeneity in output elasticities for labor and material inputs. Empirically, we have confirmed that these latter factors explain more of the difference between $\mathcal{V}_t^{\text{RHST}}$ and $\mathcal{V}_t$.

a measure of size (often employment or sales). We focus on labor markets and thus set $m = (j, \ell)$ to remain consistent with our previous analyses.

By construction, the HHI ranges from $1/F_t(m)$ to 1. A value of 1 indicates maximum concentration—the presence of only one active seller/employer in a specific market-year. If firms were equally sized, the inverse of the HHI would be equal to the number of employers $F_t(m)$ in a market m.

There are two common approaches to combining market-level concentration measures into an aggregate measure. Under the first approach, HHIs are constructed at the industry level (so that a market m is a national industry) and then aggregated through employment or sales weights. Following Autor et al. (2020), we refer to these as measures of national concentration.

In contrast to this "national" approach, Rossi-Hansberg, Sarte and Trachter (2020) have argued that market competition is sometimes better captured at the local level, which may especially be the case for labor. Therefore, markets are instead defined through sector-location cells. Formally:

$$\begin{aligned} \text{LOCAL}_t &= \sum_{j\in J}\sum_{l\in L} \omega_{jlt}\text{HHI}_{jlt} \qquad (18) \\ &= \sum_{j\in J}\sum_{l\in L} \omega_{jlt}\left[\sum_{f\in F_t(j,l)}\left(\frac{x_{flt}}{X_{F(j,l)t}}\right)^2\right] \text{ s.t. } X_{F(j,l)t} = \sum_{f'\in F_t(j,l)} x_{f'lt} \end{aligned}$$

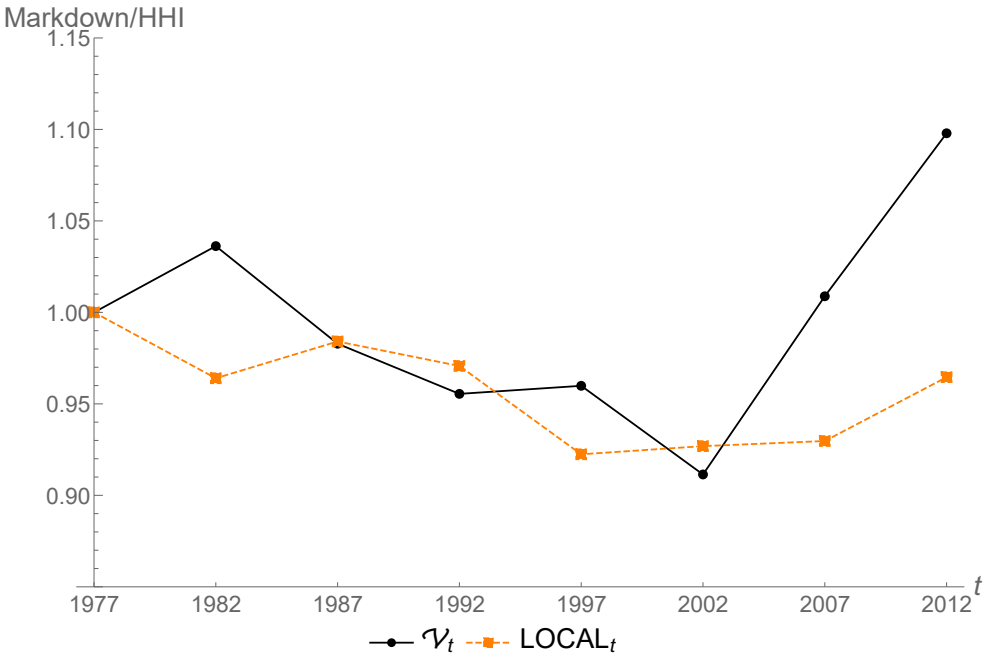
Following our reasoning on the local nature of labor markets as in section 4.1, we implement equation (18) with data on employment as our preferred measure underlying both x_{flt} and ω_{jlt}, where the latter are sector-location cell shares of total employment.[42]

Before we turn to comparisons of aggregates, we first correlate HHIs and our measure

[42]Rossi-Hansberg, Sarte and Trachter (2020) focus on product markets and apply the analogue of Equation (18) to sales for x_{jlt} but employment for ω_{jlt}. Rinz (2018), like us, focuses on labor markets and uses employment for both x_{jlt} and ω_{jlt}, but uses the Longitudinal Business Database to cover all establishments, not just those in manufacturing. While our results here consider (stock) employment concentration, we have also constructed concentration measures based on vacancies (as in Azar et al. (2020a)), job creation flows, and payroll, all of which produce qualitatively similar patterns. Results for vacancies, based on data from Burning Glass Technologies (BGT), can be found in Online Appendix O.8. Remaining results are available upon request.

of markdowns across local labor markets (sector-location cells). We find that the cross-sectional correlation between $\mathcal{V}_{jlt}$ and HHI_{jlt} is essentially zero: across years, this correlation never exceeds 0.02, and it is sometimes negative.[43] Despite this weak cross-sectional correlation, Figure 6 demonstrates that time trends in *aggregate* local concentration (LOCAL_t) and markdowns ($\mathcal{V}_t$) are qualitatively similar. Nonetheless, while both series generally decline between the late 1970s and early 2000s, the subsequent rise in the aggregate markdown occurs both sooner and faster than the uptick in employment concentration.

Figure 6: Within manufacturing, markdowns trend somewhat similarly with local employment concentration but show greater increases since the early 2000s.



Note: The solid black line shows the time series for the aggregate markdown, as in equation (14), and the dashed orange line shows the time series of local employment concentration, as in equation (18). Both are normalized to their initial respective values in 1977. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

These patterns suggest that—at least within manufacturing—cross-sectional and temporal variation in local employment concentration may not necessarily reflect variation in employer market power as measured by markdowns. While it is beyond the scope of this paper to thoroughly analyze these differences, we believe a promising area of future re-

[43] We provide full details in Appendix B.2.

search is advancing theory on the general conditions under which measures of concentration serve as sufficient statistics for wedges between wages and marginal revenue products of labor.[44]

5 Robustness

In this section, we discuss the robustness of our results. We present several exercises that address concerns related to the validity of our markdown formula in Proposition 1. We also discuss some unresolved econometric issues of the proxy variable methodology.

CHOICE OF FLEXIBLE INPUT. The production approach, as popularized by De Loecker and Warzynski (2012), comes with many advantages but is not free of criticism. One of the key identifying assumptions is the requirement for at least one flexible input. Pinpointing such an input is difficult in most publicly available data sets, as most inputs are not observed separately but rather aggregated into broad groups following accounting standards.[45] Although there is still some disagreement on what constitutes a flexible input (e.g., Traina, 2018), we follow the IO literature and assume that material inputs are flexible (Basu, 1995; De Loecker and Warzynski, 2012).

Despite this standard IO assumption, there is some evidence of monopsony in the market for material inputs. For instance, Morlacco (2020), using transaction-level data from French manufacturers, finds evidence of market power in imported intermediate inputs under the identifying assumption that domestically sourced intermediate inputs are perfectly competitive. If material inputs are subject to monopsony, then the ratio $\frac{\theta_\ell}{\alpha_\ell}\left(\frac{\theta_M}{\alpha_M}\right)^{-1}$ in Equation (3) would reflect the markdown for labor *relative to the markdown for materials*, say, ν_ℓ/ν_M. Therefore, in the presence of market power for materials, ν_M implicitly

[44]The weak empirical relationship we document may stem, in part, from different definitions of the relevant labor market. Azar, Marinescu and Steinbaum (2019a), for example, find a negative correlation between job application elasticity and HHI when markets are defined by occupation–commuting zone pairs. Data limitations unfortunately prevent us from investigating the relative roles of occupation vs. industry, or sectoral composition, in explaining these differences.

[45]Recent studies estimating markups typically rely on the Compustat database, in which variable inputs are often identified with "cost of goods sold" (COGS)—which commingles material inputs and variable and fixed labor—or "selling, general, and administrative expenses" (SGA). Because our data allow us to observe *separately* expenditures on capital, labor, material, and energy, we circumvent having to make this choice. Regrettably, neither data source—or any other with similar coverage, to our knowledge—further allows for observation of input quality or other sources of heterogeneity.

exceeds unity, and our estimates for labor markdowns would be biased toward zero, underestimating the extent of labor monopsony in U.S. manufacturing.

A plausible alternative for the flexible input is energy, as advocated by Kim (2017). He maintains that monopsony power through buyer-supplier networks may affect materials, while energy inputs are less prone to monopsony forces, as prices for energy tend to be regulated. On the other hand, Davis et al. (2013) provide robust evidence against the hypothesis that the energy input market is perfectly competitive. They find that plant-level differences within manufacturing industries in energy purchases account for a substantial fraction—at least one-third—of overall price dispersion. Furthermore, they document sizable price gaps between larger and smaller purchases, even when controlling for plant location and/or electric utility provider fixed effects. This seems to contradict the "no monopsony" condition for energy and cautions against its use as a flexible input. Moreover, material inputs have another attractive property in that they represent a much larger share of manufacturing revenues than does energy. Because our measure of markdowns requires division by the flexible input, measurement error is of lesser concern for material inputs compared to energy.[46] We view these factors as compelling evidence in favor of materials as the flexible input.

POINT IDENTIFICATION. Gandhi, Navarro and Rivers (2020) show that the standard assumptions of the proxy variable method (as we describe in subsection 2.2) are insufficient to point-identify production function parameters, and that additional sources of variation in the demand for flexible inputs are required. In turn, Flynn, Gandhi and Traina (2019) have shown that point identification can be restored if the returns to scale of the production function are known. They suggest that a baseline assumption of "constant returns to scale" forms a useful benchmark that performs well in their Monte Carlo simulations. When we impose this constant returns to scale assumption in our context, we reassuringly find our markdown estimates change relatively little (column "CRS" of Table V).[47] This corroborates the notion that our strategy yields reliable estimates of monopsony power in U.S. manufacturing.

[46] We provide evidence of both factors in Online Appendix O.1.3. If we calculate markdowns using energy as the flexible input, we find higher markups and lower markdowns, suggesting labor markdowns are low *relative to* energy markdowns. Additionally, labor markdowns calculated with energy as the flexible input have volatility nearly an order of magnitude greater than our baseline estimates, suggesting division bias.

[47] Additional details of this exercise are in Online Appendix O.3.1.

Table V: Our markdown estimates are robust to alternative assumptions, including ex-ante specified returns to scale (CRS), adjustment costs (Biennial), and including employee benefits in labor compensation (Benefits).[e]

INDUSTRY GROUP	Baseline	CRS	Biennial	Benefits
Food and Kindred Products	1.761	1.475	1.871	1.276
Textile Mill Products	1.208	1.389	3.852	1.128
Apparel and Leather	1.035	0.663	1.074	1.024
Lumber	1.540	1.746	1.508	1.223
Furniture and Fixtures	1.150	1.831	1.122	1.038
Paper and Allied Products	1.695	1.669	1.699	1.431
Printing and Publishing	1.345	0.954	1.344	1.263
Chemicals	1.623	1.765	1.671	1.429
Petroleum Refining	2.391	2.826	2.131	3.463
Plastics and Rubber	1.812	1.424	1.200	1.207
Nonmetallic Minerals	1.139	1.296	1.289	1.147
Primary Metals	1.450	1.712	1.477	1.440
Fabricated Metal Products	1.257	1.684	1.368	1.148
Nonelectrical Machinery	1.246	1.489	1.151	1.068
Electrical Machinery	1.317	1.338	1.184	1.193
Motor Vehicles	1.368	1.663	1.268	1.078
Computer and Electronics	2.296	2.786	2.320	1.669
Miscellaneous Manufacturing	1.208	2.468	1.208	1.114

[e]Markdowns are estimated under the assumption of a translog specification for gross output. For each robustness specification, we report the median of each industry group. Under the column "CRS," we display estimates under the additional assumption of constant returns to scale to address identification concerns. Results from estimating markdowns using biennial data to capture nonconvex adjustment costs are displayed under the column "Biennial." Results from including benefits in the measure of labor compensation (available only from 2002 forward) are displayed under the column "Benefits." Source: Authors' calculations from ASM/CM data in 1976–2014.

LABOR ADJUSTMENT COSTS. In our baseline specification, we assume there are no labor adjustment costs (Assumptions **II** and **IV**). Adjustment costs, however, also can potentially drive a wedge between the output elasticity of labor and its revenue share, possibly contaminating our markdown estimates as expressions of monopsony power. In a quantitative assessment of such bias, however, we find that the impact of labor adjustment costs on our estimates is minimal.

To show that adjustment costs trivially affect our baseline estimates, we proceed in two steps. First, we show that, when labor is subject to convex adjustment costs, the wedge between the marginal revenue product of labor and wages reflects both monopsony power

and adjustment costs. In particular, we show that $\frac{R'(\ell^*)}{w(\ell^*)} = (\varepsilon_S^{-1} + 1) + \mathcal{A}$, where $\mathcal{A}$ would equal zero in the absence of labor adjustment costs. Second, we derive an explicit correction term when labor adjustment costs are quadratic, as commonly suggested (Hall, 2004; Cooper, Haltiwanger and Willis, 2007). This term depends on a plant's growth in labor and its wage bill, and a parameter governing the magnitude of adjustment costs. When we calibrate the correction term over a varied range of labor adjustments and parameters drawn from the literature, we find that the resulting "corrected" estimates of markdowns are not far from baseline. In particular, the most conservative correction adjusts average markdowns by approximately 0.03, quite small relative to the baseline average markdown of 1.53.[48]

We also consider the possibility of fixed or otherwise *non*-convex adjustment costs by re-estimating our markdowns on a biennial basis. Conceptually, nonconvex adjustment costs that may affect our estimates at an annual frequency are less likely to do so at a biennial frequency, especially since the majority of plants demonstrate changes in their employment levels every year. Results from this biennial estimation, as illustrated in Table V under the column "Biennial," are again similar to baseline.

BENEFITS. Our baseline measure of labor costs is based on "wages" and covers a broad range of compensation, including base salaries and wages; bonuses; incentive, overtime, and shift differential pay; and stock grants and options. However, it does not include employer-provided benefits. Consequently, it is possible that we overestimate employers' labor market power to the extent that health and pension benefits are a significant source of overall labor compensation and are correlated with the components of markdowns. We thus re-estimate micro-level markdowns, including benefits in our compensation measure. This more inclusive measure of labor compensation is available only from 2002 onward, which results in a smaller sample than our baseline estimates.[49] The results—displayed in the last column of Table V—show markdown estimates that are slightly lower than baseline, indicating that the inclusion of benefits may be a nontrivial part of the wedge between observed wages and the marginal revenue product of labor. Nonetheless, these estimated markdowns are still above unity for each industry, demonstrating that the presence of monopsony is robust to broader measures of compensation.

[48] We provide derivations of the correction term and a detailed illustration of this exercise in Appendix C.
[49] Appendix O.4.1 provides full details on the components of labor compensation in the ASM/CM data.

REVENUE VERSUS QUANTITY ELASTICITIES. In a recent paper, Bond et al. (2021) argue that the production approach and its implementation with proxy variable estimators cannot generate unbiased estimators of market power. Their critique centers around three issues that we discuss here. The first is that, in most firm-level data sets, the quantity of physical output is not available and instead has to be proxied by deflating revenues. As noted by Klette and Griliches, 1996, this can lead to downward-biased estimates for markups. Bond et al. (2021) further argue that markups under the production approach will mechanically equal unity whenever (deflated) revenues are used to proxy for physical output.

We argue that this is indeed problematic when markups need to be identified in isolation. However, this critique does not apply to markdowns. The key insight is that markdowns are estimated through a *ratio* of elasticities. Revenue elasticities are not equal to output elasticities; however, the component (or "bias") that separates them is identical across inputs and multiplicative. As a result, the bias documented by Bond et al. (2021) cancels out via our construction of markdowns: the ratio of revenue elasticities for two inputs is equal to the ratio of output elasticities for these two same inputs. We formalize this intuition in Proposition 4 of Online Appendix O.5.

INPUTS FOR NONPRODUCTION PURPOSES. Bond et al. (2021) also argue that the production approach can lead to biased estimates of markups and markdowns when the econometrician cannot separate inputs used for purposes other than production that could still affect the quantity of output. For example, inputs could also be used to shift demand (e.g., marketing/advertising). To ensure that our estimates are not subject to this criticism, we need to show that material inputs and labor are used primarily for production purposes.

We argue that it is unlikely *prima facie* that material inputs are used to influence demand. Note that these inputs consist of raw materials, parts, containers, and supplies. Given this categorization, it is safe to assume that material inputs are used solely for production purposes.

However, it is less obvious that no labor inputs are used for shifting demand. We perform two robustness exercises to address this possibility. First, as noted in Section 3.2, our data allow us to separate labor into production and nonproduction workers, and when we estimate markdowns for these types of labor separately, we find that monopsony forces are still significant among production workers specifically. Second, and more generally, our

focus on manufacturing *plants* should render our analysis more robust to this specific critique, since we can plausibly assume that the great majority of a manufacturer's plant-level workforce is indeed employed for production. In fact, we explicitly derive a markdown counterpart for the bias characterized in Bond et al. (2021) and show that, even if we do not separate production from nonproduction labor, the components inducing bias are likely to be small for manufacturers.[50]

SCALAR UNOBSERVABLE ASSUMPTION. The last critique in Bond et al. (2021) relates to the scalar unobservable assumption (our Assumption 4). They show that, in the presence of market power, this assumption cannot be satisfied, since the econometrician is also required to observe a plant's marginal cost of production. Consequently, they suggest using production function estimators that do *not* rely on the scalar unobservable assumption, such as dynamic panel IV methods (Blundell and Bond, 2000). To evaluate this claim, we use Monte Carlo methods to compare the performance of several production function estimators. In particular, we adopt data-generating processes from Ackerberg, Caves and Frazer (2015) that are inconsistent with the econometric assumptions of the family of proxy variable estimators. Nevertheless, as we show in Onlne Appendix O.5.3, our preferred translog estimator outperforms several estimators that do not rely on the scalar unobservable assumption, including those from Blundell and Bond (2000) and Hu, Huang and Sasaki (2020). Hence, even though the scalar unobservable assumption is violated, we do not believe it causes significant problems in practice.

6 Conclusion

This paper provides a characterization of employer market power in the U.S. manufacturing sector, both in the cross-section and over time. We start by estimating markdowns—the wedge between marginal revenue products of labor and wages—at the plant-year level using the "production approach." We find that labor markets in U.S. manufacturing are far from perfectly competitive: the average plant operates in a monopsonistic environment, as it charges a markdown of 1.53. In other words, a worker employed at the average manufacturing plant earns 65 cents of each dollar generated on the margin. We also document that there is a substantial amount of dispersion in markdowns. For our whole sample, the interquartile range of markdowns is 0.618, but most of this variation is observed within

[50]The details for this exercise are found in Online Appendix O.5.

detailed industries, with an average within-industry interquartile range of 0.616. Furthermore, we find that size—whether measured as the relative share of employment in a local labor market or as geographical and sectoral scope—is associated with greater markdowns. On the other hand, we find less correlation with a revenue-based measure of productivity or an indicator for being in a high-tech industry.

We also investigate long-term trends in employer market power, via a novel measure of aggregate markdowns that is consistent with aggregate wedges, accounts for local labor markets, and uses sales-weighted harmonic averages to adjust for production heterogeneity across firms. We find that aggregate markdowns decreased between the late 1970s and early 2000s but increased sharply afterward. This nonmonotonic pattern is inconsistent with the view that the decline in the U.S. labor share (or wage stagnation) was induced by changes in labor market power. Furthermore, we show that popular measures of employment concentration do not line up well with the aggregate markdown, suggesting that the variation underlying local employment concentration does not necessarily reflect the variation underlying employer market power as measured by markdowns.[51]

While we believe that our approach makes significant strides in the estimation and trend measurement of markdowns, we have only scratched the surface in understanding how and why markdowns vary. For example, while we provide qualitative evidence of a negative correlation between the industry's rate of unionization and markdowns, we do not yet know whether the cross-industry variation in markdowns can be further rationalized by the prevalence of noncompete agreements or labor regulations (e.g., right-to-work laws). Such empirical exercises could help us further understand the determinants—and welfare implications—of employer market power.

We also acknowledge that our approach is not without shortcomings. While it is compatible with a broad array of monopsony frameworks, it rules out any model of monopsony in which firms' market power does not originate from an upward-sloping labor supply curve. Most notably, our results cannot be interpreted through the lens of models in the family of Diamond (1982) and Mortensen and Pissarides (1994). However, Dobbelaere and Mairesse

[51] Recent papers have documented that large increases in HHI driven by mergers lead to decreased wages (Arnold, 2020; Prager and Schmitt, 2021). It is unclear, though, whether this relationship holds throughout the HHI distribution, or whether the reduction in wages stems from labor market frictions other than the wedge between wages and marginal product.

(2013) show that wedges between output elasticities and revenue shares can also be used to identify *firm*-level parameters of a static Nash bargaining problem in which risk-neutral workers and firms negotiate over wages and the level of employment. These estimated parameters can be informative for characterizing employer market power in random search settings with perfectly elastic labor supply curves. Last, our econometric methodology does not explicitly allow for factor-biased technological change. While there are estimation methods that do account for labor-augmenting technological change, they do not allow for a generalized production function (Doraszelski and Jaumandreu, 2018; Raval, 2020) or labor market power (Demirer, 2020). We leave investigation of these themes for future research.

References

Abowd, J.M., F. Kramarz, and D.N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, 1999, *67* (2), 251 – 333.

Ackerberg, D., K. Caves, and G. Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 2015, *83* (6), 2411 – 2451.

Ackerberg, D.A., "Timing Assumptions and Efficiency: Empirical Evidence in a Production Function Context," Working Paper, 2020.

Arnold, D., "Mergers and Acquisitions, Local Labor Market Concentration, and Worker Outcomes," Working Paper, 2020.

Atalay, E., "Materials Prices and Productivity," *Journal of the European Economic Association*, 2014, *12* (3), 575 – 611.

Atkeson, A. and A. Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 2008, *98* (5), 1998 – 2031.

Autor, D., D. Dorn, L. Katz, C. Patterson, and J. Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *Quarterly Journal of Economics*, 2020, *135* (2), 645–709.

Azar, J., I. Marinescu, and M. Steinbaum, "Measuring Labor Market Power Two Ways," *AEA Papers and Proceedings*, 2019, *109*, 317–321.

_ , _ , and M.I. Steinbaum, "Labor Market Concentration," *Journal of Human Resources*, 2020, pp. 1218–9914R1.

_ , _ , M. Steinbaum, and B. Taska, "Concentration in US Labor Markets: Evidence from Online Vacancy Data," *Labour Economics*, 2020, *66*, 1018–1086.

_ , S. Berry, and I. Marinescu, "Estimating Labor Market Power," Working Paper, 2019.

Baqaee, D. and E. Farhi, "Productivity and Misallocation in General Equilibrium," *Quarterly Journal of Economics*, 2020, *135* (1), 105–163.

Basu, S., "Intermediate Goods and Business Cycles: Implications for Productivity and Welfare," *American Economic Review*, 1995, *85* (3), 512–531.

_ and J.G. Fernald, "Returns to Scale in U.S. Production: Estimates and Implications," *Journal of Political Economy*, 1997, *105* (2), 249 – 283.

Benmelech, E., N.K. Bergman, and H. Kim, "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?," *Journal of Human Resources*, 2020, pp. 0119–10007R1.

Berger, D., K. Herkenhoff, and S. Mongey, "Labor Market Power," *American Economic Review*, Forthcoming.

Bhaskar, V. and T. To, "Minimum Wages for Ronald McDonald Monopsonies: A Theory of Monopsonistic Competition," *Economic Journal*, 1999, *109*, 190–203.

Blundell, B. and S. Bond, "GMM Estimation with Persistent Panel Data: An Application to Production Functions," *Econometric Reviews*, 2000, *19* (3), 321–340.

Bond, S., A. Hashemi, G. Kaplan, and P. Zoch, "Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data," *Journal of Monetary Economics*, 2021, *121*, 1–14.

Bontemps, C., J.-M. Robin, and G.J. Van den Berg, "An Empirical Equilibrium Job Search Model with Search on the Job and Heterogeneous Workers and Firms," *International Economic Review*, 2001, *40* (4), 1039–1074.

Brooks, W.J., J.P. Kaboski, I.O. Kondo, Y.A. Li, and W. Qian, "Infrastructure Investment and Labor Monopsony Power," *IMF Economic Review*, 2021, *69*, 470–504.

_ , _ , Y.A. Li, and W. Qian, "Exploitation of Labor? Classical Monopsony Power and Labor's Share," *Journal of Development Economics*, 2021, *150* (102627).

Burdett, K. and D.T. Mortensen, "Wage Differentials, Employer Size, and Unemployment," *International Economic Review*, 1998, *39* (2), 257–273.

Card, D., A.R. Cardoso, J. Heining, and P. Kline, "Firms and Labor Market Inequality: Evidence and Some Theory," *Journal of Labor Economics*, 2018, *36* (1), 13–70.

_ , F. Devicienti, and A. Maida, "Rent-Sharing, Holdup, and Wages: Evidence from Matched Panel Data," *Review of Economic Studies*, 2014, *81* (1), 84–111.

Chan, M., K. Kroft, and I. Mourifie, "An Empirical Framework for Matching with Imperfect Competition," Working Paper, 2019.

Christofides, L.N. and A.J. Oswald, "Real Wage Determination and Rent-Sharing in Collective Bargaining Agreements," *Quarterly Journal of Economics*, 1992, *107* (3), 985–1002.

Cole, H. and L. Ohanian, "The U.S. and U.K. Great Depressions Through the Lens of Neoclassical Growth Theory," *American Economic Review*, 2002, *92* (2), 28–32.

Collard-Wexler, A. and J. De Loecker, "Production Function Estimation and Capital Measurement Error," Working Paper, 2020.

Colvin, A.J.S. and H. Shierhold, "Noncompete Agreements," *Economic Policy Institute Report*, 2019, pp. 1 – 16.

Cooper, R., J. Haltiwanger, and J.L. Willis, "Search Frictions: Matching Aggregate and Establishment Observations," *Journal of Monetary Economics*, 2007, *54* (S1), 56–78.

Davis, S. J., C. Grim, J. Haltiwanger, and M. Streitwieser, "Electricity Unit Value Prices and Purchase Quantities: U.S. Manufacturing Plants, 1963 - 2000," *Review of Economics and Statistics*, 2013, *95* (4), 1150–1165.

Davis, S.J., R.J. Faberman, and J. Haltiwanger, "Labor market flows in the cross-section and over time," *Journal of Monetary Economics*, 2012, *59* (1), 1–18.

Decker, R.A., J. Haltiwanger, R.S. Jarmin, and J. Miranda, "Declining Business Dynamism: What We Know and the Way Forward," *American Economic Review: Papers and Proceedings*, 2016, *106* (5), 203–207.

Demirer, M., "Production Function Estimation with Factor-Augmenting Technology: An Application to Markups," Working Paper, 2020.

Dey, M., S.N. Houseman, and A.E. Polivka, "Manufacturers' Outsourcing to Staffing Services," *ILR Review*, 2012, *65* (3), 533 – 559.

Diamond, P.A., "Wage Determination and Efficiency in Search Equilibrium," *Review of Economic Studies*, 1982, *49*, 217–227.

Dobbelaere, S. and J. Mairesse, "Panel Data Estimates of the Production Function and Product and Labor Market Imperfections," *Journal of Applied Econometrics*, January 2013, *28* (1), 1–46.

Dodini, S., M. Lovenheim, K. Salvanes, and A. Willén, "Monopsony, Skills, and Labor Market Concentration," Working Paper, 2020.

Doraszelski, U. and J. Jaumandreu, "Measuring the Bias of Technological Change," *Journal of Political Economy*, 2018, *126* (3), 1027–1084.

Edmond, C., V. Midrigan, and D.Y. Xu, "How Costly Are Markups?," Working Paper, 2021.

Eggertsson, G., J.A. Robbins, and E.G. Wold, "Kaldor's and Piketty's Facts: The Rise of Monopoly Power in the United States," *Journal of Monetary Economics*, 2021, *124*, S19–S38.

Elsby, M., B. Hobijn, and A. Şahin, "The Decline of the U.S. Labor Share," *Brookings Papers on Economic Activity*, 2013, pp. 1–52.

Flynn, Z., A. Gandhi, and J. Traina, "Identifying Market Power in Production Data," Working Paper, 2019.

Foster, L., J. Haltiwanger, and C. Syverson, "Selection on Productivity or Profitability?," *American Economic Review*, 2008, *98* (1), 394–425.

_ , _ , **and** _ , "The Slow Growth of New Plants: Learning about Demand?," *Economica*, 2016, *83* (329), 91–129.

_ , _ , **and C.J. Krizan**, "Aggregate Productivity Growth: Lessons from Microeconomic Evidence," in Charles R. Hulten, Edwin R. Dean, and Michael J. Harper, eds., *New Developments in Productivity Analysis*, Chicago: University of Chicago Press, 2001, chapter 8, pp. 303–372.

Gali, J., M. Gertler, and D.J. Lopez-Salido, "Markups, Gaps, and the Welfare Costs of Business Fluctuations," *Review of Economics and Statistics*, 2007, *89* (1), 44–59.

Gandhi, A., S. Navarro, and D.A. Rivers, "On the Identification of Gross Output Production Functions," *Journal of Political Economy*, 2020, *128* (8), 2973–3016.

Giroud, X. and H.M. Mueller, "Firms' Internal Networks and Local Economic Shocks," *American Economic Review*, 2019, *109* (10), 3617–3649.

Goldmanis, M., A. Hortacsu, C. Syverson, and O. Emre, "E-commerce and the Market Structure of Retail Industries," *Economic Journal*, 2010, *120* (545), 651–682.

Griliches, Z. and J. Mairesse, "Production Functions: The Search for Identification," in S. Strom, ed., *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, Cambridge: Cambridge University Press, 1998, pp. 169–203.

Hall, R.E., "Market Structure and Macroeconomic Fluctuations," *Brookings Papers on Economic Activity*, 1986, *2*, 285–322.

_ , "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy*, 1988, *96* (5), 921–947.

_ , "Measuring Factor Adjustment Costs," *Quarterly Journal of Economics*, 2004, *119*, 899–927.

Haltiwanger, J., R.S. Jarmin, and J. Miranda, "Who Creates Jobs? Small versus Large versus Young," *Review of Economics and Statistics*, 2013, *95* (2), 347–361.

Hsieh, C.-T. and P.J. Klenow, "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, 2009, *74* (4), 1403–1448.

Hu, Y., G. Huang, and Y. Sasaki, "Estimating production functions with robustness against errors in the proxy variables," *Journal of Econometrics*, 2020, *215* (2), 375–398.

Huckfeldt, C., "Understanding the Scarring Effects of Recessions," Working Paper, 2017.

Itskhoki, O. and B. Moll, "Optimal Development Policies with Financial Frictions," *Econometrica*, 2019, *87* (1), 139–173.

Jarosch, G., J.S. Nimczik, and I. Sorkin, "Granular Search, Market Structure, and Wages," Working Paper, 2021.

Kambourov, G. and I. Manovskii, "Occupational Specificity of Human Capital," *International Economic Review*, 2009, *50* (1), 63–115.

Karabarbounis, L., "The Labor Wedge: MRS vs. MPN," *Review of Economic Dynamics*, 2014, *17* (2), 206–223.

_ **and B. Neiman**, "The Global Decline of the Labor Share," *Quarterly Journal of Economics*, 2013, *129* (1), 61–103.

Kehrig, M., "The Cyclical Nature of the Productivity Distribution," Working Paper, 2015.

_ **and N. Vincent**, "The micro-level anatomy of the labor share decline," *Quarterly Journal of Economics*, 2021, *136* (2), 1031–1087.

Kim, R., "Price-Cost Markup Cyclicality: New Evidence and Implications," Working Paper, 2017.

Klette, T.J. and Z. Griliches, "The Inconsistency of Common Scale Estimators When Output Prices are Unobserved and Endogenous," *Journal of Applied Econometrics*, 1996, *11* (4), 343–361.

Lamadon, T., M. Mogstad, and B. Setzler, "Imperfect Competition, Compensating Differentials and Rent Sharing in the U.S. Labor Market," *American Economic Review*, 2022, *112* (1), 169–212.

Lazear, E.P. and J.R. Spletzer, "Hiring, Churn, and the Business Cycle," *American Economic Review*, 2012, *102* (3), 575–579.

Levinsohn, J.A. and A. Petrin, "Estimating Production Functions Using Inputs to Control for Unobservables," *Review of Economic Studies*, 2003, *70* (2), 317–340.

Lipsius, B., "Labor Market Concentration Does Not Explain the Falling Labor Share," Working Paper, 2018.

Loecker, J. De, "Recovering Markups from Production Data," *International Journal of Industrial Organization*, 2011, *29*, 350–355.

_ **and F. Warzynski**, "Markups and Firm-Level Export Status," *American Economic Review*, 2012, *102* (6), 2437–2471.

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

Macaluso, C., "Skill Remoteness and Post-Layoff Labor Market Outcomes," Working Paper, 2019.

Manning, A., *Monopsony in Motion*, Princeton University Press, 2003.

_ , "Imperfect Competition in the Labor Market," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, Vol. 4, North-Holland: Elsevier, 2011, chapter 11, pp. 973–1041.

_ **and B. Petrongolo**, "How Local Are Labor Markets? Evidence from a Spatial Job Search Model," *American Economic Review*, 2017, *107* (10), 2877–2907.

Marinescu, O. and R. Rathelot, "Mismatch Unemployment and the Geography of Job Search," *American Economic Journal: Macroeconomics*, 2018, *10* (3), 42–70.

McDonald, I.M. and R.M. Solow, "Wage Bargaining and Employment," *American Economic Review*, 1981, *71* (5), 896–908.

Melitz, M.J. and G.I.P. Ottaviano, "Market Size, Trade, and Productivity," *Review of Economic Studies*, 2008, *75* (1), 295–316.

Morlacco, M., "Market Power in Input Markets: Theory and Evidence from French Manufacturing," Working Paper, 2020.

Mortensen, D., "How Monopsonistic is the (Danish) Labor Market?," in P. Aghion, R. Frydman, J. Stiglitz, and M. Woodford, eds., *Knowledge, Information and Expectations in Modern Macroeconomics*, Princeton, NJ: Princeton University Press, 2003.

Mortensen, D.T. and C.A. Pissarides, "Job Creation and Job Destruction in the Theory of Unemployment," *Review of Economic Studies*, 1994, *61*, 397–415.

Olley, S. and A. Pakes, "The Dynamics of Productivity in the Telecommunications Industry," *Econometrica*, 1996, *64*, 1263–1298.

Pesaran, M. H. and R. Smith, "Estimating Long-Run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics*, 1995, *68* (1), 79–113.

Posner, E.A., G. Weyl, and S. Naidu, "Antitrust Remedies for Labor Market Power," *Harvard Law Review*, 2018, *132* (536).

Prager, E. and M. Schmitt, "Employer Consolidation and Wages: Evidence from Hospitals," *American Economic Review*, 2021, *111* (2), 397–427.

Raval, D., "Testing the Production Approach to Markup Estimation," Working Paper, 2020.

Reenen, J. Van, "The Creation and Capture of Rents: Wages and Innovation in a Panel of U.K. Companies," *Quarterly Journal of Economics*, 1996, *111* (1), 195–226.

Rinz, K., "Labor market Concentration, Earnings Inequality, and Earnings Mobility," Working Paper, 2018.

_ , "Labor Market Concentration, Earnings, and Inequality," *Journal of Human Resources*, 2020, pp. 0219–10025R1.

Robinson, J., *The Economics of Imperfect Competition*, Palgrave Macmillan, 1933.

Rossi-Hansberg, E., P.-D. Sarte, and N. Trachter, "Diverging Trends in National and Local Concentration," *NBER Macroeconomics Annual*, 2020, *35*.

Salop, S., "Monopolistic Competition with Outside Goods," *Bell Journal of Economics*, 1979, *10*, 141–156.

Schubert, G., A. Stansbury, and B. Taska, "Employer Concentration and Outside Options," Working Paper, 2021.

Seegmiller, B., "Valuing Labor Market Power: The Role of Productivity Advantages," Working Paper, 2021.

Sokolova, A. and T. Sorensen, "Monopsony in Labor Markets: A Meta-Analysis," *ILR Review*, 2020, *74* (1), 27–55.

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

Starr, E.P., J.J. Prescott, and N.D. Bishara, "Noncompete Agreements in the US Labor Force," *Journal of Law and Economics*, 2021, *64* (1), 53—84.

Syverson, C., "Market Structure and Productivity: A Concrete Example," *Journal of Political Economy*, 2004, *112* (6), 1181–1222.

_ , "Product Substitutability and Productivity Dispersion," *Review of Economics and Statistics*, 2004, *86* (2), 534–550.

_ , "Macroeconomics and Market Power: Facts, Potential Explanations and Open Questions," Technical Report, Brookings 2019.

Traina, J., "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements," Working Paper, 2018.

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

A Details on markdown estimation

A.1 Derivations

In this appendix, we formalize our arguments in the main text. In particular, we show that retrieving output elasticities and revenue shares are sufficient in order to estimate markdowns. To see this, we start with the cost minimization problem of a firm. In general, we have:

$$\min_{\mathbf{X}_{it}\in\mathbb{R}_{+}^{K}} \sum_{k=1}^{K} V_{it}^{k}(X_{it}^{k})X_{it}^{k} + \Phi_{t}^{k}\left(X_{i}^{k}, X_{it-1}^{k}\right) \quad \text{s.t.} \quad F(\mathbf{X}_{it};\omega_{it}) \geq Q_{it} \tag{19}$$

where $\mathbf{X}_{it} = \left(X_{it}^{1}, \ldots, X_{it}^{K}\right)'$ is the firm's vector of $K > 1$ production inputs with prices $\{V_{it}^{k}\}_{k=1}^{K}$. Furthermore, ω_{it} denotes a firm i's productivity level at time t, whereas a firm's production technology is denoted by $F(\mathbf{X}_{it};\omega_{it})$. Adjustment costs for some input k are captured by the function $\Phi_{t}^{k}(\cdot,\cdot)$.

To derive markdowns, we start with the insight by Hall (1986) that the wedge between a flexible input's output elasticity and its revenue share must reflect a firm's output market power (or its mark*up*; defined as its output price over marginal cost of production).

LEMMA 1. Let Assumption **I** hold. Furthermore, let Assumptions **II** – **VI** hold for some input k'. Then a firm i's markup satisfies:

$$\begin{aligned}\mu_{it} &= \frac{\partial F(\mathbf{X}_{it};\omega_{it})}{\partial X_{it}^{k'}}\frac{X_{it}^{k'}}{Q_{it}} \cdot \left(\frac{V_{it}^{k'}X_{it}^{k'}}{P_{it}Q_{it}}\right)^{-1} \\ &\equiv \frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}}\end{aligned} \tag{20}$$

Proof. Under the stated assumptions, the first-order condition for any flexible input k', associated with cost-minimization problem (19), satisfies:

$$V_{it}^{k'} = \lambda_{it}\frac{\partial F(\mathbf{X}_{it};\omega_{it})}{\partial X_{it}^{k'}}$$

where λ_{it} is the Lagrangian multiplier associated with the cost-minimization problem in (19). This shadow value of total variable costs is also known as firm i's marginal cost of production. The above equality can easily be manipulated to:

$$\frac{V_{it}^{k'} X_{it}^{k'}}{P_{it} Q_{it}} = \frac{\lambda_{it}}{P_{it}} \frac{\partial F(\mathbf{X}_{it}; \omega_{it})}{\partial X_{it}^{k'}} \frac{X_{it}^{k'}}{Q_{it}}$$

where P_{it} denotes a firm's price for its output good. Then, we get the expression for a firm i's markup $\mu_{it} = \frac{P_{it}}{\lambda_{it}}$ at time t:

$$\mu_{it} = \frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}} \tag{21}$$

where $\theta_{it}^{k'} \equiv \frac{\partial F(\mathbf{X}_{it}; \omega_{it})}{\partial X_{it}^{k'}} \frac{X_{it}^{k'}}{Q_{it}}$ and $\alpha_{it}^{k'} \equiv \frac{V_{it}^{k'} X_{it}^{k'}}{P_{it} Q_{it}}$. Thus, a firm's markup is equal to the wedge between the output elasticity and the revenue share of some input k'. Note that the existence of *only one* flexible input k' that satisfies Assumptions **II** – **VI** is sufficient to establish this result. □

Given this lemma, we can prove the main result of Proposition 1.

Proof of Proposition 1. Without loss of generality, consider the following *conditional* cost-minimization problem:

$$\min_{\ell_{it} \geq 0} w_{it}(\ell_{it}) \ell_{it} \;\; \text{s.t.} \;\; F(\ell_{it}, \mathbf{X}_{-\ell,it}^{*}; \omega_{it}) \geq Q_{it},$$

where $\mathbf{X}_{-\ell,it}^{*}$ denotes the vector of optimized inputs with the exception of labor ℓ_{it}. The associated optimality condition with Lagrangian multiplier λ_{it} can be characterized as:

$$\left[\frac{w_{it}'(\ell_{it}) \ell_{it}}{w_{it}(\ell_{it})} + 1 \right] = \lambda_{it} \cdot \frac{\frac{\partial F(\ell_{it}, \mathbf{X}_{-\ell,it}^{*}; \omega_{it})}{\partial \ell_{it}}}{w_{it}(\ell_{it})},$$

which we can rearrange as:

$$\begin{aligned} \left[\frac{w_{it}'(\ell_{it}) \ell_{it}}{w_{it}(\ell_{it})} + 1 \right] &\equiv \varepsilon_{S}^{-1}(\ell_{it}) + 1 \\ &= \frac{\lambda_{it}}{P_{it}} \cdot \frac{\partial F(\ell_{it}, \mathbf{X}_{-\ell,it}^{*}; \omega_{it})}{\partial \ell_{it}} \frac{\ell_{it}}{Q_{it}} \cdot \frac{P_{it} Q_{it}}{w_{it}(\ell_{it}) \ell_{it}} \\ &\equiv \mu_{it}^{-1} \cdot \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}}. \end{aligned} \tag{22}$$

Given our insight on a firm's markdown, we must have:

$$\frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} = \nu_{it} \cdot \mu_{it}. \tag{23}$$

Then, the result follows immediately from lemma 1. Hence, we have:

$$\nu_{it} = \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} \cdot \left(\frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}} \right)^{-1}, \tag{24}$$

which is what we wanted to show. □

Note that the result from the main text follows immediately from the above proposition whenever material inputs are assumed to be flexible. The revenue shares α_{it}^{ℓ} and α_{it}^{M} can be directly constructed from the data. To obtain markdowns (and markups), it is sufficient to estimate output elasticities only. Therefore, we need to estimate production functions.

A.2 GMM-IV estimation procedure

In the following, we will provide more details on how we obtain output elasticities. To do so, we will follow the "proxy variable" literature on production function estimation (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; De Loecker and Warzynski, 2012; Ackerberg, Caves and Frazer, 2015).

Let the production function be given by:

$$Q_{it} = F(\mathbf{V}_{it}, \mathbf{K}_{it}; \omega_{it}),$$

where we categorize inputs as flexible or nonflexible inputs, i.e., $\mathbf{X}_{it}' = (\mathbf{V}_{it}', \mathbf{K}_{it}')$. In particular, we have:

$$\begin{aligned} \mathbf{V}_{it} &= (X_{it}^{1}, \ldots, X_{it}^{V})' \\ \mathbf{K}_{it} &= (X_{it}^{V+1}, \ldots, X_{it}^{K})', \end{aligned}$$

where the first $V \geq 1$ inputs are flexible and the latter $K - V$ inputs are not fully flexible. In particular, $\mathbf{K}_{it}$ is a state variable when choosing the inputs $\mathbf{V}_{it}$. Furthermore, ω_{it} denotes

a firm's productivity. In particular, suppose that $X_{it}^1 = M_{it}$ are material inputs.

To account for measurement error, we assume that *observed* logged output satisfies $y_{it} = \ln(Q_{it}) + \varepsilon_{it}$; i.e., measurement error enters production in a multiplicative fashion. Note that the error term ε_{it} is *not* observed by firms when they have to make their optimal input decisions. Given our econometric assumptions 1–5, we can write:

$$y_{it} = f(\mathbf{v}_{it}, \mathbf{k}_{it}; \boldsymbol{\beta}) + \omega_{it} + \varepsilon_{it},$$

where $f(\mathbf{v}_{it}, \mathbf{k}_{it}; \boldsymbol{\beta}) = \ln\left(F(\mathbf{V}_{it}, \mathbf{K}_{it}; \boldsymbol{\beta})\right)$, and $\mathbf{v}_{it}$ and $\mathbf{k}_{it}$ denote componentwise natural log transformations of $\mathbf{V}_{it}$ and $\mathbf{K}_{it}$, respectively. Firm-level productivities ω_{it} are not observed by the econometrician, but are observable for firms themselves.

Unobservable productivity is the main cause of endogeneity concerns in our estimation procedure. To deal with this, we use the insight of Levinsohn and Petrin (2003). Under Assumptions 4 and 5, material demand $\ln(X_{it}^1) = m_{it}$ can be used to proxy for productivity. Note that firms choose flexible inputs given the state $\mathbf{K}_{it}$, idiosyncratic productivity ω_{it}, and some controls that can influence their decisions $\mathbf{c}_{it}$ (e.g., input prices):

$$m_{it} = m_t(\omega_{it}; \mathbf{k}_{it}, \mathbf{c}_{it}),$$

where the vector $\mathbf{c}_{it}$ denotes any additional, observable variables that can affect a plant's optimal demand for material inputs.[52] The above mapping for materials is invertible in productivity ω_{it} by Assumption 4. Following Levinsohn and Petrin (2003), a sufficient condition for invertibility is:

$$D^M = \left| \frac{\partial \mathbf{V}_{it}(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}} \quad \mathbf{H}_{2,V}^F(\mathbf{K}_{it}, \omega_{it}) \quad \ldots \quad \mathbf{H}_{V,V}^F(\mathbf{K}_{it}, \omega_{it}) \right| > 0,$$

where $\frac{\partial \mathbf{V}_{it}(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}} = \left(\frac{\partial X_{it}^1(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}}, \ldots, \frac{\partial X_{it}^V(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}} \right)'$ and $\mathbf{H}_{r,V}^F(\mathbf{K}_{it}, \omega_{it}) = \left(\frac{\partial F(\mathbf{V}_{it}, \mathbf{K}_{it}, \omega_{it})}{\partial X_{it}^r \partial X_{it}^1}, \ldots, \frac{\partial F(\mathbf{V}_{it}, \mathbf{K}_{it}, \omega_{it})}{\partial X_{it}^r \partial X_{it}^V} \right)'$ is the r^{th} column of the Hessian matrix for $F(\cdot, \mathbf{K}_{it}; \omega_{it})$ evaluated at $\mathbf{V}_{it} \in \mathbb{R}_+^V$.

[52] In the empirical implementation, $\mathbf{c}_{it}$ contains only a set of year fixed effects. Industry fixed effects are not required whenever production technology parameters are estimated industry-by-industry. However, the used methodology is flexible enough to account for other observables.

Under this assumption, the material input demand function is monotonic in productivity ω_{it}.[53] Then there exists some function $h_t(\cdot; \mathbf{k}_{it}, \mathbf{c}_{it})$ so that:

$$\omega_{it} = h_t(m_{it}; \mathbf{k}_{it}, \mathbf{c}_{it}).$$

As a result, production y_{it} can be written in terms of observables only:

$$\begin{aligned} y_{it} &= f(\mathbf{v}_{it}, \mathbf{k}_{it}; \boldsymbol{\beta}) + h_t(m_{it}; \mathbf{k}_{it}, \mathbf{c}_{it}) + \varepsilon_{it} \\ &= \phi_t(\mathbf{v}_{it}, \mathbf{k}_{it}, \mathbf{c}_{it}) + \varepsilon_{it} \\ &= \varphi_{it} + \varepsilon_{it}. \end{aligned}$$

Estimating the production technology parameters $\boldsymbol{\beta}$ is done in a three-stage fashion, which is in a similar spirit to Ackerberg, Caves and Frazer (2015). To implement our estimation procedure, we set $\mathbf{v}_{it} = m_{it}$, $\mathbf{k}_{it} = (k_{it}, \ell_{it}, e_{it})'$ and $\mathbf{c}_{it} = (d_{i,1}, \ldots, d_{i,T})'$, where $d_{i,t}$ is a fixed effect for a specific year t. Even though we will mainly focus on translog production functions, we also occasionally report results for Cobb-Douglas specifications.

STEP 1. NONPARAMETRIC ESTIMATION OF φ_{it} AND ε_{it}.

First, we estimate φ_{it} and ε_{it} nonparametrically by approximating y_{it} with a third-degree polynomial in $\tilde{\mathbf{x}}_{it} = (k_{it}, \ell_{it}, m_{it}, e_{it})'$ with interaction terms. In the case of translog production, we have:

$$\mathbf{x}_{it} = (k_{it}, \ell_{it}, m_{it}, e_{it}, k_{it}\ell_{it}, k_{it}m_{it}, k_{it}e_{it}, \ell_{it}m_{it}, \ell_{it}e_{it}, m_{it}e_{it}, k_{it}^2, \ell_{it}^2, m_{it}^2, e_{it}^2)'.$$

Let its fitted values and residuals be denoted by $\widehat{\varphi}_{it}$ and $\widehat{\varepsilon}_{it}$, respectively. These residuals are then interpreted as measurement error in observed output.

STEP 2. CONSTRUCTION OF INNOVATIONS ξ_{it} TO PRODUCTIVITY ω_{it}.

By Assumption 3, idiosyncratic productivity ω_{it} is Markovian; thus, its expected value is only a function of its lagged value. As a result, we have $\omega_{it} = g_t(\omega_{it-1}) + \xi_{it}$. Then, productivity is approximated in the data by:

[53] This follows from standard arguments for comparative statics under multiple inputs. We then apply Cramer's rule to arrive at the stated condition. Levinsohn and Petrin (2003) show a similar result for $V = 2$ in their Appendix A. In a nutshell, Assumption 5 imposes a set of regularity conditions on the cross-derivatives of the production function in $\mathbf{V}_{it}$ which are fairly mild.

$$\omega_{it}(\boldsymbol{\beta}) = \widehat{\varphi}_{it} - f(\mathbf{x}_{it}; \boldsymbol{\beta}).$$

Then we approximate $g_t(.)$ with a $\mathcal{P}^{\text{th}}$ order polynomial in its argument:

$$\begin{aligned}\omega_{it}(\boldsymbol{\beta}) &= \Omega_{it-1}(\boldsymbol{\beta})'\rho(\boldsymbol{\beta}) + \xi_{it} \\ &= \sum_{p=0}^{\mathcal{P}} \rho_p \omega_{it-1}^p(\boldsymbol{\beta}) + \xi_{it},\end{aligned}$$

where we follow De Loecker and Warzynski (2012) and set $\mathcal{P} = 3$. Thus, the innovations to productivity can be constructed as a function of $\boldsymbol{\beta}$ through:

$$\xi_{it}(\boldsymbol{\beta}) = \omega_{it}(\boldsymbol{\beta}) - \Omega_{it-1}(\boldsymbol{\beta})'\widehat{\rho}(\boldsymbol{\beta}).$$

The estimates $\widehat{\rho}(\boldsymbol{\beta}) = (\{\widehat{\rho}_p\}_{p=1}^{\mathcal{P}})'$ are simply obtained by running a least squares regression of $\Omega_{it-1}(\boldsymbol{\beta})$ on $\omega_{it}(\boldsymbol{\beta})$.

STEP 3. GMM-IV ESTIMATION OF $\boldsymbol{\beta}$.

By Assumption 2, capital is predetermined at time t, as a firm chooses it one period ahead. As a result, it is safe to assume that k_{it} is orthogonal to the innovation $\xi_{it}(\boldsymbol{\beta})$. Similarly, firms cannot observe the string of future innovations to their productivity. As a result, current input decisions (with the exception of investment in capital) must be orthogonal to shocks to their idiosyncratic productivity in the future. Define the instrument $\mathbf{z}_{it} \in \mathbb{R}^Z$ as the vector that contains one-period lagged values of every polynomial term containing ℓ_{it}, m_{it}, and e_{it} in the production technology $f(\mathbf{x}_{it}; \boldsymbol{\beta})$, but with capital preserved at its current value k_{it}. Then, we define the following system of moment conditions to identify $\boldsymbol{\beta} \in \mathbb{R}^Z$:

$$\mathbb{E}\left(\xi_{it}(\boldsymbol{\beta})\mathbf{z}_{it}\right) = \mathbf{0}_{Z\times 1}. \tag{25}$$

By construction, this system of equations defines a set of exogeneity conditions. Lagged inputs are used to instrument for current period inputs. To validate this identification strategy, we need to argue that the moment conditions in (25) also satisfy rank conditions. Our focus lies on material inputs, so we will pay particular attention to this specific input. For lagged material inputs to be a valid instrument for current material inputs, m_{it} and m_{it-1} need to be correlated. A sufficient condition would be that input prices for material inputs

are persistent over time. In fact, Atalay (2014) finds empirical evidence for this using data from the Census of Manufactures.

To obtain $\boldsymbol{\beta}$, we rely on the minimization of a quadratic loss function, which is standard in GMM estimation.[54] Thus, we get:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^Z}{\arg\min} \sum_{m=1}^{Z} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{it}(\boldsymbol{\beta}) z_{it}^{m} \right)^2 ,$$

where we have $\mathbf{z}_{it} = (z_{it}^1, \ldots, z_{it}^Z)'$.

CONSTRUCTING MARKUPS AFTER OBTAINING ESTIMATES $\widehat{\boldsymbol{\beta}}$. In general, output elasticities with respect to material inputs can depend on the level of *all* inputs, whether that level be flexible or predetermined. This implies that $\theta_{it}^M = \theta_M^{j(i)}(\tilde{\mathbf{x}}_{it}; \boldsymbol{\beta})$. Following the estimation procedure by De Loecker and Warzynski (2012), we can furthermore correct for measurement error ε_{it} in logged output. This is particularly important for data in the ASM and CM. Output prices are not available at the firm level, so output levels are obtained by deflating revenues adjusted for inventories. Unfortunately, the deflators used in the NBER-CES Manufacturing Industry Database are only available at the industry level. This causes an unavoidable bias in measuring real output.

However, De Loecker and Warzynski (2012) mention that some of the concern about this bias can be taken care of with the correction term ε_{it}. By construction, any unobserved variation in output prices orthogonal to a firm's inputs will be absorbed by the measurement-error correction term. In addition, if pricing decisions are correlated with a plant's productivity, then this specific variation will be controlled for as well, through the use of a proxy for productivity. Then, markups are constructed as:

$$\begin{aligned} \widehat{\mu}_{it} &= \widehat{\theta}_{it}^M \left(\frac{\mathtt{vm}_{it}}{\mathtt{tvs}_{it}/\widehat{\varepsilon}_{it}} \right)^{-1} \\ &= \theta_M^{j(i)}(\tilde{\mathbf{x}}_{it}; \widehat{\boldsymbol{\beta}}) \left(\frac{\mathtt{vm}_{it}}{\mathtt{tvs}_{it}/\exp(\widehat{\varepsilon}_{it})} \right)^{-1} , \end{aligned} \tag{26}$$

where $\mathtt{vm}_{it}$ and $\mathtt{tvs}_{it}$ denote a plant i's total expenditure on intermediate inputs and total value of shipments in year t, respectively. Production technologies do not differ over time,

[54] By construction, the number of parameters in $\boldsymbol{\beta}$ is equal to the amount of identifying moments. This case of "just identification" renders the specification of a weighting matrix useless.

but are allowed to vary across industries by assumption 3.[55]

To construct output elasticities explicitly, we need to take a stance on the production function. In the following, we demonstrate how to obtain output elasticities in the case of translog production.[56] Our preferred specification assumes that production is translog, for two reasons. First, the translog specification is a second-order log approximation to *any* arbitrary, differentiable production function. In fact, the Cobb-Douglas setup is nested within our translog specification. Second, output elasticities are allowed to vary with the level of any input under the translog specification. This implies that markups and markdowns have two sources of time variation: 1) time-varying output elasticities and 2) input revenue shares.

TRANSLOG PRODUCTION. Assumption 3 under translog production implies:

$$\begin{aligned} f(\mathbf{x}_{it};\boldsymbol{\beta}) = {} & \beta_K k_{it} + \beta_L \ell_{it} + \beta_M m_{it} + \beta_E e_{it} \\ & + \beta_{KL} k_{it}\ell_{it} + \beta_{KM} k_{it} m_{it} + \beta_{KE} k_{it} e_{it} + \beta_{LM} \ell_{it} m_{it} + \beta_{LE} \ell_{it} e_{it} + \beta_{ME} m_{it} e_{it} \\ & + \beta_{KK} k_{it}^2 + \beta_{LL} \ell_{it}^2 + \beta_{MM} m_{it}^2 + \beta_{EE} e_{it}^2. \end{aligned}$$

Assuming that capital is chosen one period ahead, the instrument vector becomes:

$$\begin{aligned} \mathbf{z}_{it} = \Big(& k_{it}, \ell_{it-1}, m_{it-1}, e_{it-1}, k_{it}\ell_{it-1}, k_{it} m_{it-1}, k_{it} e_{it-1}, \ell_{it-1} m_{it-1}, \ell_{it-1} e_{it-1}, m_{it-1} e_{it-1}, \\ & k_{it}^2, \ell_{it-1}^2, m_{it-1}^2, e_{it-1}^2 \Big)', \end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{14}$ is estimated for each industry j. Note that the number of parameters increases exponentially whenever more inputs are considered.[57] Markdowns are then em-

[55] Note that this assumption can be relaxed by estimating, for example, time-varying Cobb-Douglas parameters. This is easily done by restricting the estimation sample to repeated cross-sections in a subset of years. Theoretically, this should be possible for the translog case as well, but the amount of cross-sectional variation in these subsamples might not be sufficient to identify all parameters properly.

[56] Under Cobb-Douglas production, output elasticities are equal to their respective production coefficients.

[57] With a translog production function with K inputs, there are K linear terms, K quadratic components and $\binom{K}{2}$ unique input pairs. Thus, there are a total of $2K + \binom{K}{2} = \frac{K(K+3)}{2}$ terms.

pirically implemented through:

$$\widehat{\nu}_{it}^{\text{TL}} = \widehat{\theta}_{\ell}^{j(i)}(\tilde{\mathbf{x}}_{it};\widehat{\boldsymbol{\beta}}) \left(\frac{\texttt{sw}_{it}}{\texttt{tvs}_{it}}\right)^{-1} \left[\widehat{\theta}_{M}^{j(i)}(\tilde{\mathbf{x}}_{it};\widehat{\boldsymbol{\beta}}) \left(\frac{\texttt{vm}_{it}}{\texttt{tvs}_{it}/\exp(\widehat{\epsilon}_{it})}\right)^{-1}\right]^{-1}$$

s.t.

$$\widehat{\theta}_{\ell}^{j(i)}(\tilde{\mathbf{x}}_{it};\widehat{\boldsymbol{\beta}}) = \widehat{\beta}_{L}^{j(i)} + \widehat{\beta}_{KL}^{j(i)} k_{it} + \widehat{\beta}_{LM}^{j(i)} m_{it} + \widehat{\beta}_{LE}^{j(i)} e_{it} + 2\widehat{\beta}_{LL}^{j(i)} \ell_{it}$$
$$\widehat{\theta}_{M}^{j(i)}(\tilde{\mathbf{x}}_{it};\widehat{\boldsymbol{\beta}}) = \widehat{\beta}_{M}^{j(i)} + \widehat{\beta}_{KM}^{j(i)} k_{it} + \widehat{\beta}_{LM}^{j(i)} \ell_{it} + \widehat{\beta}_{ME}^{j(i)} e_{it} + 2\widehat{\beta}_{MM}^{j(i)} m_{it}.$$

B Aggregate markdowns

B.1 Aggregation of micro-level markdowns

Proof of Proposition 2. Whenever Assumptions **I – VI** are satisfied and Assumptions **II – VI** apply specifically to material inputs, then we show in lemma 1 that markups can be characterized as:

$$\begin{aligned}\mu_{it} &= \frac{\theta_{it}^{M}}{\alpha_{it}^{M}} \\ &= \theta_{it}^{M} \cdot \frac{P_{it}Q_{it}}{P_{t}^{M}M_{it}}.\end{aligned} \tag{27}$$

Similar to Edmond, Midrigan and Xu (2021), we define the **aggregate markup** as the wedge between the aggregate output elasticity of some flexible input and its revenue share. Under the assumption of material inputs being flexible, Equation (27) also holds in the aggregate; i.e., we have:

$$\mathcal{M}_{t} \equiv \theta_{t}^{M} \cdot \frac{P_{t}Y_{t}}{P_{t}^{M}M_{t}}, \tag{28}$$

where we dropped the indices for local markets (j, ℓ) for simplicity. Substituting out the price for material inputs P_{t}^{M} from (28) into (27), we obtain:

$$\mu_{it} = \frac{\theta_{it}^{M}}{\theta_{t}^{M}} \cdot \frac{P_{it}Q_{it}}{P_{t}Y_{t}} \cdot \frac{M_{t}}{M_{it}} \cdot \mathcal{M}_{t}.$$

Then, we sum across firms and rearrange to derive the aggregate markup:

$$\mathcal{M}_t = \left(\sum_{i \in F_t} \frac{\theta_{it}^M}{\theta_t^M} \cdot s_{it} \cdot \mu_{it}^{-1} \right)^{-1}, \tag{29}$$

where $s_{it} \equiv \frac{P_{it}Q_{it}}{P_tY_t}$ denotes a firm i's revenue share relative to the aggregate and we used the definition for aggregate materials $M_t = \sum_{i \in F_t} M_{it}$. Whenever production technologies are Cobb-Douglas, we have $\theta_{it}^M = \theta_t^M$ for each $i \in F_t$. Then, the aggregate markup is simply a revenue-weighted harmonic average of firm-level markups.

We use a similar insight to derive the **aggregate markdown** $\mathcal{V}_t$. Whenever Assumptions **II** and **IV – VI** hold for labor, the wedge between the output elasticity of labor and its revenue share for a firm i must reflect market power in either output or labor markets. We showed this explicitly in Proposition 1. Therefore, we have:

$$\nu_{it}\mu_{it} = \theta_{it}^L \cdot \frac{P_{it}Q_{it}}{w_{it}\ell_{it}}. \tag{30}$$

Rearranging for a firm i's wage bill and summing across firms, it follows that:

$$\begin{aligned} \sum_{i \in F_t} w_{it}\ell_{it} &= w_t L_t \\ &= P_t Y_t \cdot \sum_{i \in F_t} \theta_{it}^L \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1}, \end{aligned}$$

where the first equality follows from definition of the aggregate wage bill. We define the aggregate markdown $\mathcal{V}_t$ as that part of the wedge between the aggregate output elasticity of labor and the aggregate labor share that is not due to markups. Then, by definition, we have:

$$\mathcal{V}_t \cdot \mathcal{M}_t = \theta_t^L \cdot \frac{P_t Y_t}{w_t L_t}. \tag{31}$$

Using our previous results, we then get:

$$\mathcal{V}_t \cdot \mathcal{M}_t = \theta_t^L \cdot \left(\sum_{i \in F_t} \theta_{it}^L \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1}$$
$$= \left(\sum_{i \in F_t} \frac{\theta_{it}^L}{\theta_t^L} \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1}.$$

Apply Expression (29) for the aggregate markup and we obtain an expression for the aggregate markdown:

$$\mathcal{V}_t = \frac{\left(\sum_{i \in F_t} \frac{\theta_{it}^L}{\theta_t^L} \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1}}{\left(\sum_{i \in F_t} \frac{\theta_{it}^M}{\theta_t^M} \cdot s_{it} \cdot \mu_{it}^{-1} \right)^{-1}}. \tag{32}$$

A special case is whenever each firm i has a Cobb-Douglas technology. Then, we get:

$$\mathcal{V}_t = \frac{\left(\sum_{i \in F_t} s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1}}{\left(\sum_{i \in F_t} s_{it} \cdot \mu_{it}^{-1} \right)^{-1}}, \tag{33}$$

which amounts to a ratio of sales-weighted harmonic averages. □

B.2 Aggregate markdowns and employment concentration

We calculate the cross-sectional correlation (across local labor markets) between the aggregate markdown $\mathcal{V}_{jlt}$ and employment concentration HHI_{jlt}. The results for each census year can be found in Table VI.

Our results indicate that the correlations between labor market power and employment concentration are low. In fact, these correlations are close to zero, and for some census years even negative. When we take the average cross-market correlation across census years, we basically find a value of zero. Our conclusions do not change whenever we base our results on rank correlations (e.g., Spearman's ρ or Kendall's τ) instead. In Section 4, we found that the aggregate markdown in the spirit of Rossi-Hansberg, Sarte and Trachter (2020), calculated with Equation (16), displayed a relatively strong correlation over time, with local concentration LOCAL_t. However, our results in Table VI indicate that the cross-sectional correlations are also fairly weak under this specification.

Table VI: The correlation between employment HHIs and aggregate markdown across local labor markets is close to zero.†

Specification: TRANSLOG MARKDOWNS		
YEAR	$\rho(\mathcal{V}_{jlt}, \text{HHI}_{jlt})$	$\rho(\mathcal{V}^{\text{RHST}}_{jlt}, \text{HHI}_{jlt})$
1977	0.01656	0.00017
1982	0.00779	0.03593
1987	−0.00164	0.03528
1992	−0.01491	0.03305
1997	0.00097	0.01567
2002	0.00385	0.01444
2007	0.00440	0.00425
2012	−0.01964	0.01108
AVERAGE	−0.00033	0.01873

†Markdowns are estimated under the assumption of a translog specification for gross output. Cross-market correlations are calculated at the 3-digit NAICS county level for each census year. Aggregate markdowns are calculated according to formulas (14) and (16), whereas HHI_{jlt} denotes a market's employment Herfindahl-Hirschman Index. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

C Labor adjustment costs

In this appendix, we show that the wedge between the marginal revenue product of labor and the wage is no longer reflective of only labor market power whenever labor adjustment costs are present. This is not a trivial result, since a firm's profit maximization problem becomes dynamic when labor is subject to costly adjustments. Intuitively, this is because labor adjustment costs depend on the level of labor in the previous period. If these adjustment costs take a quadratic form, however, it is possible to "correct" our initial estimates for markdowns. When we apply these correction terms to our estimates, we obtain measures for markdowns that are only reflective of monopsony forces and not of labor adjustment costs. In the end, we find that these correction terms are quantitatively small.

The proposition below shows that labor adjustment costs can also drive a wedge between

marginal revenue products of labor and wages. Nevertheless, we can identify the "monopsony" component whenever these adjustment costs take a quadratic form.

PROPOSITION 3. Let $\mathbf{z}$ denote a firm's set of stochastic state variables and suppose revenue, labor adjustment cost, and wage schedule functions are differentiable. Then, a firm's wedge between its MRPL and wage satisfies:

$$\frac{R'(\ell^*)}{w(\ell^*)} = \left(\varepsilon_S^{-1} + 1\right) + \mathcal{A}(\ell^*, \ell_{-1}),$$

where $\mathcal{A}(\ell^*, \ell_{-1})$ equals zero whenever labor adjustment costs are absent. If, in addition, a firm is subject to convex labor adjustment costs of the form $\Phi(\ell, \ell_{-1}) = \frac{\gamma}{2}\ell\left(\frac{\ell-\ell_{-1}}{\ell_{-1}}\right)^2$ for $\gamma \geq 0$ and it discounts future profits at the rate of $\beta \in [0, 1]$, then a firm's monopsony power can be characterized as:

$$\varepsilon_S^{-1} + 1 = \frac{\frac{R'(\ell^*)}{w(\ell^*)} - \gamma \cdot \left(g_\ell(1 + g_\ell) - \beta \mathbb{E}_{\mathbf{z}'}\left[g_{\ell'}(1 + g_{\ell'})(1 + g_{\mathtt{sw}'})|\mathbf{z}\right]\right)}{1 + \frac{\gamma}{2} g_\ell^2}, \tag{34}$$

where g_ℓ, $g_{\ell'}$, and $g_{\mathtt{sw}'}$ denote current labor growth, future labor growth, and future wage bill growth, respectively.

Proof. We will consider environments in which revenue, labor adjustment costs, and wage schedules are continuously differentiable (at least in labor). Furthermore, we will restrict our attention to convex adjustment costs in labor, but we do allow for dynamic considerations (i.e., adjustment costs in labor are allowed to depend on the stock of labor in the previous period, denoted by ℓ_{-1}). Then, consider a firm's *dynamic* profit maximization problem:

$$v(\ell_{-1}; \mathbf{z}) = \max_{\ell \geq 0} R(\ell; \mathbf{z}) - w(\ell) \cdot \ell - w(\ell) \cdot \Phi(\ell, \ell_{-1}) + \beta \cdot \mathbb{E}_{\mathbf{z}'}\left[v(\ell; \mathbf{z}')|\mathbf{z}\right], \tag{35}$$

where $\Phi(\ell, \ell_{-1})$ denotes a firm's adjustment cost (in real terms) whenever it wants to change its stock of labor to $\ell \neq \ell_{-1}$, and $\beta \in [0, 1]$ is its discount factor. We will assume that the adjustment cost function is homogeneous of degree one and continuously differentiable in both arguments. Furthermore, we have that $\Phi(\ell, \ell_{-1}) > 0$ for $\ell \neq \ell_{-1}$ and zero otherwise. Similar to before, we denote the revenue function by $R(\ell; \mathbf{z}) \equiv \text{rev}(\ell; \mathbf{X}^*_{-\ell}(\ell), \mathbf{z})$, where $\mathbf{z}$ denotes a firm's (possibly stochastic) state variable, e.g. productivity. Given this setup, a

firm's optimal choice is characterized by its first-order condition:

$$\begin{aligned} R'(\ell) &= w'(\ell)\ell + w(\ell) + w(\ell)\cdot\Phi_1(\ell,\ell_{-1}) + w'(\ell)\cdot\Phi(\ell,\ell_{-1}) - \beta\cdot\mathbb{E}_{\mathbf{z}'}\left[v'(\ell)|\mathbf{z}\right] \\ &= w'(\ell)\ell + w(\ell) + w(\ell)\cdot\Phi_1(\ell,\ell_{-1}) + w'(\ell)\cdot\Phi(\ell,\ell_{-1}) + \beta\cdot\mathbb{E}_{\mathbf{z}'}\left[\Phi_2(\ell',\ell)w(\ell')|\mathbf{z}\right], \end{aligned}$$

where we applied the envelope theorem in the last equality. This can be rearranged to end up with an expression for a firm's markdown:

$$\begin{aligned} \nu &\equiv \frac{R'(\ell)}{w(\ell)} \\ &= \varepsilon_S^{-1} + 1 + \Phi_1(\ell,\ell_{-1}) + \frac{\Phi(\ell,\ell_{-1})}{\ell}\varepsilon_S^{-1} + \beta\cdot\mathbb{E}_{\mathbf{z}'}\left[\Phi_2(\ell',\ell)\frac{w(\ell')}{w(\ell)}\middle|\mathbf{z}\right] \\ &\equiv \varepsilon_S^{-1} + 1 + \mathcal{A}(\ell,\ell_{-1}), \end{aligned} \tag{36}$$

where $\mathcal{A}(\ell,\ell_{-1})$ reflects a firm's expected continuation value of adjustment cost relative to its wage level.

Without specifying the shape of the real labor adjustment cost function further, it is hard to assess the magnitude of the bias (i.e., $\mathcal{A}(\ell,\ell_{-1})$) that we are dealing with. For illustrative purposes, we use a commonly specified labor adjustment cost function $\Phi(\ell,\ell_{-1}) = \frac{\gamma}{2}\ell\left(\frac{\ell-\ell_{-1}}{\ell_{-1}}\right)^2$ (Hall, 2004; Cooper, Haltiwanger and Willis, 2007). Given this specification and after some algebra, we can simplify Equation (36) to:

$$\nu = \left(1+\frac{\gamma}{2}g_\ell^2\right)(\varepsilon_S^{-1}+1) + \gamma g_\ell(1+g_\ell) - \beta\gamma\mathbb{E}_{\mathbf{z}'}\left[g_{\ell'}(1+g_{\ell'})(1+g_{\mathtt{sw}'})|\mathbf{z}\right], \tag{37}$$

where we defined labor growth rates as $g_\ell = \frac{\ell-\ell_{-1}}{\ell_{-1}}$ and $g_{\ell'} = \frac{\ell'-\ell}{\ell}$, respectively. Furthermore, we have a firm's future growth rate in its wage bill, which equals $g_{\mathtt{sw}'} = \frac{w(\ell')\ell'}{w(\ell)\ell} - 1$. If our estimates for markdowns do not only reflect monopsony, then we can obtain "unbiased" estimates for labor market power (i.e., percentage wedges between marginal revenue products of labor and wages corrected for labor adjustment costs, as reflected by $\varepsilon_S^{-1}+1$ alone) by using Equation (37) instead. To do so, we solve for $\varepsilon_S^{-1}+1$ and obtain:

$$\varepsilon_S^{-1} + 1 = \frac{\frac{R'(\ell^*)}{w(\ell^*)} - \gamma\cdot\left(g_\ell(1+g_\ell) - \beta\mathbb{E}_{\mathbf{z}'}\left[g_{\ell'}(1+g_{\ell'})(1+g_{\mathtt{sw}'})|\mathbf{z}\right]\right)}{1+\frac{\gamma}{2}g_\ell^2},$$

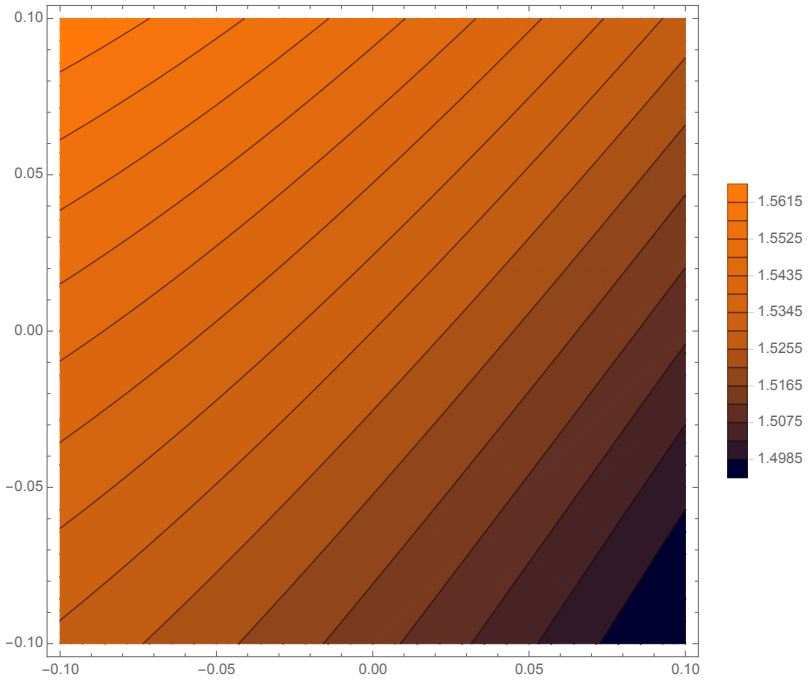
which is exactly what we wanted to show. □

We apply the above proposition by substituting out expected growth rates with their realized counterparts. In particular, our estimates for markdowns $\widehat{\nu}$ can be adjusted for labor adjustment costs as follows:

$$\frac{\varepsilon_S + 1}{\varepsilon_S} = \frac{\widehat{\nu} - \gamma \cdot [g_\ell(1 + g_\ell) - \beta g_{\ell'}(1 + g_{\ell'})(1 + g_{\mathtt{sw}'})]}{1 + \frac{\gamma}{2} g_\ell^2}. \tag{38}$$

The proposition above shows that the wedge between a firm's MRPL and the wage it pays its workers no longer only reflects monopsony power in the presence of convex labor adjustment costs. In other words, labor adjustment costs can also drive a wedge between MRPL and wages. Hence, one could be worried that our measured markdowns do not only reflect monopsony forces but also capture labor adjustment costs.

Figure 7: Corrections to markdowns from convex labor adjustment costs are quantitatively small.



Wage bill growth $g_{\mathtt{sw}'}$ is set at 2.19 percent, which is the average level of wage bill growth in U.S. manufacturing from 1987 to 2017 (BEA GDP by Industry accounts). Horizontal and vertical axes denote current and future labor growth g_ℓ and $g_{\ell'}$, respectively. The adjustment cost parameter γ is set at 0.185 (Hall, 2004).

If labor adjustment costs are quadratic, then the second part of the above proposition demonstrates that we can correct our measured markdowns so that they only reflect forces of monopsony power. This can be done if we observe a plant's growth in labor and its wage bill, and we know the parameters β and γ. Obviously, quadratic adjustment costs are not without loss of generality, but it is a specification that is often employed (see Hall, 2004; Cooper, Haltiwanger and Willis, 2007). Another advantage of this functional form is that it is governed by only one parameter. Obviously, in the absence of adjustment costs, we are back to our baseline when $\gamma = 0$ holds, as can be seen from Equation (34).

To be conservative, we choose the highest estimate for γ in Hall (2004) that is estimated with reasonable precision. This results in $\gamma = 0.185$.[58] In Figure 7, we set $\beta = 1$ and show that our measured markdowns only have to be adjusted by a maximum of 3.15 percent for a broad range of labor growth rates (varying from -10 to 10 percent). We conclude that labor adjustment costs play only a minor quantitative role and, hence, our baseline estimates must reflect labor market power.

[58] See the estimation results in table II of Hall (2004).

ONLINE APPENDIX (NOT FOR PUBLICATION)

O.1 Additional results on markdowns

O.1.1 Unionization

In the following, we provide some external validity on our markdown measures by correlating them with measures of unionization. The Annual Survey of Manufactures (ASM) and Census of Manufacturs (CM) unfortunately do not contain measures of unionization at the plant level. Instead, we leverage the Current Population Survey (CPS), which since 1984 has asked about unionization and collective bargaining status in outgoing rotation months, to construct measures of unionization at the 3-digit NAICS–state–year level. To do so, we convert the census industry codes for manufacturing in the CPS to 21 consistent, 3-digit, 2012-vintage NAICS codes using crosswalks provided by IPUMS.[59] We then run a logit regression of union coverage (union member or covered by a union) on a vector of state indicators, NAICS3 indicators, and year indicators. After collapsing the data to 3-digit NAICS–state–year cells, we fit values of union coverage based on the estimated logit coefficients. This simulated instrument adjusts for small cells (including missings) and mitigates endogeneity, although it still contains measurement error.

Due to data limitations, we can construct these measures only from 1984 onward. Hence, our sample to correlate markdowns with unionization will be somewhat smaller than our baseline sample (which starts in 1976). There are only a limited number of observations available at this narrow cell level in the CPS, so our correlations with labor market power could be noisy. To avoid this, we create a binary variable which categorizes a plant's level of unionization either above or below the median of the unionization distribution for a given year. Our results are displayed in the table below.

As expected, markdowns are negatively correlated with unionization, albeit the correlation is noisily estimated. A plant operating in a 3-digit NAICS–state cell that is in the upper half of the unionization distribution has a markdown that is about 7.5 percent lower on average. This is intuitive since plants can extract less rents in those environments in which workers are more likely to be affiliated with a union.

[59] See `https://cps.ipums.org/cps-action/variables/IND#codes_section`.

Table VII: Plant-level markdowns are negatively correlated with unionization.[†]

Dependent variable: PLANT-LEVEL TRANSLOG MARKDOWNS		
UNIONIZATION	-0.07463	-0.07628
	(0.04760)	(0.04731)
Fixed effects		
YEAR	N	Y
Weights	empwt	empwt
Observations (in millions)	10.91	10.91

[†]Markdowns are estimated under the assumption of a translog specification for gross output. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA, which approximately follows a 3-digit NAICS specification. Standard errors are clustered at the industry-state level and denoted between parentheses. Regressions are weighted by the product of employment count and ASM sampling weights. Source: Authors' calculations from ASM/CM data in 1984–2014.

O.1.2 Below-unity markdowns

Our baseline estimates on markdowns in Section 3 indicate that most plants operate in a monopsonistic environment, since markdowns are above unity. However, a relatively small fraction of our sample (approximately 11 percent) features markdowns below unity. We have verified that our core results are robust to dropping establishment-years with below-unity markdowns, but while these types of markdowns could partly be the result of statistical noise, they could also be real, especially when temporary. In the following, we rationalize why below-unity markdowns can occur under the production approach.

First, we deal with measurement error in output, but we do not account for measurement error in *inputs*. This type of measurement error can obviously impact the estimated production function coefficients. Whenever we allow for a translog specification, it is not unlikely that some of the higher order (cross- and second-order) terms are negative, which pulls estimated output elasticities below their revenue shares. Given that the overwhelming majority of observations with below-unity markdowns are between 0.75 and 1.00 (see table below), we believe moderate measurement error in inputs can likely account for some markdowns being estimated below unity.

Table VIII: Estimated plant-level markdowns in U.S. manufacturing (below-unity sample).[†]

BELOW-UNITY SAMPLE	Median	Mean	25%	75%	SD
	0.864	**0.816**	**0.748**	**0.942**	**0.173**
Sample size	$1.56 \cdot 10^5$				

[†]Markdowns are estimated under the assumption of a translog specification for gross output. The flexible input is materials. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA, which approximately follows a 3-digit NAICS specification. The sample is restricted to those plant-year observations with markdowns strictly below unity. Source: Authors' calculations from ASM/CM data in 1976–2014.

Second, our baseline results are relying on the assumption that material inputs are not subject to any monopsony forces. However, it is not unlikely that this specific assumption does not apply equally to all plants in a given industry. Think about monopolistic competition across space in the spirit of Hotelling (1929) and Salop (1979). Whenever this is the case, we are identifying monopsony for labor *relative to material inputs*. If the latter is larger than the former, then we expect to see below-unity labor markdowns.

Third, we are also assuming that labor is chosen statically. Whenever this is the case, our markdown formula based on static first-order conditions applies. Even though we show in Appendix C that labor adjustment costs are unlikely to change our estimates, we did not rule out other dynamic considerations. It might be the case that some plants in our sample are subject to a (for example) "customer capital" mechanism. Under this narrative, a plant's future demand directly depends on the amount of quantity currently sold. As a result, some plants are willing to make losses (i.e., set below-unity markdowns and/or markups) in order to sell more in the future. This reflects "investing-harvesting" incentives that are present in models of the customer base. Even though our baseline estimates do not capture these dynamic considerations, we do think they describe the data in a reasonable fashion.

Fourth and last, the estimated wedges for labor cannot be interpreted as labor market power under the classical monopsony framework whenever these wedges are below unity. However, Dobbelaere and Mairesse (2013) show that these below-unity wedges can be interpreted as labor market imperfections in a setting where risk-neutral workers and firms efficiently bargain over wages in the spirit of McDonald and Solow (1981). In fact, the estimated wedges can be used to retrieve the relevant bargaining parameters. Let $\gamma_{it} \in (0, 1)$ denote workers' bargaining power (also referred to as the "absolute extent of rent sharing"

by Dobbelaere and Mairesse, 2013). Then it can be shown that:

$$\frac{\theta_{it}^{M}}{\alpha_{it}^{M}} - \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} = \mu_{it} \cdot \frac{\gamma_{it}}{1-\gamma_{it}} \cdot \left[\frac{1-\alpha_{it}^{\ell}-\alpha_{it}^{M}}{\alpha_{it}^{\ell}}\right],$$

which we can rearrange as:

$$1 - \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} \Big/ \frac{\theta_{it}^{M}}{\alpha_{it}^{M}} = \frac{\gamma_{it}}{1-\gamma_{it}} \cdot \left[\frac{1-\alpha_{it}^{\ell}-\alpha_{it}^{M}}{\alpha_{it}^{\ell}}\right]. \tag{39}$$

Obviously, the interpretation for γ_{it} is only valid whenever relative labor wedges $\frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} / \frac{\theta_{it}^{M}}{\alpha_{it}^{M}}$ are below unity. Following Dobbelaere and Mairesse (2013), below-unity markdowns in the classical monopsony setting can also be reinterpreted as a different labor market "regime" in which there are labor market imperfections under efficient bargaining.

O.1.3 Markdowns with energy as flexible input

In our baseline estimates, we assumed that material inputs were flexible and used these inputs to identify markups. We argued in Section 5 that material inputs are more suitable than energy because (a) Davis et al. (2013) document that a large fraction of the cross-sectional dispersion in electricity prices is due to variation in purchase quantities contradicting the required "no monopsony" Assumption **III**, and (b) revenue shares for energy are much smaller when compared to material inputs; thus, measurement error in energy inputs gets amplified when estimating markdowns due to division bias. In this section, we provide some additional evidence supporting these claims.

We start by recalculating markdowns with energy as the flexible input. If there is indeed a substantial amount of monopsony in energy markets, then our estimates do not necessarily reflect labor market power alone but labor markdowns *relative to energy markdowns*, say ν_{ℓ}/ν_{E}. The evidence in Davis et al. (2013) indicates that $\nu_{E} > 1$ is likely, so we expect our markdown results with energy inputs to be lower when compared to our baseline. If monopsony in energy markets is so prevalent, in fact, it is also possible that our estimates fall below unity most of the time. This is the case whenever $\nu_{E} > \nu_{\ell}$. This is exactly what we observe in Table IX. For many industries, the median markdown is smaller than unity.

Table IX: Estimated plant-level markdowns in U.S. manufacturing (energy as a flexible input).[†]

INDUSTRY GROUP	Median	Mean	SD
Food and Kindred Products	0.559	0.758	0.825
Textile Mill Products	1.871	2.998	3.085
Apparel and Leather	0.473	0.727	0.970
Lumber	0.681	1.032	1.369
Furniture and Fixtures	0.634	0.889	1.007
Paper and Allied Products	1.118	1.553	1.632
Printing and Publishing	1.396	2.287	2.450
Chemicals	0.980	1.870	2.380
Petroleum Refining	1.963	2.258	1.781
Plastics and Rubber	1.023	1.264	1.135
Nonmetallic Minerals	0.389	0.531	0.606
Primary Metals	1.218	1.603	1.501
Fabricated Metal Products	0.656	0.846	0.889
Nonelectrical Machinery	0.310	0.376	0.276
Electrical Machinery	0.494	0.914	1.366
Motor Vehicles	0.387	0.492	0.457
Computer and Electronics	0.986	2.084	2.77
Miscellaneous Manufacturing	0.518	0.691	0.765
Whole sample	**0.618**	**0.957**	**1.350**
Sample size	$1.018 \cdot 10^6$		

[†]Markdowns are estimated under the assumption of a translog specification for gross output. The flexible input is energy. Each industry group in manufacturing corresponds to the manufacturing categorization of the U.S. Bureau of Economic Analysis (BEA) which approximately follows a 3-digit NAICS specification. Source: authors' calculations from ASM/CM data in 1976–2014.

Furthermore, energy shares in U.S. manufacturing are small. The NBER-CES Manufacturing Database indicates that revenue shares average around 2 percent.[60] In addition, the dispersion in energy shares is substantial: its 10th and 90th percentiles equal 0.59 percent and 4.26 percent, respectively. Note, however, that energy is not only more dispersed across plants, but it is also more volatile for a given plant. To show this, we have calculated the standard deviation of log inputs for each plant's life cycle. Inputs are normalized by the mean of its log level over time. The results are displayed in Table X. Because of its modest and volatile revenue share, we conjectured that markdowns estimated with energy as the flexible input would be much less accurate. Indeed, because of the volatility of expenditure

[60]The median revenue share for energy is even smaller, at 1.18 percent.

on energy inputs, measurement error in energy shares is amplified by division bias. This is reflected in the within-industry standard deviations of markdowns when estimated with energy inputs, which are significantly higher compared to our baseline estimates.

Table X: Variability of inputs.[†]

INPUT	Median	Mean	25%	75%	SD
Capital	0.0154	0.0280	0.0080	0.0339	0.0341
Labor	0.0307	0.0401	0.0171	0.0518	0.0349
Materials	0.0391	0.0493	0.0222	0.0648	0.0394
Energy	0.0625	0.0954	0.0335	0.1158	0.1331

[†]For each plant, we calculate the standard deviation of its log normalized inputs over time. Each plant's input is normalized by the mean of its log level over time. The sample is restricted to those plants that have at least three observations over their life cycle. Source: Authors' calculations from ASM/CM data in 1976–2014.

As expected, we see that energy usage is much more volatile for the average plant when compared to other inputs. Hence, it is not surprising that our markdown estimates with energy are much more volatile when compared to our baseline estimates.

O.1.4 Markups

In the following, we report our estimates for markups. Summary statistics are provided for each industry group. The results clearly indicate that there is market power in output markets: the median (and mean) markup at the plant-year level equals about 20 percent. Similar to markdowns, there is a substantial amount of variation across industry groups, though the within-industry variation of markups is substantially more limited when compared to markdowns. The interquartile range (IQR) for markups is about 16.5 percent, whereas the standard deviation for the whole sample is 18.8 percent.

While these estimates are informative for markups, it should be noted that our estimates for markups *in isolation* are faced with a bias. This is because we proxied physical output with deflated revenues, which causes a downward bias in markups (see Klette and Griliches, 1996). This has recently been reiterated by Bond et al. (2021). Thus, in a conservative sense, our estimates for markups can also be interpreted as lower bounds for market power in output markets. Note, however, that these estimates for markups are still valid when they are used in order to obtain estimates for markdowns. This is a point we emphasize in Online Appendix O.5.

Table XI: Estimated plant-level markups in U.S. manufacturing.[†]

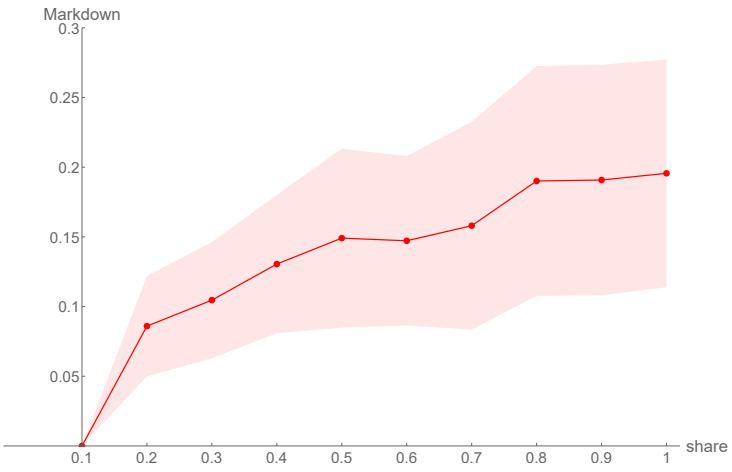
INDUSTRY GROUP	Median	Mean	IQR_{75-25}	SD
Food and Kindred Products	1.145	1.165	0.139	0.123
Textile Mill Products	1.218	1.220	0.136	0.122
Apparel and Leather	1.286	1.293	0.152	0.193
Lumber	1.056	1.055	0.115	0.107
Furniture and Fixtures	1.227	1.226	0.143	0.122
Paper and Allied Products	1.081	1.084	0.129	0.106
Printing and Publishing	1.249	1.234	0.136	0.183
Chemicals	1.330	1.368	0.243	0.214
Petroleum Refining	1.119	1.160	0.194	0.192
Plastics and Rubber	1.107	1.105	0.147	0.131
Nonmetallic Minerals	1.219	1.218	0.104	0.135
Primary Metals	1.129	1.142	0.116	0.096
Fabricated Metal Products	1.194	1.198	0.073	0.058
Nonelectrical Machinery	1.449	1.488	0.278	0.193
Electrical Machinery	1.286	1.294	0.105	0.083
Motor Vehicles	1.170	1.178	0.082	0.071
Computer and Electronics	1.023	1.018	0.197	0.180
Miscellaneous Manufacturing	1.255	1.263	0.071	0.068
Whole sample	**1.205**	**1.214**	**0.165**	**0.188**
Sample size	$1.393 \cdot 10^6$			

[†]Markups are estimated under the assumption of a translog specification for gross output. The flexible input is materials. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA, which approximately follows a 3-digit NAICS specification. Source: Authors' calculations from ASM/CM data in 1976–2014.

O.1.5 Size, age, and productivity effects

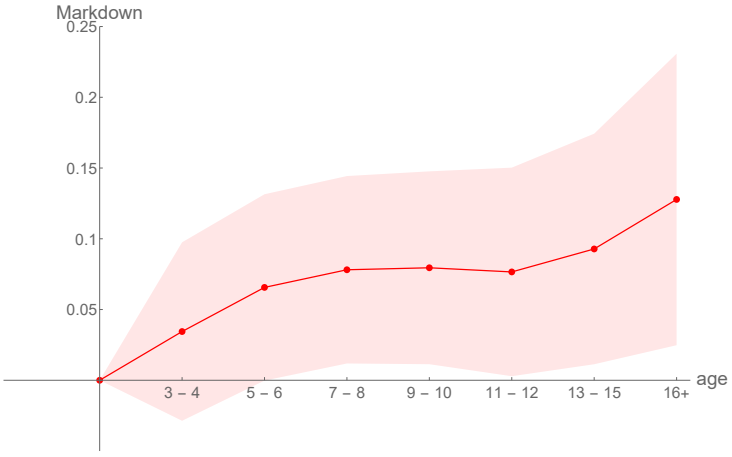
O.1.5.1 Size and age regressions without controls

Figure 8: Markdowns increase with establishment size.



Note: The figure shows point estimates and 95 percent confidence intervals of plant-specific markdowns on size (as measured by employment share) indicators, controlling for state, industry, and year fixed effects. The omitted group is the smallest size indicator, so coefficients reflect deviations relative to this baseline. The indicator labeled "0.1" is equal to unity for those plants with employment shares $s \in (0, 0.1]$. Other indicators are defined similarly. Standard errors are clustered at the industry level. Source: Authors' own calculations from ASM/CM data in 1976–2014.

Figure 9: Markdowns increase with establishment age, but this result only holds when not controlling for establishment size.



Note: The figure shows point estimates and 95-percent confidence intervals of plant-specific markdowns on age category indicators, controlling for state, industry and year fixed effects. The omitted group is the smallest age category, less than three years, so coefficients reflect deviations relative to this baseline. Standard errors are clustered at the industry level. Source: Authors' own calculations from ASM/CM data in 1976–2014.

O.1.5.2 Baseline results in tabular form

Table XII: Nonparametric estimates of markdowns on size, age, and productivity[†]

Dependent variable: MARKDOWNS					
	SIZE		AGE		TFPR
Share bin		Age bin		TFPR %	
0.1 – 0.2	0.0849	3 – 4	0.0242	1% – 5%	−0.8088
	(0.0181)		(0.0308)		(0.2842)
0.2 – 0.3	0.1030	5 – 6	0.0536	5% – 10%	−0.8162
	(0.0212)		(0.0327)		(0.3920)
0.3 – 0.4	0.1286	7 – 8	0.0637	10% – 25%	−0.7629
	(0.0254)		(0.0326)		(0.4198)
0.4 – 0.5	0.1471	9 – 10	0.0557	25% – 50%	−0.6257
	(0.0326)		(0.0333)		(0.4360)
0.5 – 0.6	0.1452	11 – 12	0.0586	50% – 75%	−0.5020
	(0.0308)		(0.0365)		(0.4383)
0.6 – 0.7	0.1560	13 – 15	0.0709	75% – 90%	−0.4031
	(0.0377)		(0.0401)		(0.4486)
0.7 – 0.8	0.1880	16+	0.0978	90% – 95%	−0.2453
	(0.0419)		(0.0514)		(0.4747)
0.8 – 0.9	0.1882			95% – 99%	0.1084
	(0.0420)				(0.5182)
0.9 – 1	0.1934			99%+	0.8046
	(0.0420)				(0.5321)
Observations (in millions)	1.393		1.393		1.393
R^2	0.2579		0.2579		0.3385

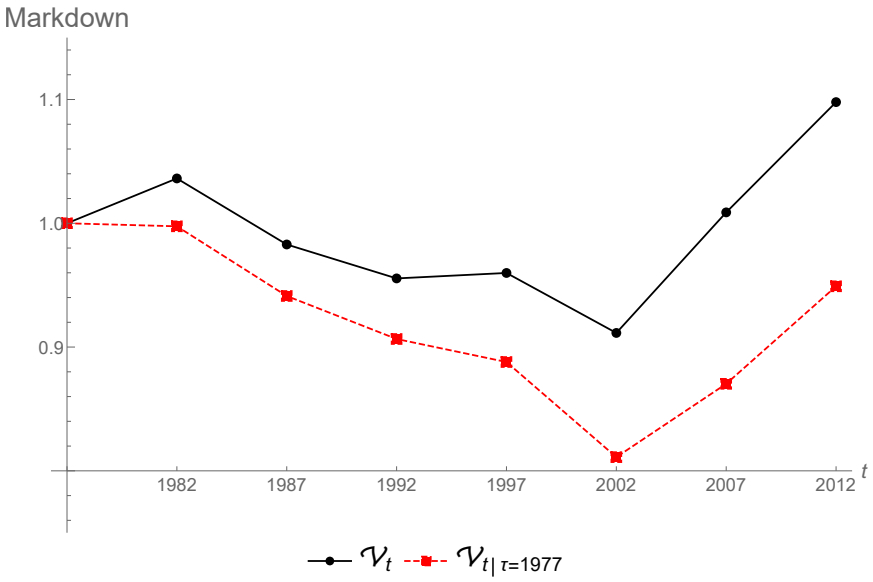
[†]All regression specifications contain fixed effects at the state, industry and year level, and are weighted by the product of employment and the ASM sampling weights. The results are almost identical when only ASM sampling weights are used instead. The specifications for size and age respectively control for the other factor. The omitted categories for the size, age, and productivity specifications are 0–0.1, 1–2 and $< 1\%$, respectively. Hence, the regression coefficients reflect deviations relative to these baselines. The indicator labeled "0.1–0.2" is equal to unity for those plants with employment shares $s \in (0.1, 0.2]$. Other indicators for the size specification are defined similarly. Standard errors, in parentheses, are clustered at the industry level. Source: Authors' calculations from ASM/CM data in 1976–2014.

O.2 Additional results on the aggregate markdown

O.2.1 Compositional effects and benefits

In this section, we provide several robustness checks on the aggregate markdown $\mathcal{V}_t$. First, we verify that the distinct time evolution of the aggregate markdown is not purely driven by compositional changes across local labor markets. To do so, we recalculate the aggregate markdown but fix its weights across local labor markets at their 1977 level. That is, we construct $\mathcal{V}_{t|\tau} \equiv \sum_{j\in J}\sum_{l\in L}\omega_{jl\tau}\mathcal{V}_{jlt}$ with $\tau = 1977$. The results can be found in Figure 10.

Figure 10: The qualitative nature of the time evolution for the aggregate markdown cannot be explained by compositional changes across local labor markets.



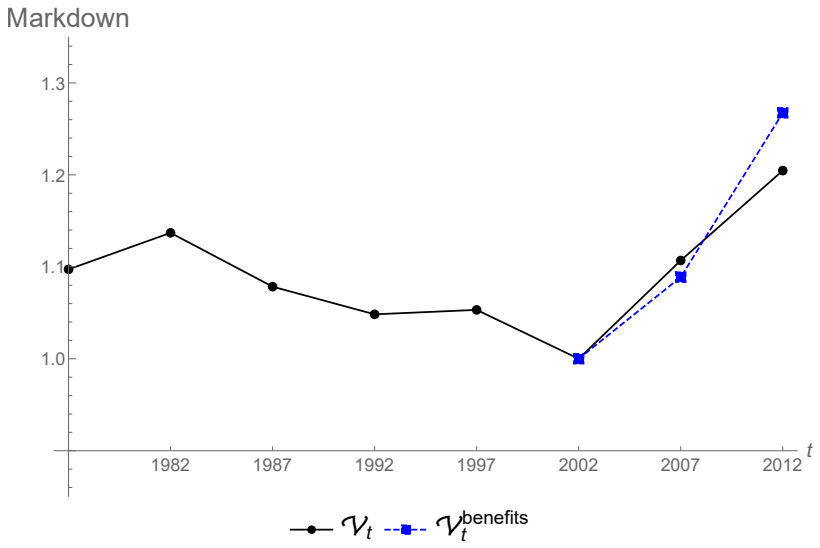
Markdowns are constructed under the assumption of translog production and aggregated according to equation (14). Our baseline measure $\mathcal{V}_t$ is depicted by the solid black line. The aggregate markdown $\mathcal{V}_{t|\tau=1977}$ (dashed red) is calculated by fixing the employment weights for local labor markets at their 1977 values. All measures are normalized relative to their initial value in 1977. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

We find that the qualitative nature of the aggregate markdown is preserved. When employment weights across local labor markets are fixed at their 1977 values, the aggregate markdown also decreases until 2002 and increases afterward. However, its decrease from 1977 to 2002 is a bit stronger than in our baseline specification. Nevertheless, we conclude that the evolution of the aggregate markdown $\mathcal{V}_t$ cannot be accounted for by changes in the

employment composition across local labor markets.

Second, our baseline specification of the aggregate markdown does not include health and pension benefits. However, these benefits are available from 2002 onward. We verify that the aggregate markdown also starkly increases whenever benefits are taken into consideration. Given that benefits are available from only 2002 onward, we normalize our series to unity in 2002. As shown in Figure 11, the aggregate markdown also increases from 2002 onward whenever benefits are included.

Figure 11: The stark increase of the aggregate markdown $\mathcal{V}_t$ (solid black) from 2002 onward is preserved whenever benefits (dashed blue) are also taken into account.

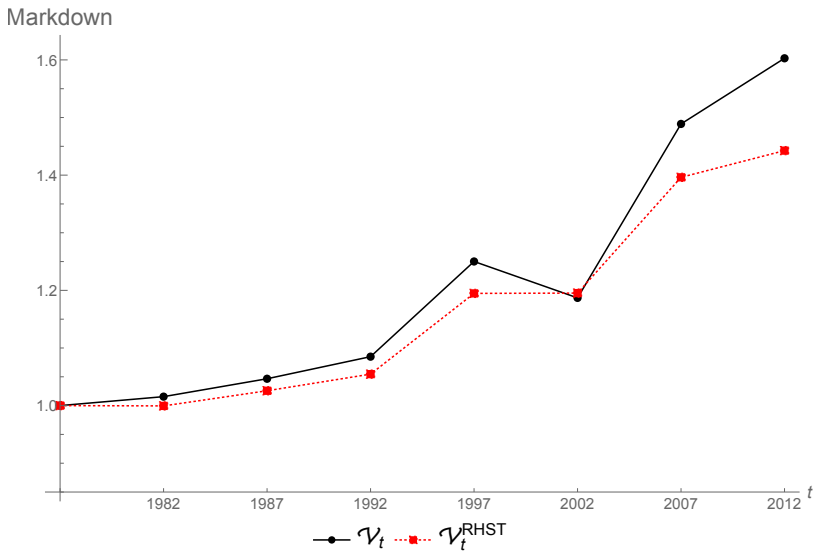


Markdowns are constructed under the assumption of translog production and aggregated according to equation (14). The aggregate markdown $\mathcal{V}_t^{\text{benefits}}$ is calculated by including health and pension benefits. All measures are normalized relative to their values in 2002. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

O.2.2 Secular trend in markdowns: Cobb-Douglas

In our baseline estimates, we specified production functions to be translog. By construction, the translog specification allows output elasticities to vary with the level of inputs. As a result, these output elasticities can vary over time as well. Under a Cobb-Douglas specification, output elasticities are constant and markdowns can only vary over time because of changes in revenue shares. In the following, we show that allowing for time-varying output elasticities is important for several measures of the aggregate markdown.

Figure 12: Time evolution of aggregate markdowns across U.S. manufacturing plants from 1977 to 2012 (Cobb-Douglas case). Unlike the baseline estimation using translog, these measures are increasing over time (cfr. Figure 4 in main text).



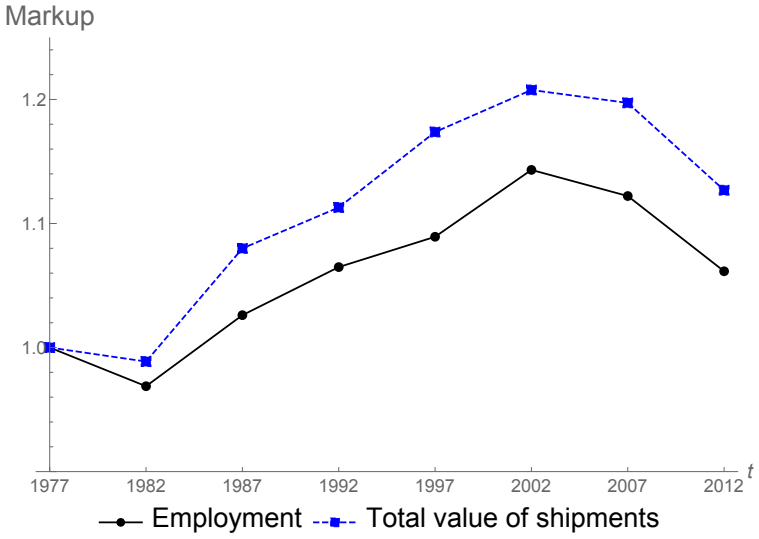
Markdowns are constructed under the assumption of Cobb-Douglas production and aggregated according to Equations (14) and (16), respectively. All measures are normalized relative to their initial value in 1977. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

We start by calculating the aggregate measures $\mathcal{V}_t$ and $\mathcal{V}_t^{\text{RHST}}$ whenever production technologies are assumed to be Cobb-Douglas. While these measures are decreasing over time (at least before 2002) under a translog specification, the opposite is true whenever markdowns are estimated under Cobb-Douglas technologies. This is illustrated in Figure 12. These differences underline that Cobb-Douglas specifications can be quite restrictive. By construction, the Cobb-Douglas specification assumes that output elasticities are constant and, hence, ignores any time variation in a plant's output elasticities. Conversely, a translog specification allows precisely for this. Our results favor the translog specification since they indicate that this time variation is quantitatively important.

O.2.3 Secular trend in markups

In this section, we present the time series for the aggregate markup. The aggregate markup at the market level is calculated according to Equation (13). Then, we aggregate markups across markets through either employment or revenue weights.

Figure 13: Time evolution of revenue- and employment-weighted markups (the black and blue line, respectively) across U.S. manufacturing plants from 1977 to 2012.



Markups are constructed under the assumption of translog production and aggregated according to equation (13). Source: Authors' own calculations from ASM/CM data in 1976–2014.

As emphasized by Bond et al. (2021), the estimation of micro-level markups with deflated revenues, instead of physical output, leads to biases that make interpretation challenging (see Online Appendix O.5). In turn, bias in the level of markups at the micro level will lead to bias in the aggregate markup. Note however, as we show formally in Online Appendix O.5.1, this concern does *not* apply to our estimation of markdowns.

Consequently, we feel that using our methodology to present markups should—at the very least—be treated cautiously by other researchers. However, presenting a trend of aggregate markups could still be useful to others even when bias is present—perhaps in comparison to markup trends created under different approaches and different biases. This trend is depicted in Figure 13.

O.2.4 Decomposition of aggregate markdowns

In the following, we will apply the decomposition by Foster, Haltiwanger and Krizan (2001) to aggregate markdowns in order to understand what was driving its changes. However, this is not straightforward because the accounting decomposition by Foster, Haltiwanger and Krizan (2001) applies to arithmetic (weighted) averages only. In the discussion

below, we will present some accounting identities that will allow us to apply the decomposition by Foster, Haltiwanger and Krizan (2001) to harmonic (weighted) averages. To do so, we start with the following lemma.

LEMMA 2. For any aggregate variable X_t, we have:

$$\Delta X_t = -\frac{\Delta X_t^{-1}}{1 + \Delta X_t^{-1}}. \tag{40}$$

Proof. By definition, we have:

$$\begin{aligned}
\Delta X_t^{-1} &= \frac{X_t^{-1} - X_{t-1}^{-1}}{X_{t-1}^{-1}} \\
&= -\frac{X_{t-1}}{X_t}\left(\frac{X_t - X_{t-1}}{X_{t-1}}\right) \\
&= -\frac{\Delta X_t}{1 + \Delta X_t}.
\end{aligned}$$

Then, the lemma follows directly by solving for ΔX_t. □

This is useful since our definition of the aggregate markdown consists of a ratio of two sales-weighted harmonic averages. That is, we have $\mathcal{V}_{jlt} \equiv \frac{V_{jlt}}{\mathcal{M}_{jlt}}$ with:

$$V_{jlt} = \left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta_{it}^L}{\theta_{jlt}^L} \cdot (\nu_{it}\mu_{it})^{-1}\right)^{-1} \tag{41}$$

$$\mathcal{M}_{jlt} = \left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta_{it}^M}{\theta_{jlt}^M} \cdot \mu_{it}^{-1}\right)^{-1}. \tag{42}$$

Note that for any weighted harmonic average X_t, we can write:

$$\widetilde{X}_t \equiv X_t^{-1} = \sum_{i \in F_t} s_{it} x_{it}^{-1} \equiv \sum_{i \in F_t} s_{it} \tilde{x}_{it}.$$

The latter is just a simple (i.e., arithmetic) weighted average. For these types of averages,

we can apply the decomposition of Foster, Haltiwanger and Krizan (2001):

$$\begin{aligned}\Delta\widetilde{X}_t &= \sum_{i\in C_t} s_{it-1}\Delta\tilde{x}_t + \sum_{i\in C_t}\left(\tilde{x}_{it-1}-\widetilde{X}_{t-1}\right)\Delta s_{it} + \sum_{i\in C_t}\Delta\tilde{x}_{it}\Delta s_{it} \\ &\quad + \sum_{i\in N_t} s_{it}\left(\tilde{x}_{it}-\widetilde{X}_{t-1}\right) - \sum_{i\in X_t} s_{it-1}\left(\tilde{x}_{it-1}-\widetilde{X}_{t-1}\right) \qquad (43) \\ &\equiv \text{WITHIN}_t + \text{BTWN}_t + \text{COV}_t + \text{ENTRY}_t - \text{EXIT}_t, \qquad (44)\end{aligned}$$

where the growth rate of $\widetilde{X}_t$ can be decomposed into within-firm, between-firm, covariance, entry and exit components, respectively. Note that the first three components can only be applied to incumbent firms (i.e., firms active in periods t and $t-1$). By definition of the aggregate markdown, we have:

$$\begin{aligned}\widetilde{V}_{jlt} \equiv V_{jlt}^{-1} &= \sum_{i\in F_t(j,l)} s_{it}\cdot\frac{\theta_{it}^L}{\theta_{jlt}^L}\cdot(\nu_{it}\mu_{it})^{-1} \\ &\equiv \sum_{i\in F_t(j,l)} s_{it}\cdot\tilde{v}_{it} \\ \widetilde{\mathcal{M}}_{jlt} \equiv \mathcal{M}_{jlt}^{-1} &= \sum_{i\in F_t(j,l)} s_{it}\cdot\frac{\theta_{it}^M}{\theta_{jlt}^M}\cdot\mu_{it}^{-1} \\ &\equiv \sum_{i\in F_t(j,l)} s_{it}\cdot\tilde{\mu}_{it}.\end{aligned}$$

Thus, we can apply the insight of Foster, Haltiwanger and Krizan (2001) in (43) to $\widetilde{V}_{jlt}$ and $\widetilde{\mathcal{M}}_{jlt}$ to obtain decompositions for $\Delta\widetilde{V}_{jlt}$ and $\Delta\widetilde{\mathcal{M}}_{jlt}$. This will aid us in understanding growth in the aggregate markdown, since we have:

$$\begin{aligned}\Delta\mathcal{V}_{jlt} &= \Delta V_{jlt} - \Delta\mathcal{M}_{jlt} \\ &= -\frac{\Delta\widetilde{V}_{jlt}}{1+\Delta\widetilde{V}_{jlt}} + \frac{\Delta\widetilde{\mathcal{M}}_{jlt}}{1+\Delta\widetilde{\mathcal{M}}_{jlt}} \qquad (45) \\ &\approx \Delta\widetilde{\mathcal{M}}_{jlt} - \Delta\widetilde{V}_{jlt}, \qquad (46)\end{aligned}$$

where the last approximation follows from the fact that we have $-\frac{x}{1+x} \simeq -x$ up to a first order for small values of x. This seems appropriate in our setting given the observed movements in aggregate markdowns. Thus, growth in the aggregate markdown, for a given

local labor market, is primarily led by those components that are more important for the growth rate of the inverse aggregate markup—i.e., $\Delta\widetilde{\mathcal{M}}$—whereas it is slowed down by those components that determine the growth rate of the inverse aggregate labor wedge—i.e., $\Delta\widetilde{V}$.

Table XIII: Decomposition of $\Delta\widetilde{\mathcal{M}}$ and $\Delta\widetilde{V}$ (cfr. Equation 46).† Movements in the aggregate markdown are not clearly driven by one specific type of reallocation.

YEAR		WITHIN$_t$	BTWN$_t$	COV$_t$	ENTRY$_t$	EXIT$_t$
1977 – 1982	$\Delta\widetilde{\mathcal{M}}$	0.3618	0.1370	0.1547	0.2042	0.1423
1977 – 1982	$\Delta\widetilde{V}$	0.3997	0.1231	0.1120	0.2162	0.1490
1982 – 1987	$\Delta\widetilde{\mathcal{M}}$	0.3724	0.1125	0.1140	0.2317	0.1694
1982 – 1987	$\Delta\widetilde{V}$	0.3386	0.1271	0.1553	0.2261	0.1528
1987 – 1992	$\Delta\widetilde{\mathcal{M}}$	0.3782	0.1131	0.1218	0.2244	0.1625
1987 – 1992	$\Delta\widetilde{V}$	0.3537	0.1236	0.1585	0.2190	0.1453
1992 – 1997	$\Delta\widetilde{\mathcal{M}}$	0.3903	0.1250	0.1164	0.2113	0.1570
1992 – 1997	$\Delta\widetilde{V}$	0.3452	0.1281	0.1753	0.2119	0.1395
1997 – 2002	$\Delta\widetilde{\mathcal{M}}$	0.3555	0.1189	0.1193	0.2408	0.1655
1997 – 2002	$\Delta\widetilde{V}$	0.3358	0.1262	0.1583	0.2307	0.1491
2002 – 2007	$\Delta\widetilde{\mathcal{M}}$	0.3777	0.1273	0.1244	0.2172	0.1534
2002 – 2007	$\Delta\widetilde{V}$	0.3363	0.1384	0.1819	0.1966	0.1469
2007 – 2012	$\Delta\widetilde{\mathcal{M}}$	0.3979	0.1281	0.1280	0.2033	0.1426
2007 – 2012	$\Delta\widetilde{V}$	0.3441	0.1449	0.1767	0.190	0.1444

†Markdowns are estimated under the assumption of a translog specification for gross output. The flexible input is materials. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA, which approximately follows a 3-digit NAICS specification. Each component is denoted in absolute values and normalized by the sum of absolute values for each component. The table reports the employment-weighted mean across local labor markets. Source: Authors' calculations from ASM/CM data in 1976–2014.

We follow Foster, Haltiwanger and Krizan (2001) and calculate the employment-weighted average across local labor markets of the *absolute* contribution for each component. By construction, we can write $\Delta\widetilde{V} = \text{WITHIN} + \text{BTWN} + \text{COV} + \text{ENTRY} - \text{EXIT}$. Then,

for each local labor market, we calculate each component's absolute contribution by taking its absolute value and dividing it by the sum of absolute values for each component. That is:

$$\widehat{x} = \frac{|x|}{|\text{WITHIN}| + |\text{BTWN}| + |\text{COV}| + |\text{ENTRY}| + |\text{EXIT}|}$$

for $x \in \{\text{WITHIN}, \text{BTWN}, \text{COV}, \text{ENTRY}, \text{EXIT}\}$.[61] Then, we report averages across local labor markets using employment weights. This is appropriate in our setting since we aggregate markdowns across local labor markets by taking employment-weighted averages in order to obtain $\mathcal{V}_t$.

Our decomposition in Equation (46) indicates that movements in the aggregate markdown are primarily determined by those components that are relatively important for $\Delta\widetilde{\mathcal{M}}$ but not for $\Delta\widetilde{V}$. However, our results in Table XIII indicate that each component is about equally important for $\Delta\widetilde{\mathcal{M}}$ and $\Delta\widetilde{V}$. As a result, we conclude that movements in the aggregate markdown are not clearly driven by one specific type of reallocation.

[61] We report absolute contributions for each component since the patterns over time for each raw component are difficult to interpret: they can switch signs over time and are also quite volatile. This is similar to Foster, Haltiwanger and Krizan (2001), who apply the decomposition to aggregate productivity in U.S. manufacturing sectors (see their Table 8.7). In fact, they mention that their results can be quite "erratic" under the used accounting decomposition.

O.2.5 Aggregate markdowns and local concentration in tabular form

Table XIV: Measures of the aggregate markdown and local concentration.[†]

Specification: TRANSLOG MARKDOWNS				
YEAR	$\mathcal{V}_t$	$\mathcal{V}_t^{\text{RHST}}$	$\mathcal{V}_t^{\text{dLEU}}$	LOCAL_t
1977	1.000	1.000	1.000	1.000
1982	1.0362	0.9653	0.9495	0.9640
1987	0.9829	0.9515	0.9392	0.9841
1992	0.9555	0.9460	0.9289	0.9707
1997	0.9599	0.9344	0.9330	0.9224
2002	0.9114	0.9322	0.9310	0.9269
2007	1.0088	0.9366	0.9815	0.9297
2012	1.0979	0.9272	1.016	0.9646

[†]Markdowns are estimated under the assumption of a translog specification for gross output. Aggregate markdowns $\mathcal{V}_t$, $\mathcal{V}_t^{\text{dLEU}}$ and $\mathcal{V}_t^{\text{RHST}}$ are calculated according to formulas (14), (15) and (16), respectively, whereas LOCAL_t denotes local concentration as calculated according to Equation (18). All values are normalized with respect to 1977. Source: Authors' own calculations from quinquennial CM data from 1977–2012.

O.3 Details on GMM-IV estimation procedure

O.3.1 Implementation of constant returns to scale restriction

We implement the "production approach" for obtaining markdowns by relying on proxy variable methods. While the induced moment conditions are easily derived and understood, Gandhi, Navarro and Rivers (2020) emphasize that point identification is not achieved when applying the methodology by De Loecker and Warzynski (2012), for example. To address this criticism, we apply the solution suggested in Flynn, Gandhi and Traina (2019). They show that the nonidentification problem can be resolved whenever a production function's return to scale is ex-ante specified. Similar to their work, we show the robustness of our markdown estimates whenever we impose a constant-returns-to-scale restriction.[62] Assuming constant returns to scale seems reasonable, since a substantial body of previous work (e.g., Basu and Fernald, 1997; Syverson, 2004a; Syverson, 2004b) has shown that constant returns to scale is a good approximation for manufacturing plants.

In the following, we will briefly describe how our estimation procedure is adjusted (for the translog case) when imposing constant returns to scale. In fact, this requires minor adjustments only. Steps 1 and 2 are unchanged, whereas we only need to add some moment conditions to step 3. To do so, we define a firm's returns to scale as follows:

$$\Sigma_{it}(\boldsymbol{\beta}) = \sum_{\iota \in \{k,\ell,m,e\}} \frac{\partial f(\mathbf{x}_{it};\boldsymbol{\beta})}{\partial \iota_{it}}. \tag{47}$$

Also, if we define the vector $\boldsymbol{\chi}_{it} = (1, \tilde{\mathbf{x}}_{it}')' = (1, k_{it}, \ell_{it}, m_{it}, e_{it})' \in \mathbb{R}^{K+1}$, then the new set of moment conditions can be compactly written as:

$$\mathbb{E}\begin{pmatrix} \xi_{it}(\boldsymbol{\beta})\mathbf{z}_{it} \\ \Sigma_{it}(\boldsymbol{\beta}) - 1 \end{pmatrix} = \mathbf{0}_{(Z+1)\times 1}. \tag{48}$$

In the case of a translog production function, we can write the constant returns to scale restriction as a linear operator:

$$\Sigma_{it}(\boldsymbol{\beta}) - 1 = (R\boldsymbol{\beta})' \boldsymbol{\chi}_{it},$$

[62] We draw similar conclusions whenever we allow for deviations around constant returns to scale.

where R is a $5 \times Z$ matrix defined as:

$$R = \begin{bmatrix} -1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Our estimation results are displayed in column 2 of Table V in the main text (see Section 5).

O.3.2 Bootstrapping procedure

The GMM-IV estimator of the proxy variable approach does not have a closed-form solution for its standard errors. Furthermore, even if we did have these standard errors for the production function coefficients, it is difficult to derive standard errors for the aggregate markdown because of its nonlinear structure. As a result, we resort to bootstrapping methods, similar to De Loecker and Warzynski (2012). In the following, we describe the bootstrap algorithm that we implemented with the census data.

Initiate bootstrap round parameter at $b = 1$.

I. For each industry group $j \in \{1, \ldots, \mathcal{J}\}$, draw a random sample with replacement from the unbalanced ASM panel containing $N_j^{[b]} = 0.9 \times N_j$ observations.

II. For each plant that has been sampled, select its entire life cycle; i.e., we engage in *panel bootstrapping* (or block-bootstrapping at the plant level). This generates the unbalanced sample $S_j^{[b]}$.

III. Obtain the estimated production function parameters $\widehat{\boldsymbol{\beta}}_j^{[b]}$ (with the two-step GMM-IV estimator from De Loecker and Warzynski, 2012) for each industry j, using data from sample $S_j^{[b]}$.

IV. For each census year τ, calculate the aggregate markdown $\widehat{\mathcal{V}}_\tau^{[b]}$ (normalized to unity in 1977) with the universe of manufacturing plants from the CM using the production function parameters $\widehat{\boldsymbol{\beta}}^{[b]} = (\widehat{\boldsymbol{\beta}}_1^{[b]\prime}, \ldots, \widehat{\boldsymbol{\beta}}_{\mathcal{J}}^{[b]\prime})'$.

V. Define $b := b + 1$ and repeat step **I**. Stop the algorithm whenever $b > B$.

Confidence interval bounds at the α-significance level for the aggregate markdown $\mathcal{V}_\tau$ can then be constructed by taking the $100 \cdot \frac{\alpha}{2}$ and $100 \cdot (1 - \frac{\alpha}{2})$ percentile of the set $\{\widehat{\mathcal{V}}_\tau^{[b]}\}_{b=1}^{B}$. We construct 95 percent confidence intervals through 500 simulations; i.e., $\alpha = 0.05$ and $B = 500$.

Note that the constructed confidence interval for the normalized aggregate markdown does not necessarily have to be symmetric around the estimated (normalized) aggregate markdown $\mathcal{V}_t$. This is because of the nonlinear structure of markdowns at the firm level and how firm-level markdowns enter the aggregate markdown in a nonlinear fashion. Note that we only sample with replacement in the ASM to estimate the production function parameters $\boldsymbol{\beta}$. However, markdowns at the firm level and the aggregate markdown are always calculated using the full sample of the CM for every census year $\tau \in \{1977, \ldots, 2012\}$. By construction, there is no confidence interval for the aggregate markdown in 1977, since this value is always normalized to unity.

Using these block-bootstrap methods, we have verified that the production function parameters $\boldsymbol{\beta}$ are statistically significant for every industry group. In particular, we find that the cross- and second-order terms of our production function specification are statistically significant, indicating the importance of the translog specification.

O.4 Benefits

O.4.1 Measures of compensation

In our baseline estimation procedure, we use a plant's total wage bill (or "payroll") as its total variable expenditure on labor. Following the instructions of form MA-10000, payroll is an overall measure of wages and salaries paid to a plant's employee(s). An employee is defined according to Internal Revenue Service Form 941, Employer's Quarterly Federal Tax Return. This includes:

- All persons on paid sick leave, paid holidays, and paid vacation during these pay periods
- Officers at this establishment, if a corporation
- Spread on stock options that are taxable to employees as wages

An employer's wage bill is defined as its payroll before deductions, excluding an employer's cost for fringe benefits. In particular, it includes:

- Employee's Social Security contributions, withholding taxes, group insurance premiums, union dues, and savings bonds
- In gross earnings: commissions, dismissal pay, paid bonuses, employee contributions to pension plans such as 401(k), vacation and sick leave pay, and the cash equivalent of compensation paid in kind
- Spread on stock options that are taxable to employees as wages
- Salaries of officers of this establishment, if a corporation
- Paid holiday, personal, funeral, jury duty, military and family leave
- Nonproduction bonuses
  - Cash profit-sharing
  - Employee recognition
  - End-of-year
  - Holiday
  - Payment in lieu of benefits—Referral
  - Other

By construction, the wage bill does not include benefits. Fortunately, the ASM/CM does include a measure of these benefits from 2002 onward. Benefits cover health insurance, pension plans, and other employer-paid benefits. The latter includes legally required benefits (e.g., Social Security, workers' compensation insurance, unemployment tax, state disability insurance programs, Medicare), benefits for life insurance, "quality of life" benefits (e.g., childcare assistance, subsidized commuting, etc.), employer contributions to pretax benefit accounts (e.g., health savings accounts), education assistance, and other benefits. In the end, our results on markdowns are not qualitatively changed whenever we use a measure for labor that includes benefits.

O.4.2 Understanding markdowns with benefits

In one of our robustness exercises, we calculated micro-level markdowns whenever benefits were also included as a part of workers' compensation. We saw from Table V that median markdowns at the industry group level slightly declined relative to our baseline results from Table I.

In this section, we verify that the differences between our baseline estimates for markdowns and those with benefits included can be rationalized by the fraction of benefits in total compensation.

Given that benefits are not included in our baseline estimates, we expect that these estimates are biased upwards. This is intuitive, since we are including only wage payments in the denominator of the markdown. As a result, the bias of our baseline estimates should increase more for those plants whose compensation to workers relies more on benefits. We measure the latter by the "benefit fraction"—i.e., total benefits relative to the sum of total benefits and wage payments.

Our hypothesis is confirmed by Table XV. Our baseline estimates, but more importantly the *difference* between our baseline estimates and those including benefits, are increasing in the benefit fraction. Our conclusions are not affected much when we take absolute differences instead. This is as expected, since our baseline estimates are larger than those estimates including benefits for the overwhelming fraction of our sample anyway.

However, it could also be argued that the sign of the benefit fraction coefficient may at first

Table XV: The fraction of benefits in total compensation accounts for the difference between baseline and markdowns with benefits.†

Dependent variable	ν_{it}	$\nu_{it} - \nu_{it}^{\text{benefit}}$	$\lvert\nu_{it} - \nu_{it}^{\text{benefit}}\rvert$
BENEFIT FRACTION	1.682 (0.3153)	1.299 (0.2057)	1.0360 (0.1642)
Fixed effects			
INDUSTRY	Y	Y	Y
STATE	Y	Y	Y
YEAR	Y	Y	Y
Weights	empwt	empwt	empwt
Observations	$4.02 \cdot 10^5$	$4.02 \cdot 10^5$	$4.02 \cdot 10^5$

†Markdowns are estimated under the assumption of a translog specification for gross output. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA, which approximately follows a 3-digit NAICS specification. Baseline markdowns are denoted by ν, whereas markdowns with benefits are denoted by ν^{benefit}. A plant's benefit fraction is defined as benefit payments divided by the sum of wage and benefit payments to workers. All regressions contain size and age controls at the plant level. Furthermore, all regressions include average earnings (i.e., total wage bill divided by employment count) as a control. Standard errors are clustered at the industry group level and denoted between parentheses. Regressions are weighted by the product of employment count and ASM sampling weights. Source: Authors' calculations from ASM/CM data in 2002–2014.

be surprising, as one might associate larger benefit shares of compensation with stronger employee bargaining power and thus expect a *lower* markdown. However, because we control for plant-level average earnings, the results in Table XV show how markdown estimates change as the benefit share changes, holding average earnings constant. To the extent that benefit shares are higher in lower-wage plants, on average, our regressions control for this mechanical relationship.

Finally, note that this sample is smaller than our base sample, since we can estimate markdowns with benefits from 2002 onward only.

O.5 Critique by Bond *et al.* (2021)

O.5.1 Deflated revenues

Unfortunately, most firm-level data sets do not have physical output available. As an alternative, physical output is typically approximated by deflating revenues with some industry-level deflator. While it could be argued that revenues are more easily comparable across firms, it does not align with the theory behind production function estimation. In fact, Klette and Griliches (1996) show that estimated production function coefficients in an imperfectly competitive environment with price heterogeneity are downwardly biased whenever physical output is approximated with deflated revenues. This immediately implies that markups are also downwardly biased under the production approach.

Bond et al. (2021) demonstrate that the problem is even more severe: using deflated revenues does not only induce a downward bias, but it results in ratio estimators, such as the one we employ in Equation (3), to be equal to unity. To see why this is the case, consider the analog of (3) using revenue elasticities:

$$\begin{aligned}\frac{\theta_{it}^{k',\text{rev}}}{\alpha_{it}^{k'}} &= \frac{\theta_{it}^{k',\text{Q}}}{\alpha_{it}^{k'}} \cdot \left(1 + \frac{\mathrm{d}P(Q_{it})}{\mathrm{d}Q_{it}} \frac{Q_{it}}{P_{it}}\right) \\ &\equiv \frac{\theta_{it}^{k',\text{Q}}}{\alpha_{it}^{k'}} \cdot (1 + \varepsilon_{P,Q,it}) \\ &\equiv \mu_{it} \cdot (1 + \varepsilon_{P,Q,it}) \\ &= 1, \end{aligned} \tag{49}$$

where the last equality follows directly from Lerner's monopoly pricing rule, i.e. $\mu_{it} = (1+\varepsilon_{P,Q,it})^{-1}$. Based on this result, Bond et al. (2021) conclude that it is basically hopeless to retrieve markups through the production approach whenever data on physical output is not available. Estimates of markups using deflated revenues that do not equal unity then indicate that Assumptions **I–VI** and/or 1–5 must be violated. While this is an issue for the estimation of mark*ups*, we argue that it does not pose any problems when estimating mark*downs*. This can be shown most clearly through the following proposition:

PROPOSITION 4. Let $\theta_{it}^{j,Q} \equiv \frac{\partial \ln(Q_{it})}{\partial \ln(X_j)}$ and $\theta_{it}^{j,\text{rev}} \equiv \frac{\partial \ln(P(Q_{it}) \cdot Q_{it})}{\partial \ln(X_j)}$ denote the output and revenue elasticities with respect to some differentiable input j, respectively. Furthermore,

let $\alpha_{it}^{j} \equiv \frac{V_{it}^{j} \cdot X_{it}^{j}}{P_{it} Q_{it}}$ denote the revenue share of input j. Then, we have:

$$\frac{\theta_{it}^{\ell,\text{rev}}}{\alpha_{it}^{\ell}} \Big/ \frac{\theta_{it}^{M,\text{rev}}}{\alpha_{it}^{M}} = \frac{\theta_{it}^{\ell,Q}}{\alpha_{it}^{\ell}} \Big/ \frac{\theta_{it}^{M,Q}}{\alpha_{it}^{M}}. \tag{50}$$

That is, it is sufficient to estimate *revenue* elasticities in order to construct markdowns on labor inputs.

Proof. We drop firm and time subscripts to ease notation. To prove the proposition, it is sufficient to show that $\frac{\theta_{\ell}^{\text{rev}}}{\theta_{M}^{\text{rev}}} = \frac{\theta_{\ell}^{Q}}{\theta_{M}^{Q}}$ is true. To do so, note that we have:

$$\begin{aligned}
\theta_{j}^{\text{rev}} &\equiv \frac{\partial \left[P(Q) \cdot Q\right]}{\partial X_{j}} \cdot \frac{X_{j}}{P(Q)Q} \\
&= \left[P'(Q)Q + P(Q)\right] \cdot \frac{\partial Q}{\partial X_{j}} \cdot \frac{X_{j}}{P(Q)Q}.
\end{aligned}$$

Then, with some abuse of notation, it immediately follows that:

$$\begin{aligned}
\frac{\theta_{\ell}^{\text{rev}}}{\theta_{M}^{\text{rev}}} &= \frac{\left[P'(Q)Q + P(Q)\right] \cdot \frac{\partial Q}{\partial \ell} \cdot \frac{\ell}{P(Q)Q}}{\left[P'(Q)Q + P(Q)\right] \cdot \frac{\partial Q}{\partial M} \cdot \frac{M}{P(Q)Q}} \\
&= \frac{\frac{\partial Q}{\partial \ell} \cdot \frac{\ell}{Q}}{\frac{\partial Q}{\partial M} \cdot \frac{M}{Q}} \\
&\equiv \frac{\theta_{\ell}^{Q}}{\theta_{M}^{Q}},
\end{aligned}$$

which is exactly what we wanted to show. □

The proposition shows that the bias occurring from proxying physical output with deflated revenues cancels out, since it appears in both the numerator and denominator (i.e., markup) of the markdown expression in a multiplicative manner. Thus, the lack of data availability on physical output would only affect our results if we were interested in estimating markups separately. As a result, the main point of critique by Bond et al. (2021) does not apply to markdowns.

O.5.2 Demand shifters

Another point of critique on the production approach in Bond et al. (2021) revolves around the assumption of inputs being solely used for the production of output (i.e., Assumption **VI**). However, in reality, some inputs can also be used for activities to shift demand, such as marketing or advertising. When inputs are also used to shift (or "influence," using the terminology of Bond et al., 2021) demand, then the markup formula in Equation (3) is no longer correct.

To see this, consider an environment in which each input X_{it}^k can be used for either the production of output $X_{it}^{k,Q}$ or to shift demand $X_{it}^{k,D}$. Then, assume that a firm's inverse demand function is of the following form:

$$P(Q_{it}, D_{it}) \text{ s.t. } D_{it} = \mathscr{D}(\mathbf{X}_{it}^D), \tag{51}$$

where all functions are differentiable in their arguments and $\mathbf{X}_{it}^D = (X_{it}^{1,D}, \ldots, X_{it}^{K,D})'$ are those parts of each input that are used for shifting demand. Hence, by construction, we have $\mathbf{X}_{it} = \mathbf{X}_{it}^D + \mathbf{X}_{it}^Q$.

If we let k' be some flexible input, then Hall's (1988) formula only holds for that part dedicated to production; i.e., we have:

$$\mu_{it} = \frac{\theta_{it}^{k',Q}}{\alpha_{it}^{k',Q}}. \tag{52}$$

Bond et al. (2021) argue that, for most data sets, we can only observe $X_{it}^{k'}$ and its expenditure, but not its components $X_{it}^{k',Q}$ and $X_{it}^{k',D}$ separately. If one would apply formula (3) to $X_{it}^{k'}$ rather than $X_{it}^{k',Q}$, we would obtain a biased estimate of the markup:

$$\mu_{it} \cdot \frac{\varepsilon_{X^{k',Q},X^{k'}}}{1 + \frac{X_{it}^{k',D}}{X_{it}^{k',Q}}}, \tag{53}$$

where $\varepsilon_{X^{k',Q},X^{k'}}$ denotes by what percentage the usage of input k' for production purposes increases if total expenditure on input k' is raised by 1 percent. If Assumption **VI** holds, then we must have $\varepsilon_{X^{k',Q},X^{k'}} = 1$ and $X_{it}^{k',D} = 0$.

In our baseline estimates, we adopt the definition for material inputs as used by the Census Bureau, which includes contract work. It is not unlikely that some of this contracted labor is used for activities such as marketing; even though it is less likely for manufacturers. However, our results are robust to using an alternative definition for materials as proposed by Kehrig (2015), in which contract work is disregarded and information on inventories for materials is used instead. Under this definition, material inputs only consist of materials and parts. Its exact definition can be taken from Section 16A1 of Form MA-10000, which we documented below for convenience.

Table XVI: Description of what constitutes "material inputs" from Section 16A1 in Form MA-10000 of ASM.

MATERIALS		PARTS	CONTAINERS	SUPPLIES	
Lumber	Cement	Pumps	Pails	Bolts, screw and nuts	Cleaning supplies
Plywood	Clay	Wheels	Drums and barrels	Drills, tools, dies, jigs and	Stationary and
Paper	Glass	Bearings	Tubes	fixtures which are charged to	office supplies
Resins	Steel sheet	Engines	Boxes and bags	current accounts	First aid and
Sulfuric acid	Steel scrap	Gears	Crates	Welding rods, electrodes and	safety supplies
Alcohols	Copper rods	Motors		acetylene	Dunnage
Rubber	Iron castings	Hardware		Lubricating oils	Water
Coking coal	Metal stampings	Compressors			
Crude petroleum	Wire				

If we impose the assumption that none of the expenditures on materials and parts are used to shift demand, which we believe to be reasonable given the table above, then there are no issues with the denominator of our markdown definition. On the other hand, the numerator of our markdown definition consists of the "labor markup." If some fraction of total labor is used to shift demand, then our markdown estimates are biased. This is formally shown in the proposition below.

PROPOSITION 5. Let there exist some input $k' \neq \ell$ that satisfies Assumptions **I–VI**. If labor ℓ does not satisfy Assumption **VI** and firms possess monopsony power but cannot discriminate between different workers, then the ratio estimator for the markdown in (3) retrieves:

$$\hat{\nu} = \left[\varepsilon_S^{-1}\frac{\ell^Q}{\ell} + 1\right] \cdot \frac{\varepsilon_{\ell^Q,\ell}}{1 + \frac{\ell^D}{\ell^Q}}, \tag{54}$$

where total labor $\ell \equiv \ell^Q + \ell^D$ is the sum of labor used for production and shifting demand, respectively. Furthermore, $\varepsilon_{\ell_Q,\ell}$ denotes the elasticity of labor used for output with respect to total labor. If labor ℓ does not satisfy Assumption **VI** but firms are allowed to discriminate between different workers, then the ratio estimator for the markdown in (3) retrieves instead:

$$\hat{\nu} = \nu_{\ell^Q} \cdot \frac{\varepsilon_{\ell^Q,\ell}}{1 + \frac{\ell^D}{\ell^Q}}, \tag{55}$$

where ν_{ℓ^Q} denotes the markdown a firm charges on its production workers.

Proof. We follow the proof of Bond et al. (2021) closely. For notational convenience, we drop firm and time subscripts. A firm's profit maximization problem reads as:

$$\max_{Q,D \geq 0} P(Q, D) \cdot Q - C_Q(Q) - C_D(D), \tag{56}$$

where $C_D(D)$ denotes the cost of reaching a level D for the demand shifter. This results in the two first order conditions:

$$(1 + \varepsilon_{P,Q})^{-1} = \mu \tag{57}$$

$$\varepsilon_{P,D} = \frac{\frac{\mathrm{d}C_D(D)}{\mathrm{d}D} \cdot D}{P(Q)Q}. \tag{58}$$

Assuming that a firm has monopsony power but faces a residual labor supply curve only for its total stock of workers, the first-order conditions for ℓ^Q and ℓ^D for the cost minimization problem give us:

$$\left[\varepsilon_S^{-1} \frac{\ell^Q}{\ell} + 1\right] \cdot \mu = \frac{\varepsilon_{Q,\ell^Q}}{\alpha_{\ell^Q}} \tag{59}$$

$$\left[\varepsilon_S^{-1} \frac{\ell^D}{\ell} + 1\right] = \frac{\frac{\mathrm{d}C_D(D)}{\mathrm{d}D} \cdot D}{P(Q)Q} \cdot \frac{\varepsilon_{D,\ell^D}}{\alpha_{\ell^D}}, \tag{60}$$

where we defined $\varepsilon_{D,\ell^D} = \frac{\partial \mathscr{D}(\mathbf{X}^D)}{\partial \ell^D} \frac{\ell^D}{\mathscr{D}(\mathbf{X}^D)}$. Then, we get:

$$\begin{aligned} \alpha_\ell &= \alpha_{\ell^Q} + \alpha_{\ell^D} \\ &= (1 + \varepsilon_{P,Q})\varepsilon_{Q,\ell^Q} \left[\varepsilon_S^{-1} \frac{\ell^Q}{\ell} + 1\right]^{-1} + \varepsilon_{P,D}\varepsilon_{D,\ell^D} \left[\varepsilon_S^{-1} \frac{\ell^D}{\ell} + 1\right]^{-1} \end{aligned} \tag{61}$$

where $\varepsilon_{Q,\ell^Q} = \theta^{\ell^Q,Q}$ is the output elasticity with respect to labor for production purposes. Similarly, we define ε_{D,ℓ^D} as the demand shifter elasticity with respect to labor for "influencing" purposes. Then, the numerator for our markdown expression in Equation (3) using *total* labor ℓ is equal to:

$$\begin{aligned}
\frac{\theta^{\ell,Q}}{\alpha_\ell} &= \frac{\varepsilon_{Q,\ell^Q} \cdot \varepsilon_{\ell^Q,\ell}}{\alpha_\ell} \\
&= \frac{\varepsilon_{Q,\ell^Q} \cdot \varepsilon_{\ell^Q,\ell}}{(1+\varepsilon_{P,Q})\varepsilon_{Q,\ell^Q}\left[\varepsilon_S^{-1}\frac{\ell^Q}{\ell}+1\right]^{-1} + \varepsilon_{P,D}\varepsilon_{D,\ell^D}\left[\varepsilon_S^{-1}\frac{\ell^D}{\ell}+1\right]^{-1}} \\
&= \frac{\varepsilon_{\ell^Q,\ell}}{\mu^{-1}\left[\varepsilon_S^{-1}\frac{\ell^Q}{\ell}+1\right]^{-1} + \frac{\varepsilon_{P,D}\varepsilon_{D,\ell^D}}{\varepsilon_{Q,\ell^Q}}\left[\varepsilon_S^{-1}\frac{\ell^D}{\ell}+1\right]^{-1}} \\
&= \mu \cdot \left[\varepsilon_S^{-1}\frac{\ell^Q}{\ell}+1\right] \cdot \frac{\varepsilon_{\ell^Q,\ell}}{1+\frac{\alpha_{\ell^D}}{\alpha_{\ell^Q}}} \\
&= \mu \cdot \left[\varepsilon_S^{-1}\frac{\ell^Q}{\ell}+1\right] \cdot \frac{\varepsilon_{\ell^Q,\ell}}{1+\frac{\ell^D}{\ell^Q}}.
\end{aligned} \tag{62}$$

If there exists some input $k' \neq \ell$ that satisfies Assumptions **I**–**VI**, then we get an unbiased estimate for markups. As a result, we must have:

$$\begin{aligned}
\hat{\nu} &= \frac{\theta^{\ell,Q}}{\alpha_\ell}\left(\frac{\theta^{k',Q}}{\alpha_{k'}}\right)^{-1} \\
&= \left[\varepsilon_S^{-1}\frac{\ell^Q}{\ell}+1\right] \cdot \frac{\varepsilon_{\ell^Q,\ell}}{1+\frac{\ell^D}{\ell^Q}},
\end{aligned} \tag{63}$$

which covers the case whenever a firm faces a residual labor supply curve as function of only its *total* stock of workers. This is similar to the case in Bond et al. (2021), in which it is assumed that production and nonproduction workers are compensated at an identical wage rate. The derivation for the case in which a firm faces different residual labor supply curves for its production and nonproduction workers is almost identical. Note that a firm can then charge different markdowns for different workers. We only need to replace $\left[\varepsilon_S^{-1}\frac{\ell^Q}{\ell}+1\right]$ with $\left[\varepsilon_{S,\ell^Q}^{-1}+1\right] \equiv \nu_{\ell^Q}$ and $\left[\varepsilon_S^{-1}\frac{\ell^D}{\ell}+1\right]$ with $\left[\varepsilon_{S,\ell^D}^{-1}+1\right] \equiv \nu_{\ell^D}$. Then, Expression (62) becomes:

$$\hat{\nu} = \nu_{\ell^Q} \cdot \frac{\varepsilon_{\ell^Q,\ell}}{1+\frac{\ell^D}{\ell^Q}}, \tag{64}$$

which is exactly what we wanted to show. □

If labor was used for production only, then we must have $\varepsilon_{\ell_Q,\ell} = 1$, $\ell = \ell^Q$ and $\ell^D = 0$, and our markdown estimates would feature no bias(es), since $\hat{\nu} = \nu$. Bond et al. (2021) point out that bias-free estimates can be obtained if labor inputs used for production and "influencing demand" were observed separately. Even though our baseline estimates are somewhat subject to this point of critique in Bond et al. (2021), our markdown results for production and nonproduction workers (which are estimated separately) corroborate our baseline results. It supports the observation that it is unlikely that manufacturers spend a large fraction of their workforce for nonproduction purposes (see Dey, Houseman and Polivka, 2012). As a result, it is reasonable in our setting to have $\varepsilon_{\ell^Q,\ell} \approx 1$ and $\frac{\ell^D}{\ell^Q} \approx 0$.

O.5.3 Scalar unobservable assumption

The last point of critique in Bond et al. (2021) relates to the scalar unobservable assumption of the proxy variable methodology. Bond et al. (2021) argue that this assumption cannot be satisfied whenever firms possess market power. Whenever this is the case, the econometrician also needs to observe a firm's marginal cost of production. This point is formally illustrated below through a simple example.

PROPOSITION 6. If a monopolist is faced with some differentiable, downward-sloping demand curve and is endowed with a Cobb-Douglas production technology, then there exist parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_\omega, \alpha_k, \alpha'_{\mathbf{p}}, \alpha_{\text{MC}})'$ such that its optimal input demand schedule for materials under market power satisfies:

$$\begin{aligned} m_{it}(k_{it}, \omega_{it}, mc^*_{it}) &= \alpha_0 + \alpha_\omega \cdot \omega_{it} + \alpha_k \cdot k_{it} + \alpha'_{\mathbf{p}}\mathbf{p}_t + \alpha_{\text{MC}}(p^*_{it} - \ln(\mu^*_{it})) \\ &= \alpha_0 + \alpha_\omega \cdot \omega_{it} + \alpha_k \cdot k_{it} + \alpha'_{\mathbf{p}}\mathbf{p}_t + \alpha_{\text{MC}} \cdot mc^*_{it}. \end{aligned} \tag{65}$$

That is, the optimal input demand schedule for materials depends on idiosyncratic productivity and a firm's marginal cost of production.

Proof. The monopolist's profit maximization problem becomes:

$$\max_{K_{it}, L_{it}, M_{it} \geq 0} P_t(Q_{it})Q_{it} - C(Q_{it}) \;\text{ s.t. }\; Q_{it} = \exp(\omega_{it}) K_{it}^{\beta_K} L_{it}^{\beta_L} M_{it}^{\beta_M}. \tag{66}$$

It is easy to show that the firm's optimal input demand schedule for materials is:

$$M_{it} = \left(\frac{W_t}{\beta_L}\right)^{\frac{\beta_L}{\beta_L+\beta_M}} \cdot \left(\frac{P_t^M}{\beta_M}\right)^{-\frac{\beta_L}{\beta_L+\beta_M}} \cdot \left(\frac{Q_{it}}{\exp(\omega_{it})K_{it}^{\beta_K}}\right)^{\frac{1}{\beta_L+\beta_M}}, \tag{67}$$

which leads to the Cobb-Douglas cost function (conditional on a given level of output and capital):

$$C(Q_{it}) = (\beta_L + \beta_M) \cdot \left(\frac{W_t}{\beta_L}\right)^{\frac{\beta_L}{\beta_L+\beta_M}} \left(\frac{P_t^M}{\beta_M}\right)^{\frac{\beta_M}{\beta_L+\beta_M}} \cdot \left(\frac{Q_{it}}{\exp(\omega_{it})K_{it}^{\beta_K}}\right)^{\frac{1}{\beta_L+\beta_M}}. \tag{68}$$

Following the Lerner index pricing formula, a firm's optimal output is pinned down by:

$$\begin{aligned} \mu_{it}^* &\equiv \frac{\varepsilon_D(Q_{it}^*)}{\varepsilon_D(Q_{it}^*) - 1} \\ &= \frac{P_t(Q_{it}^*)}{C'(Q_{it}^*)}, \end{aligned} \tag{69}$$

which, using (68) for the marginal cost of production, we can rearrange as:

$$\begin{aligned} C'(Q_{it}) &= \left(\frac{W_t}{\beta_L}\right)^{\frac{\beta_L}{\beta_L+\beta_M}} \left(\frac{P_t^M}{\beta_M}\right)^{\frac{\beta_M}{\beta_L+\beta_M}} \cdot \left(\frac{1}{\exp(\omega_{it})K_{it}^{\beta_K}}\right)^{\frac{1}{\beta_L+\beta_M}} Q_{it}^{\frac{1-\beta_L-\beta_M}{\beta_L+\beta_M}} \\ &= P_t(Q_{it})\mu_{it}^{-1}. \end{aligned} \tag{70}$$

Using (70), we solve for the optimal level of output Q_{it}^*:

$$Q_{it}^* = \frac{P_{it}^*}{\mu_{it}^*} \cdot \left(\frac{W_t}{\beta_L}\right)^{-\frac{\beta_L}{1-\beta_L-\beta_M}} \left(\frac{P_t^M}{\beta_M}\right)^{-\frac{\beta_M}{1-\beta_L-\beta_M}} \left(\exp(\omega_{it})K_{it}^{\beta_K}\right)^{\frac{1}{1-\beta_L-\beta_M}}. \tag{71}$$

Plugging (71) into (67) and taking natural logs, there exist values for α_0, α_ω, α_k, $\alpha_{\mathbf{p}}$, and α_{MC} such that the optimal input demand schedule for materials *under market power* satisfies:

$$\begin{aligned} m_{it}(k_{it}, \omega_{it}, mc_{it}^*) &= \alpha_0 + \alpha_\omega \cdot \omega_{it} + \alpha_k \cdot k_{it} + \alpha_{\mathbf{p}}' \mathbf{p}_t + \alpha_{\text{MC}}(p_{it}^* - \ln(\mu_{it}^*)) \\ &= \alpha_0 + \alpha_\omega \cdot \omega_{it} + \alpha_k \cdot k_{it} + \alpha_{\mathbf{p}}' \mathbf{p}_t + \alpha_{\text{MC}} \cdot mc_{it}^*. \end{aligned} \tag{72}$$

As a result, a firm's input demand schedule for materials becomes a direct function of its marginal cost of production whenever it has pricing power. □

The proposition illustrates that the econometrician needs to observe both firm-level productivity and its marginal cost of production, contradicting the scalar unobservable assumption. As a result, Bond et al. (2021) argue that other estimators, in particular those that do *not* rely on the scalar unobservable assumption, should be used in order to estimate production function parameters. In particular, they refer to the estimator from Blundell and Bond (2000).

In the following, we evaluate a set of proxy variable estimators and two estimators that do not rely on the scalar unobservable assumption. Regarding the latter two, we use the dynamic panel IV estimator from Blundell and Bond (2000) and the estimator from Hu, Huang and Sasaki (2020). To evaluate the performance of all estimators, we apply them to simulated data. In particular, we adopt the third data-generating process (DGP) from Ackerberg, Caves and Frazer (2015) (or "ACF – DGP3") which is least favorable to the family of proxy variable estimators. The latter paper only allows for gross output specifications in which materials enter in a Leontief fashion. We replicate ACF to the letter, but we also look at the performance of production function estimators whenever gross output is also Cobb-Douglas in materials. This requires us to specify a process for material prices. We follow ACF and assume it follows an AR(1) process in natural logs, i.e. $\ln(P_t^M) = \varphi_M \cdot \ln(P_{t-1}^M) + \varepsilon_{it}^M$. Online Appendix O.5.4 contains more details on what changes occur whenever we allow material inputs to enter production in a Cobb-Douglas fashion. We use the same parameter values as ACF unless otherwise specified.

ACF – DGP3: FULL COBB-DOUGLAS PRODUCTION. We start out with the case in which material inputs enter the production function in a non-Leontief fashion: i.e., $Y_{it} = \exp(\omega_{it}) K_{it}^{\beta_k} L_{it}^{\beta_\ell} M_{it}^{\beta_m}$ with $\beta_m \in (0, 1)$. Similar to the wage process in ACF – DGP3, we assume that prices for material inputs are idiosyncratic and follow an AR(1) process. This introduces two additional parameters compared to Ackerberg, Caves and Frazer (2015). We set all of the parameters in an identical fashion to the latter paper unless otherwise noted. Obviously, we cannot do this for the parameters φ_M and σ_M^2.

To solve this issue, we set $\varphi_M = 0.799$ based on evidence from Atalay (2014) and $\sigma_M^2 =$

$\sigma_W^2 = 0.1^2$.[63] Furthermore, we have to adjust the production function parameters to reflect a gross output (rather than a value added) specification. We choose $\beta_k = 0.1$, $\beta_\ell = 0.25$, and $\beta_m = 0.65$, which reflect data from the ASM/CM. To stay close to Ackerberg, Caves and Frazer (2015), we allow for optimization errors in labor by setting $\sigma_{\xi^\ell} = 0.37$. The results are not affected qualitatively by this choice, though. The simulation results of the production function estimation procedure can be found in Table XVII.

The translog specification approximates the Cobb-Douglas production function in the best manner. Each cross and second-order term is not statistically significant (at the 5 percent level). Note that the parameters are estimated with some bias, but this is to be expected since the scalar unobservable assumption is violated. Furthermore, the production function is not of the Leontief form: Ackerberg, Caves and Frazer (2015) have pointed out that the family of proxy variable estimators then generates biased results. Nevertheless, the estimated parameters are very close to their true values. In fact, the true parameters are contained within the 95 percent confidence intervals generated with the Monte Carlo simulations. Somewhat surprisingly, the estimation results for the Cobb-Douglas specification are less precisely estimated when compared to its translog counterpart.

Our simulation results also indicate that other estimators from the proxy variable family do not perform as well. In particular, the coefficient for material inputs is always heavily underestimated. To assess the importance of the scalar unobservable assumption, we test the performance of the estimators mentioned in Blundell and Bond (2000) (DPD-IV) and Hu, Huang and Sasaki (2020) (HHS).

As can be seen from the table below, the DPD-IV estimator from Blundell and Bond (2000) performs quite poorly, even when allowing for measurement error in output. In particular, capital coefficients are estimated to be implausibly large. This estimator is predicated upon several layers of differencing, and we suspect this approach eliminates the variation that is necessary for identification.

[63] Higher values for σ_M^2 will increase the standard errors of our estimates but do not affect the point estimates themselves by much. All of the remaining parameters are set to their identical values in the appendix section of Ackerberg, Caves and Frazer (2015). There are two exceptions. First, we set ρ and σ_ω^2 at 0.9 and 0.2^2 instead of 0.7 and 0.3^2. We believe this reflects the U.S. data in a better fashion. Second, we leave adjustment cost parameters to be static—i.e., they do not evolve dynamically over time. However, the latter does not affect our results much and is without much loss of generality.

Table XVII: Monte Carlo results with ACF – DGP3 under nontrivial Cobb-Douglas specification.† Our preferred estimator, DLW-TL, outperforms alternative estimators.

	β_k	β_ℓ	β_m	θ_ℓ/θ_m
$\boldsymbol{\beta}_0$	0.10	0.25	0.65	0.3846
DLW-TL	0.1097	0.2212	0.6231	0.2922
	(0.0381)	(0.0553)	(0.0566)	(0.4605)
	$\beta_{k\ell}$	β_{km}	$\beta_{\ell m}$	
	0.0428	−0.0156	0.0154	
	(0.0371)	(0.0386)	(0.0374)	
	β_{k^2}	β_{ℓ^2}	β_{m^2}	
	0.0237	−0.0167	0.0493	
	(0.0501)	(0.0205)	(0.0270)	
	β_k	β_ℓ	β_m	θ_ℓ/θ_m
DLW-CD	0.0394	0.1013	0.5682	0.1817
	(0.0386)	(0.01887)	(0.0453)	(0.0450)
LP-CD	0.1078	0.2214	0.1438	1.6303
	(0.0382)	(0.0074)	(0.0317)	(0.5031)
ACF-CD	0.0689	0.2219	0.1005	2.2255
	(0.0072)	(0.0075)	(0.0077)	(0.2478)
BB-CD: MA(0)	0.3538	0.2137	0.1069	1.9722
	(0.1775)	(0.0649)	(0.0678)	(29.3905)
BB-CD: MA(1)	0.2976	0.2254	0.1061	0.3362
	(0.2358)	(0.0749)	(0.0751)	(35.5410)
HHS-CD (capital only)	0.1414	0.1103	0.4587	−0.0214
	(0.5750)	(0.5617)	(0.7696)	(6.1008)
HHS-CD (capital and labor)	0.0981	0.2329	0.6675	0.1931
	(0.3302)	(0.3200)	(0.5247)	(2.4675)

†We estimate production function parameters through the two-step proxy variable estimator of De Loecker and Warzynski (2012) (denoted by DLW-CD and DLW-TL), the two-step proxy variable estimator of Levinsohn and Petrin (2003) (LP-CD), the two-step proxy variable estimator of Ackerberg, Caves and Frazer (2015) (ACF-CD), the dynamic panel estimator of Blundell and Bond (2000) (BB-CD) and the two-step GMM-IV estimator of Hu, Huang and Sasaki (2020) (HHS-CD). Starting values of the GMM-IV minimization processes for the proxy variable estimators are based on the true parameters of the DGP. Samples are generated based on DGP3 in Ackerberg, Caves and Frazer (2015) in which input prices are serially correlated, labor is chosen before materials and investment, and labor is subject to optimization error. However, production is generated through a Cobb-Douglas specification in capital, labor, and material inputs. Furthermore, capital adjustment costs are heterogeneous but static. The table displays the mean of each estimated parameter across $S = 1000$ simulations. Standard errors, which are displayed in parentheses, are based on the standard deviation of each estimated parameter across the simulations.

We also focus on the estimator of Hu, Huang and Sasaki (2020). It is commonly assumed that labor is chosen simultaneously with material inputs. As a result, the policy function for material inputs should only contain capital as a state variable. When we impose this in our moment conditions, we see from Table XVII that the estimator from Hu, Huang and Sasaki (2020) is quite biased; performing worse than the Cobb-Douglas specification of De Loecker and Warzynski (2012). Note, however, that labor for production in period t is chosen at $t - b$ in DGP3 of Ackerberg, Caves and Frazer (2015). Thus, the model is correctly specified whenever labor is included as a state variable. The table below shows that the methodology of Hu, Huang and Sasaki (2020) does produce consistent estimates under this scenario. However, its standard errors are an order of magnitude larger than our preferred estimator.

Last, note that output elasticities are an explicit function of inputs under translog production. Thus, it could be argued that output elasticities are incorrectly estimated despite the small estimates for cross- and higher-order terms under the translog specification. It appears that this is not the case, as can be seen from the last column in Table XVII. In fact, output elasticities are also most accurately estimated under the translog estimator from De Loecker and Warzynski (2012).

ACF – DGP3: LEONTIEF PRODUCTION. For completeness, we assess the reliability of the translog specification under the *exact same* DGP3 of Ackerberg, Caves and Frazer (2015). Material inputs enter production in a Leontief fashion instead; i.e., we have $Y_{it} = \min\left\{\exp\left(\omega_{it}\right)\beta_0 K_{it}^{\beta_k} L_{it}^{\beta_\ell}, \beta_m M_{it}\right\}$. Thus, we replicate the simulated data from Ackerberg, Caves and Frazer (2015) to the letter in this case. Most contributions in the production function literature run their Monte Carlo simulations on value-added specifications; rather than gross output specifications as in the previous section.

To assess the reliability of the estimator used in our paper, we adapt it to estimate value-added production functions instead, which allows us to directly compare it to other production function estimation methodologies in the literature. We compare the Cobb-Douglas and translog estimators of De Loecker and Warzynski (2012) with the Cobb-Douglas estimators in Blundell and Bond (2000), Levinsohn and Petrin (2003), Ackerberg, Caves and Frazer (2015), and Hu, Huang and Sasaki (2020). The results can be found in Table XVIII.

Table XVIII: Monte-Carlo results with ACF – DGP3 under Leontief specification: value-added estimation.[2]

	β_k	β_ℓ	$\beta_{k\ell}$	β_{k^2}	β_{ℓ^2}
$\boldsymbol{\beta}_0$	0.40	0.60			
DLW-TL	0.4040	0.6109	0.0013	−0.0028	−0.0010
	(0.0060)	(0.0099)	(0.0017)	(0.0028)	(0.0001)
DLW-CD	0.3878	0.6048			
	(0.0170)	(0.0077)			
LP-CD	0.5839	0.4732			
	(0.0194)	(0.0076)			
ACF-CD	0.4063	0.5953			
	(0.0166)	(0.0079)			
BB-CD: MA(0)	0.2277	0.8974			
	(0.1008)	(0.0675)			
BB-CD: MA(1)	0.1501	0.8339			
	(0.1902)	(0.0737)			
HHS-CD (capital only)	0.3161	0.3634			
	(0.1186)	(0.2028)			
HHS-CD (capital and labor)	0.4144	0.6142			
	(0.0803)	(0.1126)			

[2]We estimate production function parameters through the two-step proxy variable estimator of De Loecker and Warzynski (2012) (denoted by DLW-CD and DLW-TL), the two-step proxy variable estimator of Levinsohn and Petrin (2003) (LP-CD), the two-step proxy variable estimator of Ackerberg, Caves and Frazer (2015) (ACF-CD), the dynamic panel estimator of Blundell and Bond (2000) (BB-CD), and the two-step estimator of Hu, Huang and Sasaki (2020) (HHS-CD). Starting values of the GMM-IV minimization processes for the proxy variable estimators are based on the true parameters of the DGP. Samples are generated based on DGP3 in Ackerberg, Caves and Frazer (2015), in which input prices are serially correlated, labor is chosen before materials and investment, and labor is subject to optimization error. Furthermore, capital adjustment costs are heterogeneous but static. The table displays the mean of each estimated parameter across $S = 1000$ simulations. Standard errors, which are displayed in parentheses, are based on the standard deviation of each estimated parameter across the simulations.

Unlike the results for a gross output specification, the whole family of proxy variable estimators (with the exception of Levinsohn and Petrin, 2003) produces consistent estimates. As in the previous section, we see that the DPD-IV estimator from Blundell and Bond (2000) still performs poorly; however, its bias is less severe than before. Moreover, we see that the estimator from Hu, Huang and Sasaki (2020) does produce consistent estimates, but it is crucial that the model (in particular, the state variables of the policy function for material inputs) is correctly specified. Hence, it appears that the estimator from Hu, Huang and Sasaki (2020) is quite sensitive to model misspecification. Also, its standard errors are an order of magnitude larger than our preferred translog estimator by De Loecker and Warzynski (2012).

O.5.4 Derivation of ACF – DGP3 process

In the following, we adapt the DGP in Ackerberg, Caves and Frazer (2015) to allow for material inputs to enter production in a Cobb-Douglas fashion. Conceptually, this does not change much, but the expressions, in particular the investment function, become more complicated. To ensure the validity of our results, we verify that the limits of our expressions (in which $\beta_m \to 0$ holds) coincide with those presented in Ackerberg, Caves and Frazer (2015) and Collard-Wexler and De Loecker (2020). Furthermore, we will also run our Monte Carlo experiments with the exact same DGP in Ackerberg, Caves and Frazer (2015).

We adapt the third data-generating process (DGP3) in Ackerberg, Caves and Frazer (2015) and allow for material inputs to enter the production function through a Cobb-Douglas specification. Hence, production Y_{it} is generated through:

$$Y_{it} = \exp(\omega_{it})\beta_0 K_{it}^{\beta_k} L_{it}^{\beta_\ell} M_{it}^{\beta_m}. \tag{73}$$

In the remainder of this section, we will set $\beta_0 = 1$. In the following, we will focus on DGP3 of Ackerberg, Caves and Frazer (2015): labor is chosen before material inputs, without full knowledge of productivity ω_{it}. Instead, the firm observes some intermediate level of productivity ω_{it-b} between time periods $t-1$ and t.

Wages are idiosyncratic and stochastic. In particular, we assume that (natural log) productivity, wages, and prices for material inputs follow AR(1) processes:

$$\omega_{it} = \rho \cdot \omega_{it-1} + \varepsilon_{it}^{\omega} \tag{74}$$

$$\ln(W_{it}) = \varphi_W \cdot \ln(W_{it-1}) + \varepsilon_{it}^{W} \tag{75}$$

$$\ln(P_{it}^{M}) = \varphi_M \cdot \ln(P_{it-1}^{M}) + \varepsilon_{it}^{M}, \tag{76}$$

where $\varepsilon_{it}^{e} \sim N(0, \sigma_e^2)$ for $e \in \{\omega, W, M\}$ and all shocks are independent across firms and time. To avoid the functional dependence problem, labor is chosen at time $t - b$ for some $b \in (0, 1)$ when the firm observes only some intermediate productivity ω_{it-b}. This level of productivity evolves smoothly; i.e., it satisfies:

$$\omega_{it-b} = \rho^{1-b}\omega_{it-1} + \xi_{it}^{A} \tag{77}$$

$$\omega_{it} = \rho^{b}\omega_{it-b} + \xi_{it}^{B}. \tag{78}$$

By construction, the variances of these innovations satisfy $V(\rho^b \xi_{it}^A + \xi_{it}^B) = V(\varepsilon_{it}^{\omega}) = \sigma_{\omega}^2$. We assume that investment and material inputs are chosen at time t. To solve the firm's problem, we use a backward induction strategy. At time t, given a level of capital K_{it} and labor L_{it}, a firm i chooses its optimal level of material inputs:

$$\max_{M_{it} \geq 0} P_{it}\text{exp}(\omega_{it}) K_{it}^{\beta_k} L_{it}^{\beta_\ell} M_{it}^{\beta_m} - P_{it}^{M} M_{it}$$

Assuming that output and input markets are perfectly competitive, the first order condition for M_{it} characterizes its optimal level:

$$\beta_m P_{it}\text{exp}(\omega_{it}) K_{it}^{\beta_k} L_{it}^{\beta_\ell} M_{it}^{\beta_m - 1} = P_{it}^{M}$$

Thus, we get:

$$\begin{aligned} M_{it}^{*} &\equiv \mathscr{M}_{it}(\omega_{it}, K_{it}; L_{it}) \\ &= \beta_m^{\frac{1}{1-\beta_m}} \exp\left(\frac{\omega_{it}}{1-\beta_m}\right) P_{it}^{\frac{1}{1-\beta_m}} K_{it}^{\frac{\beta_k}{1-\beta_m}} L_{it}^{\frac{\beta_\ell}{1-\beta_m}} (P_{it}^{M})^{-\frac{1}{1-\beta_m}} \end{aligned} \tag{79}$$

At time $t - b$, a firm i takes ω_{it-b} (and *not* the level of productivity ω_{it}) as given and internalizes that its labor decision affects its choice for material inputs at time t. Hence, its

maximization problem is given by:

$$\begin{aligned}
&\max_{L_{it}\geq 0} P_{it}\mathbb{E}_{it-b}\left[\exp(\omega_{it})K_{it}^{\beta_k}L_{it}^{\beta_\ell}\mathscr{M}_{it}(\omega_{it},K_{it};L_{it})^{\beta_m}\middle|\omega_{it-b}\right]-W_{it}L_{it}\\
&=\max_{L_{it}\geq 0} P_{it}^{\frac{1}{1-\beta_m}}\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]K_{it}^{\frac{\beta_k}{1-\beta_m}}L_{it}^{\frac{\beta_\ell}{1-\beta_m}}\beta_m^{\frac{\beta_m}{1-\beta_m}}(P_{it}^M)^{-\frac{\beta_m}{1-\beta_m}}-W_{it}L_{it}
\end{aligned}$$

The first order condition for labor is then characterized by:

$$\frac{\beta_\ell}{1-\beta_m}P_{it}^{\frac{1}{1-\beta_m}}\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]K_{it}^{\frac{\beta_k}{1-\beta_m}}L_{it}^{\frac{\beta_\ell-1+\beta_m}{1-\beta_m}}\beta_m^{\frac{\beta_m}{1-\beta_m}}(P_{it}^M)^{-\frac{\beta_m}{1-\beta_m}}=W_{it}$$

Then, optimal labor L_{it}^* satisfies:

$$\begin{aligned}
L_{it}^* &\equiv \mathscr{L}_{it}(\omega_{it-b},K_{it})\\
&=\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\left(\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\times\\
&\quad K_{it}^{\frac{\beta_k}{1-\beta_m-\beta_\ell}}P_{it}^{\frac{1}{1-\beta_m-\beta_\ell}}(P_{it}^M)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}W_{it}^{-\frac{(1-\beta_m)}{1-\beta_m-\beta_\ell}}
\end{aligned}\tag{80}$$

Note that the expression for $\lim_{\beta_m\to 0} L_{it}^*$ equals:

$$\beta_\ell^{\frac{1}{1-\beta_\ell}}P_{it}^{\frac{1}{1-\beta_\ell}}\left(\mathbb{E}_{it-b}\left[\exp\left(\omega_{it}\right)\middle|\omega_{it-b}\right]\right)^{\frac{1}{1-\beta_\ell}}W_{it}^{-\frac{1}{1-\beta_\ell}}K_{it}^{\frac{\beta_k}{1-\beta_\ell}}$$

which coincides with the term for labor on p. 2443 in Ackerberg, Caves and Frazer (2015). To see this, note that $\lim_{\beta_m\to 0}\beta_m^{\beta_m}=1$. To simplify, we can also write the expression for labor as:

$$\begin{aligned}
L_{it}^* &\equiv \mathscr{L}_{it}(\omega_{it-b},K_{it})\\
&=\left(\frac{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{1-\beta_m}\cdot\beta_m^{\beta_m}\cdot P_{it}\cdot\left(\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]\right)^{1-\beta_m}}{(P_{it}^M)^{\beta_m}W_{it}^{1-\beta_m}}\right)^{\frac{1}{1-\beta_m-\beta_\ell}}K_{it}^{\frac{\beta_k}{1-\beta_m-\beta_\ell}}
\end{aligned}\tag{81}$$

Given these optimal choices, we can define the following lemmas.

LEMMA 3. Under DGP1 of ACF and $\beta_k + \beta_\ell + \beta_m = 1$, revenues at the optimum can be written as:

$$P_{it}Y_{it}^* = \left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} \beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}} \exp\left(\frac{\omega_{it}}{1-\beta_m}\right) \mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} K_{it}$$
$$\times (P_{it}^M)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}} \cdot P_{it}^{\frac{1}{1-\beta_m-\beta_\ell}} \cdot W_{it}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} \tag{82}$$

Proof. We plug the optimal choices for material inputs and labor at time t and $t-b$, respectively, in the revenue function. Then, we get:

$$\begin{aligned}
P_{it}Y_{it}^* &= P_{it}\exp(\omega_{it})K_{it}^{\beta_k}\mathscr{L}_{it}(\omega_{it-b},K_{it})^{\beta_\ell}\mathscr{M}_{it}(\omega_{it},K_{it};\mathscr{L}_{it}(\omega_{it-b},K_{it}))^{\beta_m} \\
&= P_{it}^{\frac{1}{1-\beta_m}}\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)K_{it}^{\frac{\beta_k}{1-\beta_m}}\mathscr{L}_{it}(\omega_{it-b},K_{it})^{\frac{\beta_\ell}{1-\beta_m}}\beta_m^{\frac{\beta_m}{1-\beta_m}}(P_{it}^M)^{-\frac{\beta_m}{1-\beta_m}} \\
&= P_{it}^{\frac{1}{1-\beta_m}}\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\beta_m^{\frac{\beta_m}{1-\beta_m}}(P_{it}^M)^{-\frac{\beta_m}{1-\beta_m}} \\
&\times K_{it}^{\frac{\beta_k}{1-\beta_m}+\frac{\beta_k}{1-\beta_m-\beta_\ell}\frac{\beta_\ell}{1-\beta_m}}\left(\frac{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{1-\beta_m}\cdot\beta_m^{\beta_m}\cdot P_{it}\cdot\left(\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]\right)^{1-\beta_m}}{(P_{it}^M)^{\beta_m}W_{it}^{1-\beta_m}}\right)^{\frac{1}{1-\beta_m-\beta_\ell}\frac{\beta_\ell}{1-\beta_m}} \\
&= \exp\left(\frac{\omega_{it}}{1-\beta_m}\right)K_{it}\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m}\left[1+\frac{\beta_\ell}{1-\beta_m-\beta_\ell}\right]}(P_{it}^M)^{-\frac{\beta_m}{1-\beta_m}\left[1+\frac{\beta_\ell}{1-\beta_m-\beta_\ell}\right]} \\
&\times P_{it}^{\frac{1}{1-\beta_m}\left[1+\frac{\beta_\ell}{1-\beta_m-\beta_\ell}\right]}\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}W_{it}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} \\
&= \left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\cdot K_{it} \\
&\times \exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\cdot(P_{it}^M)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}\cdot P_{it}^{\frac{1}{1-\beta_m-\beta_\ell}}\cdot W_{it}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}
\end{aligned} \tag{83}$$

which is exactly what we wanted to show. □

LEMMA 4. Under DGP1 of ACF and $\beta_k + \beta_\ell + \beta_m = 1$, revenues net of payments to labor

at the optimum can be written as:

$$P_{it}Y_{it}^{*}-W_{it}L_{it}^{*}=(P_{it}^{M})^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}\cdot P_{it}^{\frac{1}{1-\beta_m-\beta_\ell}}\cdot W_{it}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}K_{it}$$
$$\times\exp\left(\frac{\rho\omega_{it-1}}{1-\beta_m}\right)\exp\left(\frac{\rho^b\xi_{it}^A}{1-\beta_m}\right)\left\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\xi_{it}^B}{1-\beta_m}\right)\right.$$
$$\left.-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\frac{\sigma_{\xi^B}^2}{2}\right)\right\} \tag{84}$$

Proof. By applying lemma 1 and the optimal equation for labor (80), we get:

$$\begin{aligned}
P_{it}Y_{it}^{*}-W_{it}L_{it}^{*} &= P_{it}^{\frac{1}{1-\beta_m-\beta_\ell}}(P_{it}^{M})^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}W_{it}^{-\frac{(1-\beta_m)}{1-\beta_m-\beta_\ell}}K_{it}\left(\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\\
&\times\left\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\right.\\
&\left.-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]\right\}\\
&= P_{it}^{\frac{1}{1-\beta_m-\beta_\ell}}(P_{it}^{M})^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}W_{it}^{-\frac{(1-\beta_m)}{1-\beta_m-\beta_\ell}}K_{it}\left(\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\\
&\times\exp\left(\frac{\rho\omega_{it-1}}{1-\beta_m}\right)\left\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\varepsilon_{it}^{\omega}}{1-\beta_m}\right)\right.\\
&\left.-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\rho^b\xi_{it}^A}{1-\beta_m}\right)\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\frac{\sigma_{\xi^B}^2}{2}\right)\right\}\\
&= P_{it}^{\frac{1}{1-\beta_m-\beta_\ell}}(P_{it}^{M})^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}W_{it}^{-\frac{(1-\beta_m)}{1-\beta_m-\beta_\ell}}\left(\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}K_{it}\\
&\times\exp\left(\frac{\rho\omega_{it-1}}{1-\beta_m}\right)\exp\left(\frac{\rho^b\xi_{it}^A}{1-\beta_m}\right)\left\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\xi_{it}^B}{1-\beta_m}\right)\right.\\
&\left.-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\frac{\sigma_{\xi^B}^2}{2}\right)\right\} \qquad (85)
\end{aligned}$$

where we exploited the fact that $\varepsilon_{it}^{\omega}=\rho^b\xi_{it}^A+\xi_{it}^B$. Then, we have showed exactly what we wanted. □

These lemmas will be extremely useful for characterizing the optimal investment function.

This is shown in the proposition below.

PROPOSITION 7. Let the environment of DGP1 in ACF hold with $Y_{it} = \exp(\omega_{it})\beta_0 K_{it}^{\beta_k} L_{it}^{\beta_\ell} M_{it}^{\beta_m}$ and $\beta_k + \beta_\ell + \beta_m = 1$. Whenever the price for output is the numéraire, the optimal investment function equals:

$$
\begin{aligned}
I_{it}^{*} = \frac{\beta}{\varphi_{it}} \beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}} & \left[\left(\frac{\beta_\ell}{1-\beta_m} \right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} - \left(\frac{\beta_\ell}{1-\beta_m} \right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}} \right] \cdot \sum_{\tau=0}^{\infty} \Bigg\{ [\beta(1-\delta)]^{\tau} \\
& \times \exp\left[\frac{\rho^{\tau+1}\omega_{it}}{1-\beta_m-\beta_\ell} - \frac{\beta_\ell \cdot \varphi_W^{\tau+1}}{1-\beta_m-\beta_\ell} \ln\left(W_{it}\right) - \frac{\beta_m \cdot \varphi_M^{\tau+1}}{1-\beta_m-\beta_\ell} \ln\left(P_{it}^{M}\right) \right. \\
& + \frac{1}{2}\left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell} \right]^2 \sigma_W^2 \sum_{s=0}^{\tau} \varphi_W^{2(\tau-s)} + \frac{1}{2}\left[\frac{\beta_m}{1-\beta_m-\beta_\ell} \right]^2 \sigma_M^2 \sum_{s=0}^{\tau} \varphi_M^{2(\tau-s)} \\
& + \frac{1}{2}\left[\frac{1}{1-\beta_m-\beta_\ell} \right]^2 \rho^2 \sum_{s=1}^{\tau} \rho^{2(\tau-s)} \sigma_\omega^2 + \frac{\sigma_{\xi^A}^2}{2} \cdot \rho^{2b} \cdot \left(\frac{1}{1-\beta_m-\beta_\ell} \right)^2 \\
& \left. + \left(\frac{1}{1-\beta_m} \right) \left(\frac{1}{1-\beta_m-\beta_\ell} \right) \frac{1}{2} \sigma_{\xi^B}^2 \right] \Bigg\}
\end{aligned}
\tag{86}
$$

Proof. By definition, a firm i's optimal level of investment I_{it}^{*} solves the following problem:

$$
\begin{aligned}
V(\mathbf{x}_{it}) = \max_{I_{it}, M_{it} \geq 0} \Bigg\{ & P_{it} Y_{it}^{*} - W_{it} L_{it}^{*} - P_{it}^{M} M_{it} - \frac{\varphi_{it} I_{it}^2}{2} + \beta \mathbb{E}_{it}\left[V(\mathbf{x}_{it+1}) | \mathbf{x}_{it} \right] \\
& \text{s.t. } K_{it+1} = (1-\delta) K_{it} + I_{it} \Bigg\}
\end{aligned}
$$

Investment I_{it} is characterized by its first order condition:

$$
\varphi_{it} I_{it} = \beta \mathbb{E}_{it}\left[\frac{\partial V(\mathbf{x}_{it+1})}{\partial K_{it+1}} \middle| \mathbf{x}_{it} \right] \tag{87}
$$

We exploit the envelope condition to characterize the partial derivative $\frac{\partial V(\mathbf{x}_{it+1})}{\partial K_{it+1}}$. More precisely, we have:

$$
\frac{\partial V(\mathbf{x}_{it})}{\partial K_{it}} = \frac{\partial}{\partial K_{it}}\left[P_{it} Y_{it}^{*} - W_{it} L_{it}^{*} \right] + (1-\delta)\beta \mathbb{E}_{it}\left[\frac{\partial V(\mathbf{x}_{it+1})}{\partial K_{it+1}} \middle| \mathbf{x}_{it} \right]
$$

$$
\begin{aligned}
&= \frac{P_{it}}{K_{it}}\exp(\omega_{it})K_{it}^{\beta_k}\mathscr{L}_{it}(\omega_{it-b},K_{it})^{\beta_\ell}\mathscr{M}_{it}(\omega_{it},K_{it};\mathscr{L}_{it}(\omega_{it-b},K_{it}))^{\beta_m} - \frac{W_{it}}{K_{it}}\mathscr{L}_{it}(\omega_{it-b},K_{it}) \\
&+ (1-\delta)\beta\mathbb{E}_{it}\left[\frac{\partial V(\mathbf{x}_{it+1})}{\partial K_{it+1}}\middle|\mathbf{x}_{it}\right] \\
&= (P_{it}^M)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}\cdot W_{it}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\mathbb{E}_{it-b}\left[\exp\left(\frac{\omega_{it}}{1-\beta_m}\right)\middle|\omega_{it-b}\right]^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\rho\omega_{it-1}}{1-\beta_m}\right) \\
&\times\exp\left(\left[\frac{1}{1-\beta_m}\right]\left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell}+1\right]\rho^b\xi_{it}^A\right)\left\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\xi_{it}^B}{1-\beta_m}\right)\right. \\
&\left.-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\frac{\sigma_{\xi^B}^2}{2}\right)\right\} + (1-\delta)\beta\mathbb{E}_{it}\left[\frac{\partial V(\mathbf{x}_{it+1})}{\partial K_{it+1}}\middle|\mathbf{x}_{it}\right] \\
&= (P_{it}^M)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}\cdot W_{it}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell}\right]\sigma_{\xi^B}^2\right) \\
&\times\exp\left(\frac{\rho\omega_{it-1}}{1-\beta_m-\beta_\ell}\right)\exp\left(\frac{\rho^b\xi_{it}^A}{1-\beta_m-\beta_\ell}\right)\left\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\xi_{it}^B}{1-\beta_m}\right)\right. \\
&\left.-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\frac{\sigma_{\xi^B}^2}{2}\right)\right\} + (1-\delta)\beta\mathbb{E}_{it}\left[\frac{\partial V(\mathbf{x}_{it+1})}{\partial K_{it+1}}\middle|\mathbf{x}_{it}\right]
\end{aligned}
\tag{88}
$$

where the second to the third equality follows from lemma 2 and applying P_{it} as the numéraire. We go from the third to the last equality by expanding the conditional expectation and collecting common terms. Note that, by assumption, investment and material inputs are chosen at time t after labor was determined in period $t-b$. Hence, we must take revenues *net of labor payments* (which are a function of physical capital K_{it}) when applying the envelope condition. Combining expresssions (87) and (88), we obtain:

$$
\begin{aligned}
\varphi_{it}I_{it} &= (P_{it}^M)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}\cdot W_{it}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell}\right]\sigma_{\xi^B}^2\right) \\
&\times\exp\left(\frac{\rho\omega_{it-1}}{1-\beta_m-\beta_\ell}\right)\exp\left(\frac{\rho^b\xi_{it}^A}{1-\beta_m-\beta_\ell}\right)\left\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\xi_{it}^B}{1-\beta_m}\right)\right. \\
&\left.-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^2\frac{\sigma_{\xi^B}^2}{2}\right)\right\} + \beta(1-\delta)\mathbb{E}_{it}\left(\varphi_{it+1}I_{it+1}|\mathbf{x}_{it}\right)
\end{aligned}
\tag{89}
$$

Iterating expression (89) forward, we get:

$$I_{it}^{*} = \frac{\beta}{\varphi_{it}} \times \mathbb{E}_{it}\Bigg[\sum_{\tau=0}^{\infty}\left[\beta(1-\delta)\right]^{\tau} W_{it+1+\tau}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\left(P_{it+1+\tau}^{M}\right)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}} \exp\left(\left[\frac{1}{1-\beta_m}\right]^{2}\left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell}\right]\sigma_{\xi^B}^{2}\right)$$
$$\times \exp\left(\frac{\rho\cdot\omega_{it+\tau}}{1-\beta_m-\beta_\ell}\right)\exp\left(\frac{\rho^{b}\xi_{it+1+\tau}^{A}}{1-\beta_m-\beta_\ell}\right)\Bigg\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\xi_{it+1+\tau}^{B}}{1-\beta_m}\right)$$
$$-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\left[\frac{1}{1-\beta_m}\right]^{2}\frac{\sigma_{\xi^B}^{2}}{2}\right)\Bigg\}\Bigg|\mathbf{x}_{it}\Bigg] \tag{90}$$

Apply the expectations operators on the productivity shocks, then we have:

$$I_{it}^{*} = \frac{\beta}{\varphi_{it}} \times \sum_{\tau=0}^{\infty}\left[\beta(1-\delta)\right]^{\tau}\mathbb{E}_{it}\left[W_{it+1+\tau}^{-\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\left(P_{it+1+\tau}^{M}\right)^{-\frac{\beta_m}{1-\beta_m-\beta_\ell}}\exp\left(\frac{\rho\omega_{it+\tau}}{1-\beta_m-\beta_\ell}\right)\Bigg|\mathbf{x}_{it}\right]$$
$$\times \exp\left(\frac{1}{1-\beta_m}\frac{1}{1-\beta_m-\beta_\ell}\frac{\sigma_{\xi^B}^{2}}{2}\right)\exp\left(\frac{\rho^{2b}}{(1-\beta_m-\beta_\ell)^{2}}\frac{\sigma_{\xi^A}^{2}}{2}\right)\Bigg\{\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}$$
$$-\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}}\beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}}\Bigg\} \tag{91}$$

The latter step is valid since it is assumed that productivity shocks are orthogonal to shocks to input prices; across time and firms. By assumption, productivity ω_{it} follows an AR(1) process. Hence, we must have:

$$\begin{aligned}\mathbb{E}_{it}\left[\exp(\omega_{it+\tau})|\mathbf{x}_{it}\right] &= \mathbb{E}_{it}\left[\exp\left(\rho^{\tau}\omega_{it}+\sum_{s=1}^{\tau}\rho^{\tau-s}\varepsilon_{it+s}^{\omega}\right)\Bigg|\mathbf{x}_{it}\right]\\ &= \mathbb{E}_{it}\left[\exp\left(\sum_{s=1}^{\tau}\rho^{\tau-s}\varepsilon_{it+s}^{\omega}\right)\Bigg|\mathbf{x}_{it}\right]\exp\left(\rho^{\tau}\omega_{it}\right)\end{aligned}$$

We can apply the same logic to input prices, so we can rewrite expression (91) as:

$$
\begin{aligned}
I_{it}^{*} = & \frac{\beta}{\varphi_{it}} \left[\left(\frac{\beta_\ell}{1-\beta_m} \right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} \beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}} - \left(\frac{\beta_\ell}{1-\beta_m} \right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}} \beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}} \right] \\
& \times \sum_{\tau=0}^{\infty} \left\{ \left[\beta(1-\delta)\right]^{\tau} W_{it}^{-\frac{\beta_\ell \cdot \varphi_W^{\tau+1}}{1-\beta_m-\beta_\ell}} \cdot \prod_{s=0}^{\tau} \exp\left(\frac{\sigma_W^2}{2} \left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell} \right]^2 \cdot \varphi_W^{2s} \right) \right. \\
& \times \left(P_{it}^{M} \right)^{-\frac{\beta_m \cdot \varphi_M^{\tau+1}}{1-\beta_m-\beta_\ell}} \cdot \prod_{s=0}^{\tau} \exp\left(\frac{\sigma_M^2}{2} \left[\frac{\beta_m}{1-\beta_m-\beta_\ell} \right]^2 \cdot \varphi_M^{2s} \right) \\
& \times \exp\left(\frac{\rho^{\tau+1}\omega_{it}}{1-\beta_m-\beta_\ell} \right) \cdot \prod_{s=1}^{\tau} \exp\left(\frac{\sigma_\omega^2}{2} \left[\frac{\rho}{1-\beta_m-\beta_\ell} \right]^2 \cdot \rho^{2(\tau-s)} \right) \\
& \left. \times \exp\left(\frac{1}{1-\beta_m} \frac{1}{1-\beta_m-\beta_\ell} \frac{\sigma_{\xi^B}^2}{2} \right) \exp\left(\frac{\rho^{2b}}{(1-\beta_m-\beta_\ell)^2} \frac{\sigma_{\xi^A}^2}{2} \right) \right\}
\end{aligned}
$$

Collecting terms, the above expression can be rewritten as:

$$
\begin{aligned}
\hat{I}_{it}^{*}(\beta_m) = & \frac{\beta}{\varphi_{it}} \beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}} \left[\left(\frac{\beta_\ell}{1-\beta_m} \right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} - \left(\frac{\beta_\ell}{1-\beta_m} \right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}} \right] \\
& \times \sum_{\tau=0}^{\infty} \left[\beta(1-\delta)\right]^{\tau} \exp\left\{ \frac{\rho^{\tau+1}\omega_{it}}{1-\beta_m-\beta_\ell} - \frac{\beta_\ell \cdot \varphi_W^{\tau+1}}{1-\beta_m-\beta_\ell} \ln\left(W_{it}\right) - \frac{\beta_m \cdot \varphi_M^{\tau+1}}{1-\beta_m-\beta_\ell} \ln\left(P_{it}^{M}\right) \right. \\
& + \frac{1}{2} \left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell} \right]^2 \sigma_W^2 \sum_{s=0}^{\tau} \varphi_W^{2(\tau-s)} + \frac{1}{2} \left[\frac{\beta_m}{1-\beta_m-\beta_\ell} \right]^2 \sigma_M^2 \sum_{s=0}^{\tau} \varphi_M^{2(\tau-s)} \\
& + \frac{1}{2} \left[\frac{1}{1-\beta_m-\beta_\ell} \right]^2 \rho^2 \sum_{s=1}^{\tau} \rho^{2(\tau-s)} \sigma_\omega^2 + \frac{\sigma_{\xi^A}^2}{2} \cdot \rho^{2b} \cdot \left(\frac{1}{1-\beta_m-\beta_\ell} \right)^2 \\
& \left. + \left(\frac{1}{1-\beta_m} \right) \left(\frac{1}{1-\beta_m-\beta_\ell} \right) \frac{1}{2} \sigma_{\xi^B}^2 \right\}
\end{aligned} \tag{92}
$$

which is exactly what we wanted to show. Note that the case in Ackerberg, Caves and Frazer (2015) can be derived as a limit of $\beta_m \to 0$. Then, we get:

$$\lim_{\beta_m \to 0} \hat{I}^*_{it}(\beta_m) = \frac{\beta}{\varphi_{it}} \sum_{\tau=0}^{\infty} \Bigg\{ \left[\beta(1-\delta)\right]^{\tau} \left[\beta_\ell^{\frac{\beta_\ell}{1-\beta_\ell}} - \beta_\ell^{\frac{1}{1-\beta_\ell}} \right] \\ \times \exp\left[\frac{\rho^{\tau+1}\omega_{it}}{1-\beta_\ell} - \frac{\beta_\ell \cdot \varphi_W^{\tau+1}}{1-\beta_\ell} \ln\left(W_{it}\right) \right. \\ + \sum_{s=0}^{\tau} \frac{\sigma_W^2}{2} \left[\frac{\beta_\ell}{1-\beta_\ell}\right]^2 \cdot \varphi_W^{2(\tau-s)} + \frac{1}{2}\left(\frac{1}{1-\beta_\ell}\right)^2 \rho^2 \cdot \sum_{s=1}^{\tau} \sigma_\omega^2 \cdot \rho^{2(\tau-s)} \\ \left. + \frac{1}{2}\left(\frac{1}{1-\beta_\ell}\right)^2 \rho^{2b} \cdot \sigma_{\xi^A}^2 + \left(\frac{1}{1-\beta_\ell}\right) \frac{\sigma_{\xi^B}^2}{2} \right] \Bigg\} \tag{93}$$

which is the equivalent of the expression for investment in Ackerberg, Caves and Frazer (2015) on their page 2446.[64] Our expression in (93) becomes identical to the one in Ackerberg, Caves and Frazer (2015) whenever $\sigma^2_{\xi^\ell} \to 0$ and we have $\beta_0 = 1$. Note that it is straightforward to allow for measurement error in labor. Whenever we have $L^{\text{err}}_{it} = L^*_{it}\exp(\xi^\ell_{it})$ such that $\mathbb{E}_{it}\left[\exp(\xi^\ell_{it})\right] = \sigma^2_{\xi^\ell}/2$, then the optimal investment function is characterized as:

$$I^*_{it} = \frac{\beta}{\varphi_{it}} \beta_m^{\frac{\beta_m}{1-\beta_m-\beta_\ell}} \left[\left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{\beta_\ell}{1-\beta_m-\beta_\ell}} \exp\left(\frac{\beta_\ell^2 \sigma^2_{\xi^\ell}}{2}\right) - \left(\frac{\beta_\ell}{1-\beta_m}\right)^{\frac{1-\beta_m}{1-\beta_m-\beta_\ell}} \exp\left(\frac{\sigma^2_{\xi^\ell}}{2}\right) \right] \\ \times \sum_{\tau=0}^{\infty} \left[\beta(1-\delta)\right]^{\tau} \exp\Bigg\{ \frac{\rho^{\tau+1}\omega_{it}}{1-\beta_m-\beta_\ell} - \frac{\beta_\ell \cdot \varphi_W^{\tau+1}}{1-\beta_m-\beta_\ell} \ln\left(W_{it}\right) - \frac{\beta_m \cdot \varphi_M^{\tau+1}}{1-\beta_m-\beta_\ell} \ln\left(P^M_{it}\right) \\ + \frac{1}{2}\left[\frac{\beta_\ell}{1-\beta_m-\beta_\ell}\right]^2 \sigma_W^2 \sum_{s=0}^{\tau} \varphi_W^{2(\tau-s)} + \frac{1}{2}\left[\frac{\beta_m}{1-\beta_m-\beta_\ell}\right]^2 \sigma_M^2 \sum_{s=0}^{\tau} \varphi_M^{2(\tau-s)} \\ + \frac{1}{2}\left[\frac{1}{1-\beta_m-\beta_\ell}\right]^2 \rho^2 \sum_{s=1}^{\tau} \rho^{2(\tau-s)} \sigma_\omega^2 + \frac{\sigma^2_{\xi^A}}{2} \cdot \rho^{2b} \cdot \left(\frac{1}{1-\beta_m-\beta_\ell}\right)^2 \\ + \left(\frac{1}{1-\beta_m}\right)\left(\frac{1}{1-\beta_m-\beta_\ell}\right) \frac{1}{2}\sigma^2_{\xi^B} \Bigg\} \tag{94}$$

[64] The only components that are different are the terms containing $\sigma^2_{\xi^A}$ and σ^2_ω. Ackerberg, Caves and Frazer (2015) find $\frac{1}{2}\left(\frac{1}{1-\beta_\ell}\right)^2 \rho_b^2 \cdot \sum_{s=1}^{\tau} \sigma_\omega^2 \cdot \rho_b^{2(\tau-s)}$ and $\frac{1}{2}\left(\frac{1}{1-\beta_\ell}\right)^2 \rho_b^2 \rho^{2\tau} \cdot \sigma^2_{\xi^A}$ instead $\frac{1}{2}\left(\frac{1}{1-\beta_\ell}\right)^2 \rho^2 \cdot \sum_{s=1}^{\tau} \sigma_\omega^2 \cdot \rho^{2(\tau-s)}$ and $\frac{1}{2}\left(\frac{1}{1-\beta_\ell}\right)^2 \rho^{2b} \cdot \sigma^2_{\xi^A}$. Given that ρ_b is not a well-defined object and ρ^τ cannot appear in those terms associated with $\sigma^2_{\xi^A}$, it is relatively safe to assume that the expressions in Ackerberg, Caves and Frazer (2015) are typos.

Given these observations, we remain. □

Note that DGP1 in Ackerberg, Caves and Frazer (2015) with optimization error in labor is equivalent to their DGP3. To complete the description of the data-generating process, we need to specify how we initalize capital, productivity and wages.

$$K_{i0} = \exp(-10) \simeq 0.0000454 \tag{95}$$

$$\omega_{i0} = \sigma_{\omega} \cdot \varepsilon_{i0} \tag{96}$$

$$W_{i0} = \sigma_{W} \cdot \varepsilon_{i0} \tag{97}$$

where $\varepsilon \sim N(0,1)$. Note that all firms start with almost zero stock of capital. Finally, we follow Ackerberg, Caves and Frazer (2015) and Collard-Wexler and De Loecker (2020), and inject measurement error in output and material inputs. More precisely, we have:

$$\ln(Y_{it}) = \ln(Y_{it}^{*}) + \varepsilon_{it}^{Y} \tag{98}$$

$$\ln(M_{it}) = \ln(M_{it}^{*}) + m_{E} \cdot \varepsilon \tag{99}$$

where $\varepsilon_{it}^{Y} \sim N(0, \sigma_{Y}^{2})$ and m_{E}^{2} is the cross-sectional variance of demeaned levels of M_{it}^{*}.

O.6 Counterfactual exercises

Our baseline estimates in section 3 imply median markdowns of 1.53. This is well in line with the meta-study by Sokolova and Sorensen (2020): our results fall around the median in their distribution of estimates for the elasticity of labor supply. Nevertheless, we further investigate the magnitude of our estimates with two sets of counterfactual exercises in the spirit of Brooks et al. (2021b).[65]

PROFIT SHARE. In our first exercise, we verify that the majority of variable profits are not accounted for by markdowns, ensuring that our markdowns are not implausibly large. To do this, note that variable profits as a fraction of revenues (also referred to as the profit share) s_π are defined as:

$$\begin{aligned} s_\pi &\equiv 1 - \alpha_K - \alpha_\ell - \alpha_M - \alpha_E \\ &= 1 - \alpha_K - \theta_\ell \cdot \nu^{-1} \cdot \mu^{-1} - \theta_M \cdot \mu^{-1} - \alpha_E \end{aligned} \tag{100}$$

where we applied our results from proposition 1 in the second equality. Then, conditional on profits only stemming from labor market power, the counterfactual profit share satisfies:

$$s_{\pi|\mu=1} = 1 - \alpha_K - \theta_\ell \cdot \nu^{-1} - \theta_M - \alpha_E \tag{101}$$

Summary statistics on profit shares and their counterfactual counterparts can be found in table XIX.

Table XIX: Actual and counterfactual profit shares.†

PROFIT SHARE	Median	Mean	25%	75%	SD
Actual	0.203	0.190	0.101	0.303	0.227
Counterfactual	0.081	0.072	0.004	0.159	0.203
Sample size	$1.393 \cdot 10^6$				

†Markdowns are estimated under the assumption of a translog specification for gross output. The flexible input is materials. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA which approximately follows a 3-digit NAICS specification. Actual profit shares are defined as variable profits relative to revenues whereas counterfactual profit shares are constructed by setting markups to unity. By doing so, we follow the counterfactual experiments of Brooks et al. (2021b). Source: Authors' calculations from ASM/CM data in 1976–2014.

[65] We thank an anonymous referee for these helpful suggestions.

We find that, for the median plant, the majority of variable profits are actually accounted for by markups. Approximately $0.081/0.203 = 40$ percent of the median plant's profits are due to labor market power which we deem as reasonable.

AGGREGATE LABOR SHARE. In the following, we evaluate the time evolution of the aggregate labor share in the absence of labor market power. Following Brooks et al. (2021b) and Kehrig and Vincent (2021), we define the labor share as payments to labor relative to value added:

$$\eta_t^\ell \equiv \frac{\sum_{i \in F_t} w_{it} \ell_{it}}{\sum_{i \in F_t} p_{it} y_{it} - p_{it}^M m_{it} - p_{it}^E e_{it}} \tag{102}$$

For this empirical exercise, we implement the definition for value added from Kehrig and Vincent (2021). The key difference, when compared to the standard definition from the Census Bureau, lies in the use of inventories for material inputs and purchased services used as intermediate inputs. These components are included by Kehrig and Vincent (2021) but not by the Census Bureau. By construction, therefore, the Census Bureau's definition for value added is smaller than that of Kehrig and Vincent (2021) which immediately implies that labor shares under the latter must be larger. However, intermediate services are not available at the plant level. Instead, Kehrig and Vincent (2021) impute the ratio of purchased services to sales at the industry level. They show that including intermediate services only has an impact on the level of the labor share and does not affect its time evolution. As a result, we will simply ignore purchased services for intermediate use.

Let a firm i's wage bill share (relative to the national economy) be equal to:

$$\omega_{it}^\ell \equiv \frac{w_{it} \ell_{it}}{\sum_{k \in F_t} w_{kt} \ell_{kt}} \tag{103}$$

Then, given our definitions of markups and markdowns, we can use equations (102) and (103) to derive the economy's aggregate labor share:

$$\begin{aligned} \left(\eta_t^\ell\right)^{-1} &= \sum_{i \in F_t} \frac{p_{it} y_{it} - p_{it}^M m_{it} - p_{it}^E e_{it}}{\sum_{i \in F_t} w_{it} \ell_{it}} \\ &= \sum_{i \in F_t} \frac{p_{it} y_{it} - p_{it}^M m_{it} - p_{it}^E e_{it}}{w_{it} \ell_{it}} \cdot \omega_{it}^\ell \end{aligned}$$

$$= \sum_{i \in F_t} \left(\frac{1 - \alpha_{it}^M - \alpha_{it}^E}{\alpha_{it}^L} \right) \cdot \omega_{it}^\ell$$
$$= \sum_{i \in F_t} \left(\frac{\nu_{it}\mu_{it}}{\theta_{it}^L} - \frac{\nu_{it}\theta_{it}^M}{\theta_{it}^L} - \frac{\alpha_{it}^E}{\alpha_{it}^L} \right) \cdot \omega_{it}^\ell$$

Hence, we can write the labor share η_t^ℓ as:

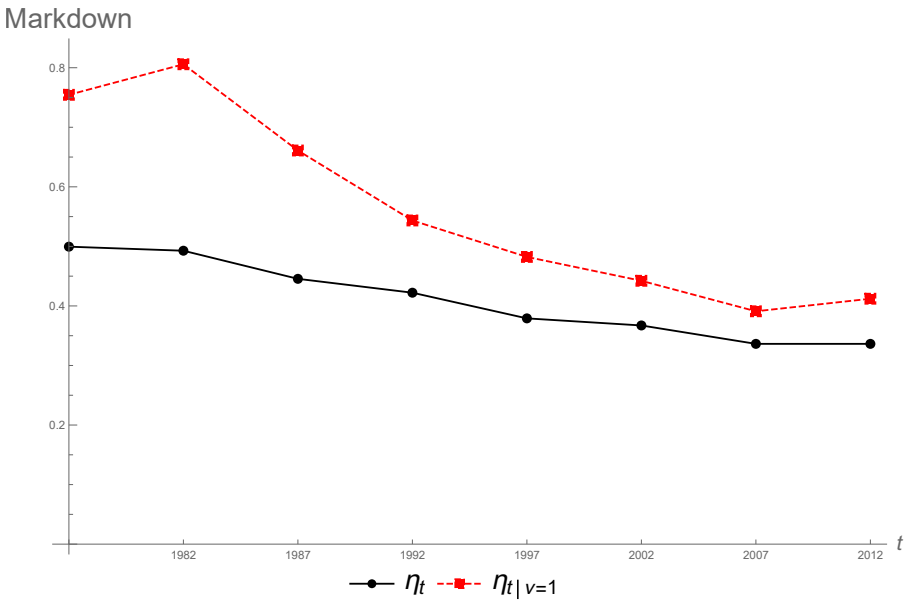
$$\eta_t^\ell = \left(\sum_{i \in F_t} \left[\nu_{it} \cdot \left(\frac{\mu_{it} - \theta_{it}^M}{\theta_{it}^L} \right) - \frac{\alpha_{it}^E}{\alpha_{it}^L} \right] \cdot \omega_{it}^\ell \right)^{-1} \tag{104}$$

According to Brooks et al. (2021b), the counterfactual labor share without *monopsony* power would then be equal to:

$$\eta_{t|\nu=1}^\ell = \left(\sum_{i \in F_t} \left[\frac{\mu_{it} - \theta_{j(i)t}^M}{\theta_{j(i)t}^L} - \frac{\alpha_{it}^E}{\alpha_{it}^L} \right] \cdot \omega_{it}^\ell \right)^{-1} \tag{105}$$

Our results are displayed in figure 14. We find that our constructed labor share declines from about 50 percent to 33 percent from 1977 to 2012. The counterfactual series implies that the labor share declined from about 75 percent in 1977 to 41 percent in 2012. Hence, the fall in the labor share would be even more pronounced in the absence of monopsony power through the lens of the counterfactual exercise in Brooks et al. (2021b). If markdowns were implausibly large, then we would expect the counterfactual labor share to be unreasonably high as well. Our counterfactual exercise does not seem to indicate that this is the case.

Figure 14: Actual and counterfactual aggregate labor shares.



Actual labor shares are defined as the aggregate wage bill divided by total value added. Counterfactual labor shares are calculated according to equation (105) in which markdowns are set to unity following Brooks et al. (2021b). Source: Authors' own calculations from quinquennial CM data from 1977–2012.

O.7 Labor market models with $\varepsilon_S \geq 0$

O.7.1 Wage posting à la Burdett-Mortensen

For ease of notation, we drop a particular firm f's index. In the wage posting model of Burdett and Mortensen (1998), a firm's law of motion for its stock of labor is given by:

$$L_t = (1 - s(w_t))L_{t-1} + R(w_t) \tag{106}$$

where $s(\cdot)$ and $R(\cdot)$ denote the separation and recruiting functions, respectively. Note that these are allowed to explicitly depend on the posted wage. In a stationary setting, we must have $L_t = \frac{R(w_t)}{s(w_t)}$. Assuming that these functions are differentiable, it is straightforward to show that labor supply elasticities satisfy:

$$\varepsilon_S = \varepsilon_{Rw} - \varepsilon_{sw} > 0$$

where $\varepsilon_{Rw,t}$ and $\varepsilon_{sw,t}$ denote separation and recruiting elasticities, respectively. The above object is strictly positive since higher wages encourage hiring and lead to fewer separations, i.e. $\varepsilon_{Rw} > 0$ and $\varepsilon_{sw} < 0$.

Formally, the separation rate is induced by some exogenous job destruction process and poaching. In particular, we have $s(w) = \delta + \lambda(1 - F(w))$. Then, $-\varepsilon_{sw} = \lambda f(w) > 0$ follows directly from the fact that probability distribution functions are non-negative. Recall that the equilibrium wage distribution function has full support on $[0, \overline{w}]$ in the baseline framework of Burdett and Mortensen (1998). Furthermore, recruitment satisfies $R(w) = R^u + \lambda \cdot \int_0^w L(x)dF(x)$ where R^u is the stock of recruits from the pool of unemployment. Note that this does not vary across wage levels w since workers' values of unemployment are normalized to zero in Burdett and Mortensen (1998). Hence, unemployed workers accept any given offer. Given this structure, it is straightforward to derive that $\varepsilon_R = \lambda \cdot \frac{f(w)L(w)w}{R(w)} > 0$. While we focus here on the canonical model of Burdett and Mortensen (1998), upward-sloping labor supply curves are also present in more generalized settings such as Bontemps, Robin and Van den Berg (2001) and Mortensen (2003).

O.7.2 Additive Random Utility Models (ARUM)

In this section, we consider a class of additive random utility models as described in Chan, Kroft and Mourifie (2019). We do so because their setup nests a variety of labor market models which we will discuss below. There are K types indexed by k which each have a mass of m_k such that $\sum_{k=1}^{K} m_k = 1$. An individual worker i with type k (which is allowed to be multidimensional) is faced with the problem of choosing among a set of employers $\mathcal{J} = \{1, 2, \ldots, J\}$. Worker choice is informed by non-pecuniary benefits, wage compensation, and some idiosyncratic term. A worker's outside option is denoted by "employer" 0. Its maximization problem is characterized by:

$$\max_{j \in \mathcal{J} \cup \{0\}} u_{kj} + w_{kj} + \varepsilon_{ij} = \max_{j \in \mathcal{J} \cup \{0\}} v_{kj} + \varepsilon_{ij}$$

The surplus function is defined as:

$$\mathcal{S}(\mathbf{v}_k) = \mathbb{E}\left[\max_{j \in \mathcal{J} \cup \{0\}} v_{kj} + \varepsilon_{ij}\right]$$

Then, Chan, Kroft and Mourifie (2019) characterize the labor supply function as:

$$\begin{aligned} L_{kj} &= m_k \cdot \Pr\left(v_{kj} + \varepsilon_{kj} \geq v_{kj'} + \varepsilon_{ij'}, \text{ for all } j' \in \mathcal{J} \cup \{0\}\right) \\ &= m_k \cdot \frac{\partial \mathcal{S}(\mathbf{v}_k)}{\partial v_{kj}} \end{aligned} \tag{107}$$

Chan, Kroft and Mourifie (2019) show that this object exists whenever ε_{ij} is independent of v_{kj} and is absolutely continuous with respect to the Lebesgue measure. Furthermore, the surplus function is convex in $\mathbf{v}_k$ under those assumptions. Hence, labor supply schedules are non-decreasing. Therefore, we have:

$$\varepsilon_S^{kj} = \frac{m_k}{L_{kj}} \frac{\partial^2 \mathcal{S}(\mathbf{v}_k)}{\partial^2 v_{kj}} w_{kj} \geq 0$$

The generalized setting of Chan, Kroft and Mourifie (2019) is quite convenient as it nests the setups of Card et al. (2018) and Lamadon, Mogstad and Setzler (2022). This can be done by appropriately defining worker types and assuming that idiosyncratic shocks are drawn from an Extreme Value Type I distribution.

O.7.3 Monopsonistic competition

In the simplest setting, upward-sloping labor supply curves are generated purely through preferences, even in the absence of strategic complementarities across firms. For instance, this would be true in a setting in which a representative household supplies a bundle of differentiated labor $\mathbf{L}_t = \{L_{it}\}_{i=1}^{K}$ and has preferences over some composite consumption bundle C_t.

Suppose the household's preferences are summarized by some function $u(C_t, \mathbf{L}_t)$ that is continuously differentiable in its arguments. Then, the schedule of labor supply functions is determined by a system of non-linear equations consisting of $\frac{(K+1)K}{2}+1$ equations. Intuitively, labor supply schedules are upward sloping whenever substitution effects dominate their income counterparts.

HORIZONTAL JOB DIFFERENTIATION. Under this class of models, workers are heterogeneous in their preferences over non-wage characteristics of a job. A simple way to capture this idea is to assume that a worker's utility is increasing in wages and decreasing in distance to work. Then, wages act as a compensating differential. Examples are Bhaskar and To (1999) and Staiger, Spetz and Phibbs (2010) who adopt frameworks in the spirit of Salop (1979).[66]

DOUBLE-NESTED CES PREFERENCES (ATKESON-BURSTEIN). Berger, Herkenhoff and Mongey (Forthcoming) consider a monopsonistic environment in the tradition of Atkeson and Burstein (2008). With some abuse of notation, preferences are characterized by:

$$u\left(C_t - \frac{1}{\overline{\varphi}^{\frac{1}{\varphi}}}\frac{\mathbf{L}_t^{1+\frac{1}{\varphi}}}{1+\frac{1}{\varphi}}\right) \text{ with } \mathbf{L}_t = \left(\int_0^1 \mathbf{L}_{jt}^{\frac{\theta+1}{\theta}} dj\right)^{\frac{\theta}{\theta+1}} \text{ and } \mathbf{L}_{jt} = \left(\sum_{f=1}^{F_j} n_{fjt}^{\frac{\eta+1}{\eta}}\right)^{\frac{\eta}{\eta+1}}$$

Thus, preferences follow the GHH specification in consumption and labor whereas labor is a double-nested CES composite. This gives rise to labor supply elasticities of the

[66] In particular, Staiger, Spetz and Phibbs (2010) assume that firms are uniformly distributed around a circle of measure one. Whenever the measure of firms N is fixed and workers' utility is increasing (decreasing) in their wage (distance to work), a firm i's labor supply function can be characterized as $L_i = \alpha + \tau^{-1}\left[w_i - \left(\frac{w_{i-1}+w_{i+1}}{2}\right)\right]$ where $\tau > 0$ denote travel costs (denoted in units of utility) per unit distance. Given this structure, we must have $\varepsilon_S > 0$.

form:

$$\varepsilon_S = \frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta}\right) \cdot s > 0$$

where $s \in [0, 1]$ is a firm's share of the industry's total payroll. The latter is guaranteed to be positive whenever $\eta > \theta$ which is the more natural assumption.

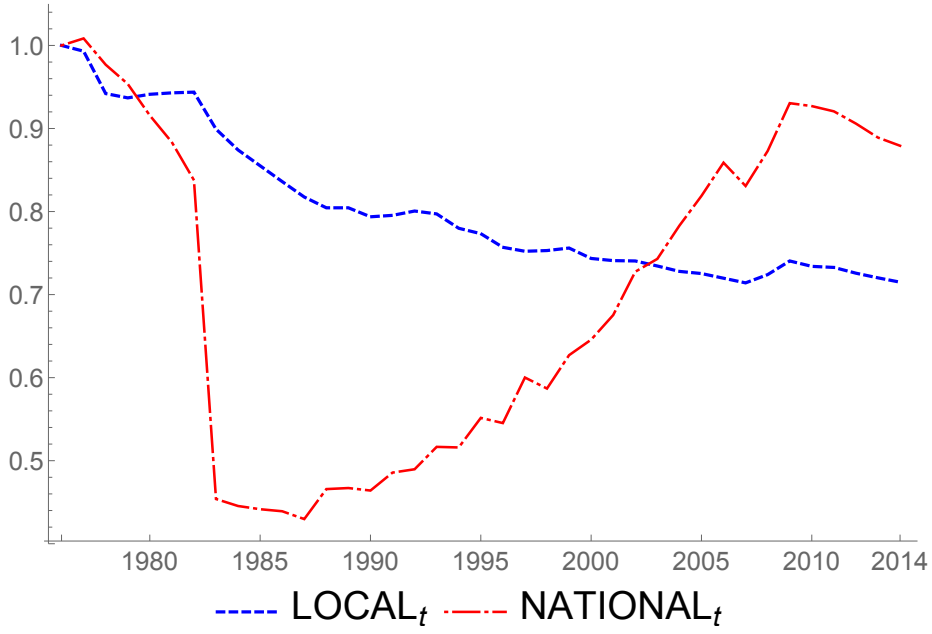
O.8 Concentration indices

NATIONAL CONCENTRATION. We construct national employment concentration, following Autor et al. (2020), as follows:

$$\begin{aligned} \text{NATIONAL}_t &= \sum_{j \in J} \omega_{jt} \text{HHI}_{jt} \\ &= \sum_{j \in J} \omega_{jt} \left[\sum_{f \in F_t(j)} \left(\frac{x_{ft}}{X_{F(j)t}} \right)^2 \right] \quad \text{s.t.} \quad X_{F(j)t} = \sum_{f' \in F_t(j)} x_{f't} \end{aligned} \tag{108}$$

Hence, national concentration is a weighted average of industry-level HHIs. We implement this measure by using employment weights and by calculating HHI_{jt} at the 3-digit NAICS-year level. The results are displayed in the figure below.

Figure 15: National employment concentration has been increasing since the early 1980s.



HHI levels are normalized relative to their initial value in 1976. Source: Authors' own calculations from LBD data from 1976–2014.

Consistent with Autor et al. (2020), we find that national employment concentration has been rising since the early 1980s. If we look at the whole available period of 1976 – 2014, then it is clear that national concentration has not been rising monotonically. In

fact, it was declining from 1976 till 1981 with a particularly sharp drop in 1982 which is consistent with Rinz (2018). While it is tempting to explain this almost continuous drop as measurement error, it is unlikely to be the case with administrative data. Furthermore, Rinz (2018) has argued that it is mainly driven by telecommunications industries and refers to a Department of Justice case in 1982 in which AT&T was required to divest itself of local telephone companies.

Regardless of the rationale behind this drop, it is clear that the time series for national employment concentration does not follow the patterns of our constructed markdown $\mathcal{V}_t$ in the least. Hence, we conclude that caution should be exercised when proxying market power with measures of concentration.

CONCENTRATION IN VACANCIES. We use two sources of data to investigate labor market concentration: employment data from the Longitudinal Business Database (LBD)—as seen in the main body—and vacancy data from Burning Glass Technologies (BGT).

The BGT data is a unique source of micro-data that contains approximately 160 million electronic job postings in the U.S. economy spanning the years 2007 and 2010–2017. These job postings were collected and assembled by BGT, an employment analytics and labor market information company, that examines over 40,000 online job boards and company websites to aggregate the job postings, parse, and deduplicate them into a systematic, machine-readable form, and create labor market analytics products. With the breadth of this coverage, the resulting database purportedly captures the near-universe of jobs posted online, estimated to be near 80 percent of total job ads. Using BGT vacancy data allows us to compute the concentration of job openings, thus zeroing in on concentration in local labor demand and computing an index of concentration that reflects how many employers are active in the hiring process in a local market.

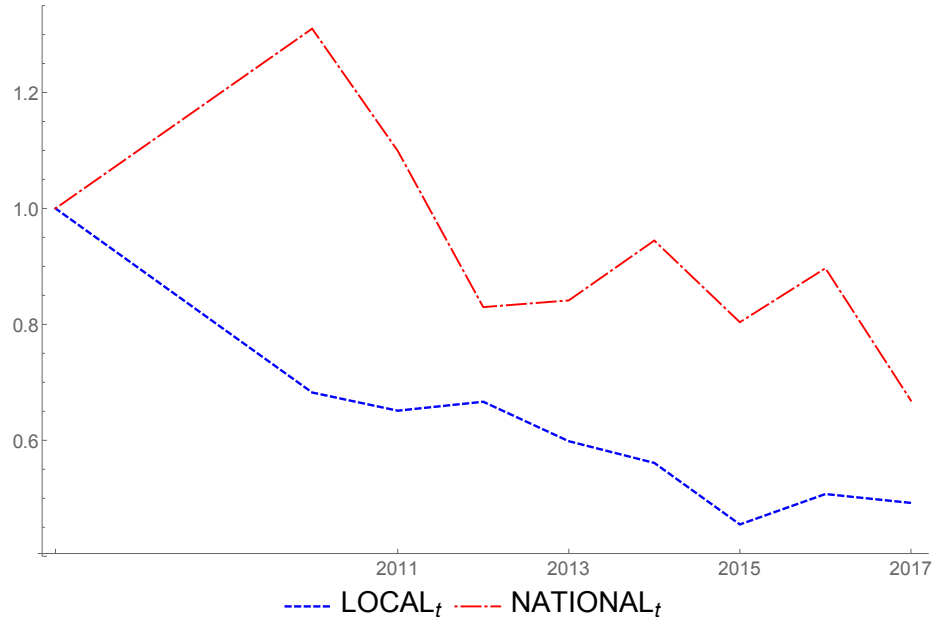
The BGT data has both extensive breadth and detail. Unlike sources of vacancy data that are based on a single job board such as `careerbuilder.com` or `monster.com`, BGT data span multiple job boards and company sites. The data are also considerably richer than sources from the Bureau of Labor Statistics (BLS), such as JOLTS (Job Openings and Labor Turnover Survey).[67] In addition to detailed information on occupation, geography,

[67] Although JOLTS asks a nationally representative sample of employers about vacancies they wish to fill in the near term, the data are typically available only at aggregated levels, and do not allow for a detailed

and employer for each vacancy, BGT data contain thousands of specific skills standardized from open text in each job posting. BGT data thus allow for a detailed analysis of vacancy flows within and across occupations, firms, and labor market areas, enabling us to document trends in employers' concentration at a very granular level.

The data, however, is not perfect. Although roughly two-thirds of hiring is replacement hiring, we expect vacancies to be somewhat skewed towards growing areas of the economy (Davis, Faberman and Haltiwanger, 2012; Lazear and Spletzer, 2012). Additionally, the BGT data only covers online vacancies. Even though vacancies for available jobs have increasingly appeared online rather than in traditional sources, it is a valid concern that the types of jobs posted online are not representative of all openings. Hershbein and Kahn (2018) provide a detailed description of the industry-occupation mix of vacancies in BGT relative to JOLTS: although BGT postings are disproportionately concentrated in occupations and industries that require greater skill, the distributions are stable across time, and the aggregate and industry trends in BGT track BLS sources closely.

Figure 16: National and local trends in the concentration of job postings.



HHI levels are normalized relative to their initial value in 2007. Observations from the Great Recession (2008–2009) are not available and are interpolated from 2007 to 2010. Source: BGT (2007, 2010–2017).

In the BGT data, we define a local labor market as an occupation-metro area pair. We define

taxonomy of local labor markets.

occupations at the 4-digit SOC level, for a total of 108 groups derived from the BLS 2010 SOC system, which aggregates "occupations with similar skills or work activities" (BLS, 2010). While our definition of occupations is considerably less detailed than the job titles available in the BGT data, we believe it offers an appropriate balance between accurately capturing the competitiveness of a market and identifying the demand for different bundles of skills.[68] Nevertheless, our results hold true for other classifications.[69] Metropolitan areas correspond to the 2013 Core-Based Statistical Areas (CBSA) with a population over 50,000. As a result, there are 382 metro areas in our final BGT dataset. In the end, we identify 41,256 local labor markets in the BGT data.

We regard vacancies concentration as the closest measure to the concentration faced by job seekers in a specific (local or national) labor market. We construct local and national concentration measures of vacancies using BGT data. Market-level HHIs are aggregated through their respective vacancy shares.[70] Figure 16 plots the time series of the aggregate local and national concentration of vacancies and shows that local concentration is markedly decreasing over time. Specifically, the local HHI of vacancies drops in the postrecession period 2010–2017 by approximately 20 percent. The decrease is even more dramatic if we consider the change between 2007 and 2017—though it is to be noted that the BGT data is not available during 2008–2009. Note that the pattern for the national concentration of vacancies is comparable to its employment counterpart.

[68] Indeed, too fine an occupational classification would mechanically lead to a small number of firms posting jobs in each market. This would bias our estimates of labor market concentration upward. On the other hand, too broad an occupational classification would erase important distinctions between heterogeneous skills used in different occupations. Even though many studies find that broad occupational changes are not uncommon in U.S. labor markets (Huckfeldt, 2017; Macaluso, 2019), especially for laid-off workers, we choose the 4-digit SOC level as a useful compromise.

[69] Examples of 4-digit SOC occupations among Production ones are Food Processing Workers (5130), Assemblers and Fabricators (5120), Textile, Apparel, and Furnishings Workers (5160), and Plant and System Operators (5180).

[70] Our results are quantitatively unaffected whenever we use employment shares instead.