

Micro and Macro Perspectives on Production-Based Markups

John Fernald
Amit Gandhi
Dimitrije Ruzic
James Traina*

July 2026

Abstract

We review the “production approach” to estimating markups, the ratio of price to marginal cost. The approach is scalable. It requires no model of consumer demand or market structure and applies broadly across firms, industries, and time. We organize our review around a simple insight: the production-based markup is a residual. Like the Solow residual, it is clean in theory but potentially contaminated by misspecification and mismeasurement. This framing helps explain why small differences in implementation can produce starkly different results from the same data. In some cases, estimated markups have risen sharply. In others, they have not. We provide conceptual rationales for these disagreements, offer practical guidance on data and estimation, and call for greater transparency about how much of the variation attributed to markups may instead reflect technology.

JEL Codes: D24, D43, E22, E23, L11, L16, O33, O47

*Fernald: john.fernald@insead.edu. Gandhi: amitgandhi@gmail.com. Ruzic: dimitrije.ruzic@insead.edu. Traina: james.traina@nyu.edu. Manuscript in preparation for the Journal of Economic Literature. We thank Susanto Basu, Arshia Hashemi, Dan Sichel, seminar participants at INSEAD, the Federal Reserve Banks of San Francisco and Richmond, and the University of Arizona, as well as our editor David Romer and several excellent anonymous referees for helpful comments and feedback.

1 Introduction

We review the production approach to estimating markups. Intuitively, markups are defined as prices relative to marginal costs, and marginal costs reflect the production constraints facing firms. The increased availability of firm-level microdata and new econometric methods make it easier to estimate production functions and infer markups. Our review focuses on this recent firm-level literature, with a goal of bridging micro and macro perspectives. Questions of market power matter for many fields, from industrial organization (IO) to international trade to macroeconomics. We highlight the challenges in moving from firm-level analyses to industry and economy-wide conclusions.

Hall (1986, 1988, 1990) introduced the production approach. He showed that cost minimization implies the markup μ can be written in a remarkably simple way as the ratio of a flexible input's output elasticity γ to the cost of that input as a share of total revenue, s . That is, $\mu = \gamma/s$. Markups can thus be inferred from producer behavior without specifying the demand system or market structure. This makes it attractive compared to typical methods from IO that are tailored to narrow industries. The production approach scales across firms, industries, and time.

Hall and a body of work that followed apply his insight to industry data, with output and all inputs in growth rates. De Loecker and Warzynski (2012) highlight challenges in applying Hall's industry approach to firm-level data. They propose a variant of Hall's method that focuses on a single input and uses modern methods for estimating firm production functions.

In recent years, interest in market power has soared. The interest reflects both the possibility that markups have risen sharply and the prospect that rising markups offer a unified explanation for a range of macroeconomic puzzles (De Loecker et al., 2020). These include declining labor shares, rising economic profits, slow productivity growth, weak investment, and declining dynamism.¹

More broadly, modern macroeconomics rests on a foundation of imperfect competition. In endogenous growth models, firms charge markups to cover innovation costs (Romer, 1990). In New Keynesian models, firms are monopolistically competitive price-setters, not perfectly competitive price-takers (Rotemberg and Woodford, 1999). Firm-level markup estimates, applied at scale, allow economists to incorporate rich heterogeneity, strengthening the microfoundations of macro analysis. IO and trade increasingly need the same scalability. IO is

¹References to these puzzles include, respectively, Elsby et al. (2013), Karabarbounis and Neiman (2014); Barkai (2020), Karabarbounis and Neiman (2019); Fernald et al. (2025); Gutiérrez and Philippon (2017), Crouzet and Eberly (2019); and Davis et al. (2007). Syverson (2025) provides a complementary survey covering both markups and markdowns.

moving from individual markets to cross-industry patterns (Döpfer et al., 2025), and trade needs markup estimates across countries and industries (Feenstra and Weinstein, 2017).

However, the empirical evidence for rising markups is far from settled. Some specifications, including De Loecker et al. (2020), imply dramatic increases in recent decades. Others, including Traina (2018), Foster et al. (2026), and Benkard et al. (2025b), do not.

To put the disagreements in context, our review emphasizes that production-based markups are a residual. Using the equilibrium first-order condition for a flexible input, the markup absorbs the entire gap between the output elasticity and the revenue share. Unlike a regression error or structural demand shock, this residual is itself the object of interest. Three things can go wrong: wrong first-order condition, wrong output elasticity, or wrong revenue share. All three deserve scrutiny.

The residual framing has parallels to Solow (1957). The Solow residual absorbs all output growth not explained by input growth. Under standard conditions, it is a clean measure of technical change. But a large literature shows that non-technological factors contaminate the Solow residual as a measure of technology.² Similarly, (mis)measurement and (mis)specification can contaminate production-based markups as a measure of market power. Differences across papers—even those using the same data—can therefore yield different residuals. Contamination is the price of scalability. The approach avoids specifying demand or market structure, but what it omits shows up in the markup.

Even if we resolve all the empirical disagreements, the implications of market power depend on its cause. The production approach we introduce in Section 2 is a measurement tool; it is silent on structural drivers. The first-order condition is an equilibrium relationship. It holds regardless of what generates the markup, so demand and market structure need not be specified. But for some questions, we need the structure. For example, are rising markups “bad,” reflecting barriers to entry, or “good,” reflecting returns to innovation? Answering that question requires structure that the production approach was never designed to provide. The IO literature offers that structure for narrow markets. The production approach offers breadth. Both perspectives are needed. We return to these issues at the end of the paper.

Unfortunately, the devil is in the implementation. Implementing the production approach requires substantive choices, each of which layers on auxiliary assumptions. The “garden of forking paths” (Borges, 1941; Gelman and Loken, 2014) in Section 3 documents the

²Non-technological factors in the Solow residual include markups (Hall, 1986, 1988, 1990), factor utilization (Basu et al., 2006; Fernald, 2014), aggregation (Basu and Fernald, 1997; Baqaee and Farhi, 2020), and mismeasurement (Corrado et al., 2009; Byrne et al., 2016; Syverson, 2017).

sensitivity across two such choices. One is which input to treat as flexible, usually labor or materials. Another is how much variation to allow in output elasticities across firms and over time. Several studies use the same data and estimation techniques to show these choices matter. The results are stark. Some specifications imply sharply rising markups; others find flat or even declining markups.

Section 4 asks whether the source of empirical disagreements is conceptual. The first-order condition may omit wedges that contaminate the markup estimate, including adjustment costs, monopsony power, or bargaining frictions that drive a gap between observed input prices and the allocative costs to which firms actually respond. The production function may also allow too little variation in output elasticities, forcing genuine technological heterogeneity into the markup residual instead. That is, assumptions that restrict elasticity variation can misattribute that variation to markups. A related conceptual fork is the choice between gross-output and value-added production functions. As Section 4.3 discusses, the theoretical and empirical case for value-added production functions is weak (Basu and Fernald, 1995; Gandhi et al., 2017; Ruzic, 2026).

Sections 5 and 6 show how data limitations and estimation choices jointly affect markup estimates. Section 5 focuses on the data constraint: missing inputs, misclassification of variable and fixed costs, and limited sample coverage can all distort revenue shares or bias estimated elasticities. Section 6 examines the econometric challenge of recovering output elasticities when inputs respond to unobserved productivity and prices. Estimation in this setting is harder than often recognized because the standard toolkit was largely built for perfect competition as well as better data than we typically have.

Despite these challenges, understanding and measuring market power remains important. Markups shape resource allocation, welfare, and the propagation of shocks—and no other method estimates them at comparable scale. The challenges are not reasons to set aside the production approach. They are reasons to invest in improving it.

Section 7 closes with three calls to arms, all centered on the tension between market power and technology. First, we call for transparency—systematic reporting of how much revenue-share variation is attributed to markups versus output elasticities. Second, we call for testing, such as comparing production-based markups against demand-based estimates from IO, quasi-experimental evidence, and simulations. Third, we call for more work mapping firm-level markup and technological heterogeneity into macroeconomic models. Getting the implementation right matters precisely because the approach is so broadly applicable.

2 The production approach to markup estimation

Production functions and markups are inextricably linked through input demand. Intuitively, firms with market power raise prices and restrict output, and so require fewer inputs. Hall (1986, 1988, 1990) showed how to use this insight to estimate markups assuming little other than cost minimization. The rest of the firm's profit maximization problem is not needed.

This section first derives the cost-minimizing first-order condition. It then reviews growth-rate implementations, turns to the new microdata literature, and closes by previewing the implementation choices that drive empirical disagreements.

2.1 The unifying first-order condition

Consider a production function for firm i in period t :

$$Y_{it} = F(X_{it}^1, \dots, X_{it}^N, A_{it}), \quad (1)$$

where Y_{it} is gross output, $X_{it}^1, \dots, X_{it}^N$ are inputs, and A_{it} denotes productivity.

The firm minimizes the cost of producing a target level of gross output $\bar{Y}_{it}$. An input is fully flexible if it is chosen in the current period and affects only current-period costs. Let $\mathcal{V}$ denote the set of fully flexible inputs. The firm takes input prices W_{it}^j and non-flexible inputs $\{X_{it}^j\}_{j \notin \mathcal{V}}$ as given. The cost-minimization problem is:

$$\begin{aligned} \min_{\{X_{it}^j\}_{j \in \mathcal{V}}} \quad & \sum_{j \in \mathcal{V}} W_{it}^j X_{it}^j \\ \text{s.t.} \quad & \bar{Y}_{it} = F(X_{it}^1, \dots, X_{it}^N, A_{it}). \end{aligned} \quad (2)$$

The cost-minimizing first-order condition (FOC) for optimal input demand is:

$$W_{it}^j = \lambda_{it} \frac{\partial F}{\partial X_{it}^j},$$

where the Lagrange multiplier λ_{it} is the firm's marginal cost: the shadow value of relaxing the production constraint by one unit. The same marginal cost λ_{it} enters the FOC for any input the firm can freely adjust. To interpret this condition, we define three terms:

- $\mu_{it} \equiv \frac{P_{it}}{\lambda_{it}}$ is the firm's markup of output price P_{it} over marginal cost,
- $\gamma_{it}^j \equiv \frac{\partial F}{\partial X_{it}^j} \frac{X_{it}^j}{Y_{it}}$ is the elasticity of output with respect to input X_{it}^j , and

- $s_{it}^j \equiv \frac{W_{it}^j X_{it}^j}{P_{it} Y_{it}}$ is the input's share in revenue.

With these definitions, the FOC implies the key relationship at the heart of the production approach:

$$\gamma_{it}^j = \mu_{it} s_{it}^j. \quad (3)$$

For any fully flexible input satisfying this FOC, market power ($\mu_{it} > 1$) implies that the output elasticity γ_{it}^j exceeds its revenue share s_{it}^j . The markup is a property of the firm, so the same μ_{it} enters the FOC for every input. Hence, any flexible input can, in principle, be used to identify the markup.³

Throughout, we primarily work with a special case of eq. (1) that separates inputs into capital K_{it} , labor L_{it} , and intermediate inputs (materials) M_{it} :

$$Y_{it} = F(K_{it}, L_{it}, M_{it}, A_{it}). \quad (4)$$

For a multi-product firm, Y_{it} is an index over outputs; each input is an index over heterogeneous types, and μ_{it} is the markup for that aggregate rather than for a specific product. A flexible input $X_{it} \in \{K_{it}, L_{it}, M_{it}\}$ satisfies FOC eq. (3). We next trace out how Hall and others implemented this approach using data in growth rates. We then turn to its more recent implementation in microdata, which focuses on a single input and, where necessary, estimates production relationships in levels.

2.2 The production approach in growth rates

In a series of papers, Hall (1986, 1988, 1990) pioneered the production approach by applying FOC eq. (3) to all inputs simultaneously. His motivation was that when output rises in a boom for reasons beyond a firm's own productivity, the gap between output growth and input growth reveals the markup. As Hall (1986, p. 285) writes,

Macroeconomic fluctuations reveal a good deal about market structure. Students of industrial organization have not generally exploited cyclical movements in their research; they have concentrated almost entirely on cross-sectional analysis.

In implementing this idea, Hall followed Solow (1957) and log-differentiated the production function eq. (4). Using FOC eq. (3) to replace output elasticities with revenue shares and

³Section 4.1 discusses the case of inputs that face additional frictions, such as adjustment costs or input-market power. As we describe there, in principle one can still estimate markups from such inputs by modeling and measuring the additional frictions.

the markup yields the “Hall equation”:

$$\Delta y_{it} = \mu_{it}(s_{it}^K \Delta k_{it} + s_{it}^L \Delta l_{it} + s_{it}^M \Delta m_{it}) + \Delta a_{it} \equiv \mu_{it} \Delta x_{it} + \Delta a_{it}, \quad (5)$$

where lowercase letters denote logs, $\Delta x_{it}^j \equiv \Delta \log X_{it}^j$, and Δx_{it} is revenue-share-weighted input growth.⁴ Appropriately estimated, the relationship between input and output growth reveals the markup μ_{it} . Because the Hall equation applies FOC eq. (3) to all inputs, the approach must grapple with inputs such as capital that may not be fully flexible. Basu and Fernald (2001) and Basu et al. (2001) generalize the cost-minimization problem eq. (2) to account for dynamic inputs and survey the literature that estimates variants of the Hall equation with industry data. Estimation typically assumes a time-invariant markup.

Most estimates using industry data find gross-output markups on the order of 1.03 to 1.06 (for example, Basu and Fernald, 1997). Ahmad et al. (2024) update the growth-rate approach with data from 1949 through 2016 and find markups remain modest. These results are at odds with recent firm-level findings. Some differences could arise from using industry rather than firm data. However, De Loecker and Warzynski (2012) use firm-level data and also find relatively modest markups using the Hall growth-rate approach; they find larger ones using their preferred implementation (discussed in Section 2.3 and the remainder of the paper).

The Hall literature provides three lessons for the more-recent literature.

First, one wants robustness to the functional form of the production function eq. (1) or eq. (4). Because eq. (5) is derived by log-differentiating the production function, in principle it provides a local approximation valid for any production function. There is no need to choose a specific functional form. The functional-form robustness is slightly attenuated by the econometric need to have a time-invariant markup parameter (though factor shares can be allowed to change over time, so it is more flexible than just a first-order approximation.)

Second, markups are not the same as pure economic profits. For example, a firm with fixed or overhead costs can charge a markup yet earn no economic profits, as we now formalize. Following Varian (1984, p. 68) and Syverson (2025), the degree of returns to scale (RTS), Γ_{it} , is the ratio of average cost to marginal cost: $\Gamma_{it} = AC_{it}/MC_{it}$. Pure economic profits, Π_{it} , are the difference between revenues ($P_{it}Y_{it}$) and costs ($\text{Cost}_{it} = \sum_j W_{it}^j X_{it}^j$). Let the rate of pure economic profit in revenue be $s_{it}^\Pi \equiv \Pi_{it}/(P_{it}Y_{it})$. Basu and Fernald (1997) derive a

⁴Hall (1990), equation 5.29. Factor-specific technology growth is subsumed into Δa_{it} by normalizing the elasticity of output with respect to technology, $\partial \log F / \partial \log A_{it}$, to one. Section 4.2 discusses non-factor-neutral technology. Classic growth accounting (Solow, 1957; Jorgenson and Griliches, 1967) imposes $\mu_{it} = 1$. If $\mu_{it} > 1$, the Solow residual equals $(\mu_{it} - 1)\Delta x_{it} + \Delta a_{it}$ and varies systematically with input growth.

useful relationship between markups, returns to scale, and the profit rate:

$$s_{it}^{\Pi} = 1 - \frac{\Gamma_{it}}{\mu_{it}} = \frac{\mu_{it} - \Gamma_{it}}{\mu_{it}}. \quad (6)$$

Economic profits arise when the markup, μ_{it} , exceeds returns to scale, Γ_{it} .⁵ For example, fixed or overhead costs raise average but not marginal costs (so $\Gamma_{it} > 1$); a firm with $\mu_{it} = \Gamma_{it}$ prices above marginal cost but just breaks even. Thus, high markups need not imply high profits; and rising profit shares need not imply rising markups.

Third, because of eq. (6), the production-approach literature following Hall (1990) gave equal weight to markups and returns to scale. Macro modeling that assumed imperfect competition also typically assumed increasing returns in the form of fixed costs to keep profit rates from being too large (Rotemberg and Woodford, 1999). The newer literature, by contrast, typically estimates markups while treating returns to scale and output elasticities as nuisance parameters.

All three lessons remain relevant as we interpret the new literature.

Most of the literature following the Hall growth-rate approach uses industry data. A few papers, however, use firm microdata, such as Klette (1999) and Dobbelaere and Mairesse (2013). Klette (1999) uses a panel of Norwegian establishments, and rather than taking growth rates over time, takes a local approximation in the cross-section in order to flexibly approximate any functional form.

2.3 The new literature on production-based markups

The production approach to markup estimation has expanded rapidly with the increasing availability of firm microdata. Following De Loecker and Warzynski (2012), the new literature uses firm or establishment data, focuses on a single fully flexible input, and (typically) estimates production functions in levels.

De Loecker (2011b) and De Loecker and Warzynski (2012) argue that the growth-rate approach is poorly suited to microdata. First-differencing exacerbates measurement error in short panels (Griliches and Mairesse, 1998); and the aggregate-demand instruments used in the growth-rate literature have little explanatory power in firm cross sections. (Of course,

⁵Returns to scale depend on the scale elasticity of the production function F_{it} (which determines the slope of marginal cost) as well as any fixed or overhead costs of production. Section 4.1 describes when the relationship in eq. (6) survives adjustment costs. Monopsony power requires an extra term, since profits can arise from input-price markdowns as well as from output-market markups.

as Section 6 discusses, not all econometric methods to address endogeneity concerns use instrumental variables.)

De Loecker and Warzynski (2012) also aim to estimate markups firm-by-firm. In contrast, the Hall approach assumes that markups are constant over time; the variants used by Klette (1999) and Dobbelaere and Mairesse (2013) impose that the markups are the same for all firms in an industry.

De Loecker and Warzynski (2012) outline how focusing on a single input—rather than a weighted bundle of inputs as in the Hall approach—can help overcome these shortcomings. The approach has three steps:

- correctly select a flexible input X_{it}^j ,
- measure its revenue share $\hat{s}_{it}^j$, and
- specify a production function and estimate the output elasticity $\hat{\gamma}_{it}^j$.

The estimated markup then follows directly from FOC eq. (3): $\hat{\mu}_{it} = \hat{\gamma}_{it}^j / \hat{s}_{it}^j$.

The first two steps are linked. We need a fully flexible input so that the input’s revenue share can potentially be observed in the data at the allocative market price. For quasi-fixed inputs such as capital—chosen in a prior period or subject to adjustment costs—the revenue share depends on an unobservable shadow price, making it more difficult to recover the allocative cost the FOC requires.

For the third step, estimating the production function generally requires including all inputs to avoid omitted-variable bias, even though only one input’s elasticity is needed for the markup. Consider a Cobb-Douglas production function in capital, labor, and materials:

$$y_{it} = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it} + a_{it}. \quad (7)$$

For example, omitting capital—even if it is not the input used to measure the markup—leads to omitted-variable bias. Firms that use more capital also tend to use more labor and materials, so the omission will load on their coefficients.

An important question in the third step is how flexibly to model the production technology. The choice of functional form determines how many degrees of freedom output elasticities have to explain revenue shares. The Cobb-Douglas function eq. (7) is the simplest choice, but it restricts output elasticities to be common across firms and time. We can relax this by allowing elasticities to vary across time or industries, or by moving to a more flexible functional form altogether. A translog production function adds squared and cross terms

in the inputs, giving it more flexibility. As a second-order approximation to any arbitrary production technology, it lets elasticities vary with input intensity.⁶

We can also relax assumptions on productivity. One such generalization allows for input-augmenting technical change, where productivity has both a Hicks-neutral component A_{it}^H and, say, a labor-augmenting component A_{it}^L . The production function then becomes $Y_{it} = A_{it}^H F(K_{it}, A_{it}^L L_{it}, M_{it})$, which in constant elasticity of substitution (CES) form is:

$$Y_{it} = A_{it}^H \left[(1 - \alpha_l - \alpha_m) K_{it}^{\frac{\sigma-1}{\sigma}} + \alpha_l (A_{it}^L L_{it})^{\frac{\sigma-1}{\sigma}} + \alpha_m M_{it}^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}. \quad (8)$$

With this technology, output elasticities depend on both input intensity and the input-augmenting productivity. We return to this choice and its consequences in the next sections.

The production approach is broadly applicable precisely because it rests on cost minimization rather than a specific model of demand or competition. Demand-side approaches, such as Berry et al. (1995), require detailed product-level data, an explicit model of consumer demand, and assumptions about strategic interaction and firm conduct. Such approaches can be highly informative within a narrowly defined industry but require extensive customization for each setting. Cost minimization, by contrast, is a more general principle. Regardless of how complex or dynamic the profit-maximization problem may be, firms typically aim to produce at the lowest possible cost. This breadth makes the production approach attractive for cross-industry and aggregate analysis.

The breadth does come at a cost. The production-based markup is a residual. It takes whatever value makes FOC eq. (3) hold for the observed revenue share and estimated elasticity. Any misspecification—of the FOC itself, of the production function, or of the econometric method used to estimate it—gets absorbed into the estimated markup. So does any mismeasurement of the revenue share. The markup is, in this sense, the residual claimant of the production approach. This is both the approach’s strength—it requires minimal assumptions—and its central vulnerability, which the rest of this review unpacks.

2.4 Key conceptual choices in implementation

In principle, implementing the production approach laid out in Section 2.3 is simple. Find a flexible input, measure its revenue share, estimate its output elasticity, and compute the ratio. In practice, each step involves choices that can markedly affect estimated markups.

⁶With two inputs, the baseline translog with Hicks-neutral technology is $y_{it} = \alpha_0 + \alpha_K k_{it} + \alpha_L l_{it} + \frac{1}{2}\beta_{KK}(k_{it})^2 + \beta_{KL}k_{it}l_{it} + \frac{1}{2}\beta_{LL}(l_{it})^2 + a_{it}$. The output elasticity for labor is then $\gamma_{it}^L = \alpha_L + \beta_{KL}k_{it} + \beta_{LL}l_{it}$, which varies across firms with their input intensity.

We highlight three key conceptual choices below.

Which input to use. FOC (3) implies any flexible input should yield the same markup. In practice, markups estimated from different inputs often diverge sharply—in both levels and trends. Section 3 documents these divergences using direct comparisons across inputs within the same datasets. Section 4.1 explains why they might arise.

How flexibly to model output elasticities. Having chosen an input, the researcher must specify a production function and estimate its output elasticity. A Cobb-Douglas function restricts elasticities to be common across firms; a translog or CES function allows them to vary with input intensity. When the assumed functional form is too restrictive, genuine variation in technology gets forced into the markup residual—making measured markups rise or fall even when true market power is unchanged. In contrast, an excessively flexible specification, estimated on limited data, could let estimated output elasticities absorb variation that truly reflects market power. Section 3 documents how sensitive markup trends are to this choice. Section 4.2 discusses the resulting bias and argues that in current practice the operative concern more often runs the other way: parametric restrictions leave output elasticities too little room.

Value added versus gross output. Macroeconomists often want a “value-added markup” because GDP aggregates value added across producers. But value-added and gross-output markups are conceptually distinct and cannot be used interchangeably in calibrating macro models. A separate issue is estimation. Basu and Fernald (1995) show that markups invalidate value added as a production measure except under the Leontief assumption that intermediates are used in fixed proportions to gross output—an assumption at odds with the empirical evidence. Section 4.3 takes up both issues.

Despite these choices—and the sensitivity to them documented throughout this review—the production approach remains the most broadly applicable tool available for measuring firm-level markups across industries and countries. That is precisely why getting the choices right, and being transparent about them, matters so much.

3 The garden of forking paths

Existing empirical evidence suggests that the main implementation choices have sharp implications for the measurement of markups. What seems like a clean, mechanical implementation quickly gives way to a garden of forking paths.

This section highlights two important forks in that garden: (1) which input to use, and

(2) how flexibly to specify technology. These are not the only forks. But each has been studied empirically within a fixed dataset, and each materially affects the estimated markups. Later sections explore additional forks, including the specification of the first-order condition (Section 4), data construction (Section 5), and econometric method (Section 6).

We present evidence from studies that compare choices within the same dataset. We defer most interpretation to Section 4. A recurring finding is uncertainty about markup levels and trends. Some specifications imply rising markups, but many plausible ones do not.

3.1 First fork: Which input to use?

Table 1a presents results from three papers that compare markup estimates across inputs within a given dataset. Any fully flexible input that satisfies the conditions in the cost-minimization problem (2) should yield the same markup. In practice, markup estimates using different inputs often diverge in both levels and trends, and the divergences are too large to dismiss as sampling error.

The input choice can reverse the sign of markup trends. Raval (2023) uses manufacturing census data from five countries to show labor- and materials-based markup estimates diverge, implying at least one set of identifying assumptions is wrong. In the United States, estimated labor-based markups nearly double between 1970 and 2010, while estimated materials-based markups fall by half. In Chile and Colombia, labor-based markups fall while materials-based markups rise. Within all five countries, labor- and materials-based markup estimates are negatively correlated across firms, in both levels and trends. These patterns persist across specifications and estimation methods.

The input choice also shapes cross-sectional comparisons. Doraszelski and Jaumandreu (2019) compare exporters and non-exporters in Spain and find the conclusion about which group charges higher markups flips depending on the input used. Exporters appear to charge higher markups when labor is used, but lower markups when materials are used instead.

Even within a single broad input category, the choice of subset matters. The second Raval (2023) block in Table 1a compares energy and non-energy materials. In Colombia, estimated markups based on energy fall 70 percent while markups based on non-energy materials rise 20 percent. In India and Indonesia, markup trends switch signs entirely. These patterns matter because input flexibility is often assumed rather than demonstrated. Implementations of the production approach frequently use materials on the grounds that they are likely more flexible than labor; yet even within materials, markup estimates diverge.

Table 1a: Markup Estimates Can Reverse with the Input Used

First Fork: Which input to use for the production approach?		
Production-Based Markups using Different Inputs		
Raval (2023)	Labor	Materials
U.S. (1970–2010)	Increase of 90%	Decrease of 50%
Chile (1979–1996)	Decrease of 20%	Increase of 15%
Colombia (1978–1991)	Decrease of 30%	Increase of 10%
India (1998–2014)	Decrease of 40%	No change
Indonesia (1991–2000)	Decrease of 10%	Increase of 5%
Doraszelski and Jaumandreu (2019): Export-status comparison, not trend	Labor	Materials
Spain (1990–2012)	Export-status effect: positive	Export-status effect: negative
Raval (2023)	Energy	Non-Energy (Raw) Materials
Chile (1979–1996)	Increase of 20%	Increase of 15%
Colombia (1978–1991)	Decrease of 70%	Increase of 20%
India (1998–2014)	Decrease of 25%	Increase of 5%
Indonesia (1991–2000)	Increase of 40%	Decrease of 5%
Traina (2018)	Cost of Goods Sold	Operating Expenses
Compustat, 1950–2016	Increase: 1.19 to 1.45	Roughly flat: 1.15 to 1.17

Sources and notes: Figures 2 and 5 from Raval (2023), Table 1 from Doraszelski and Jaumandreu (2019), and Figure 2 from Traina (2018). The Raval (2023) entries report approximate percentage changes in average markups, rounded to the nearest 5 percent. The Doraszelski and Jaumandreu (2019) row reports the sign of the Table 1 export-status coefficient, not a time trend. The Traina (2018) entries report approximate markup levels.

Similar sensitivity appears in financial data. Compustat, a dataset of U.S. public firms, is widely used for its broad sectoral coverage. Studies using financial statement data must choose how to map accounting costs to flexible inputs. One option is Cost of Goods Sold (COGS), which collects costs that accountants can trace to units of production, such as materials and direct labor. Another is Operating Expenses (OPEX), which adds Selling, General, and Administrative expenses (SG&A), a residual category covering marketing, management, and other business services. Which of these accounting categories should be interpreted as representing a flexible input is the subject of some debate.

Traina (2018) finds the choice of accounting cost measure is decisive. Using COGS alone implies markups rose from 1.19 to 1.45 between 1950 and 2016, a 26 percentage point increase. Using OPEX, which includes SG&A, implies markups rose from 1.15 to 1.17, an increase of just 2 percentage points. Including SG&A wipes out most of the rise. This sensitivity has both a measurement and an econometric interpretation. From a measurement perspective, SG&A may simply be a variable cost that belongs in the revenue share. From an econometric perspective, even if SG&A is quasi-fixed, omitting it from the production function biases the estimated elasticity of the included input upward, just as omitting capital does. We develop these interpretations in Section 5.

Benkard et al. (2025b) document a related concern: De Loecker et al. (2020)’s replication code drops observations with missing capital or SG&A data—27 percent of the sample—even though SG&A is not used in their markup calculation. Firms with missing SG&A systematically have higher COGS ratios and lower measured markups, creating selection that biases the trend upward. Both findings implicate SG&A in the measured markup increase, though through different mechanisms: input measurement in Traina (2018) and sample construction in Benkard et al. (2025b). The interpretation remains contested. De Loecker et al. (2025) reply and defend their original sample selection. Benkard et al. (2025a) respond that the restrictions remain consequential.

The choice of input shapes substantive conclusions about markups. The logic of the production approach implies that any flexible input should yield the same markup. But different inputs—whether labor, materials, energy, or accounting aggregates—often imply divergent and even contradictory trends.

3.2 Second fork: How flexibly to model output elasticities?

Having chosen an input that is arguably flexible, the researcher must decide how to estimate its output elasticity. The choice involves a tradeoff. Restrictive functional forms, such as

Cobb-Douglas with industry-level coefficients, leave little scope for elasticities to vary across firms. Hence, genuine technological heterogeneity gets forced into the markup residual, leading estimated markups to rise or fall even when true market power is unchanged. In contrast, highly flexible forms—firm-and-time-specific elasticities, or richly parameterized translog functions estimated on limited data—might instead capture variation in estimated output elasticities that truly reflects market power. Where on this spectrum a researcher lands shapes how much revenue-share variation is attributed to technology versus to markups.

Table 1b documents the consequences of different elasticity estimation strategies in three studies that apply multiple approaches to the same data.

Table 1b: Markup Trends Can Reverse with the Elasticity Specification

Second Fork: How to estimate the output elasticity for a given input?			
De Loecker et al. (2020)			
Compustat, 1980–2016			
	Cost Share	Cobb-Douglas	Overhead (PF2)
	Rises: 1.25–1.75	Rises: 1.21–1.61	Rises: 0.98–1.32
Foster et al. (2026)			
U.S. Manuf., 1977–2012			
	Cost Share	Cobb-Douglas	Translog
2-digit industry	Large increase	Increase: +8%	Increase: +23%
4-digit industry	Moderate increase	Decrease: −4%	Decrease: −12%
Demirer (2025)			
		Cobb-Douglas	Factor-augmenting
U.S. (1961–2018)		Rises: about 26 pp	Rises: 1.26–1.38
Chile (1979–1996)		Flat: 1.35–1.37	Falls: 1.28–1.25
Colombia (1978–1991)		Flat: 1.41–1.40	Falls: 1.30–1.26
India (1998–2014)		Flat: 1.32–1.34	Rises: 1.18–1.28
Turkey (1983–2000)		Rises: 1.25–1.33	Falls: 1.23–1.12

Sources and notes: Figures I, XII.A, and XIII.A from De Loecker et al. (2020); Figures 1, 2, and 3 and Table 2 from Foster et al. (2026); and Figures 6 and OA-5 from Demirer (2025). The De Loecker et al. (2020) Cobb-Douglas entry is text-stated; their cost-share and overhead entries are approximate read-offs from the cited figure panels. Most Demirer (2025) country entries report approximate markup levels at the start and end of each sample period; the U.S. Cobb-Douglas entry reports the text-stated increase. The Foster et al. (2026) Cobb-Douglas and translog entries report text-stated changes over 1977–2012; their cost-share entries are qualitative read-offs from Figure 3, which reports 1980–2014 log differences. De Loecker et al. (2020) use cost of goods sold as the flexible input. Foster et al. (2026) use materials. Demirer (2025) uses the combination of labor and materials; the factor-augmenting specification incorporates labor-augmenting productivity in a CES production function as in eq. (8).

One way to forgo more elaborate estimation and specification of technology is to proxy output elasticities with cost shares, that is, the payments to each factor as a share of total costs. As Section 4.2 discusses, with some additional assumptions including constant returns to scale, these cost shares equal output elasticities. No additional estimation is required. De Loecker et al. (2020) implement a version of this by fixing the output elasticity at its median cost share across firms, applying the same value to every firm in a given industry and period. They find this yields markup trends broadly similar to their Cobb-Douglas benchmark—markups rising from around 1.25 to 1.75 under cost shares, compared to 1.21 to 1.61 under Cobb-Douglas—suggesting the upward trend is not merely an artifact of the estimated elasticity rising over time. Even when they extend the production function to include a fixed input, the upward trend in markups remains.⁷

The similarity between De Loecker et al. (2020)’s cost-share and Cobb-Douglas estimates shows that the estimated output elasticity is not doing much independent work across the two specifications. But it also reflects that using a common cost share gives output elasticities essentially no additional degrees of freedom to explain cross-firm variation in revenue shares relative to Cobb-Douglas. An alternative, which we discuss in Section 4.2, is to estimate a separate output elasticity for each firm and period. This alternative gives elasticities more room to explain the data before the remaining variation is left to the markup term. Cost-share proxies carry their own assumptions, however, and the choice between them and more elaborate elasticity estimates is itself a fork in the garden.

Even within a single functional form, the level of industry aggregation at which elasticities are estimated can reverse markup trends entirely. Foster et al. (2026) use U.S. Census of Manufactures plant data from 1977 to 2012, with materials as the flexible input. Under Cobb-Douglas, markups rise about 8 percent at the 2-digit industry level but fall about 4 percent at the 4-digit level.⁸ The mechanism is simple: coarser industry groupings restrict how much output elasticities can vary across firms.

Moving to a more flexible functional form changes the picture further, and its effect interacts with the aggregation level. Under translog at the 4-digit level, Foster et al. (2026) find markups decline about 12 percent—where Cobb-Douglas at the same level showed a slight

⁷This De Loecker et al. (2020) implementation includes SG&A as a separate input to COGS and identifies markups from the output elasticity of COGS. Traina (2018)’s OPEX specification from Table 1a estimates markups using an output elasticity of a combined OPEX input. These discordant results are at the heart of the challenge of mapping accounting concepts into inputs appropriate for production-function estimation.

⁸Industry-level KLEMS data average over within-industry firm heterogeneity in output elasticities and markups. But much of the production approach’s value is that it recovers firm- or plant-level markups. The Foster et al. (2026) results underscore that 2-digit and 4-digit industry aggregation yield different markup trends from the same plant-level data.

decline and cost shares a moderate rise. At the 2-digit level, translog produces a rise of about 23 percent, far larger than the 8 percent under Cobb-Douglas. The underlying plant-level revenue shares are identical across all specifications. Hence, any divergence reflects how different estimation choices assign variation between output elasticities and markups through FOC (3), not differences in the data.

Assumptions about the nature of productivity also matter. Demirer (2025) finds that incorporating labor-augmenting rather than Hicks-neutral productivity can substantially change estimated markup trends. In U.S. manufacturing, Cobb-Douglas with Hicks-neutral technology implies markups rose by about 26 percentage points between 1961 and 2018; the factor-augmenting specification implies a more modest increase from 1.26 to 1.38. In Turkey, the trend reverses sign outright, rising under Cobb-Douglas but falling under the factor-augmenting specification. The rationale, developed further in Section 4.2, is that labor-augmenting productivity gives output elasticities additional freedom to explain changes in relative factor costs, leaving less variation in revenue shares $\widehat{s}_{it}^j$ to load onto the markup.

In sum, the choice of technology specification shapes substantive conclusions about markups just as the choice of input does. The logic of the production approach implies that output elasticities should absorb genuine technological variation. But restrictive functional forms, industry aggregation, or assumptions about the nature of technological progress can all limit how much variation in revenue shares gets attributed to technology. The variation that estimated elasticities cannot absorb instead gets misattributed to the markup.

4 Conceptual rationales for divergences in the garden

This section develops two potential explanations for the discrepancies above: the central FOC could be missing non-markup frictions or the estimated production technology could be insufficiently flexible (or, less often, too flexible). Both can generate the puzzles from the garden, and both have support in the literature. They are not the only explanations—issues with data (Section 5) and econometrics (Section 6) can also affect markup estimates.

4.1 Do we have the right first-order condition?

The key FOC eq. (3) assumes static cost minimization with input prices taken as given. When these assumptions fail, the effective input price that governs the firm’s decisions differs from the observed market price. The gap introduces an unobserved wedge τ_{it}^j between the output elasticity γ_{it}^j and the revenue share s_{it}^j (Doraszelski and Jaumandreu, 2019):

$$\gamma_{it}^j = \mu_{it} s_{it}^j (1 + \tau_{it}^j). \quad (9)$$

Suppose we ignore the wedge. Even with correct values for γ_{it}^j and s_{it}^j , the production approach recovers $\mu_{it}(1 + \tau_{it}^j)$, not the true markup μ_{it} . Suppose also that different inputs face different wedges. They will then yield different markups, as in Table 1a.

A fundamental implication of eq. (9) is that a single FOC cannot separately identify two unknowns. Observed revenue shares s_{it}^j reflect both the true markup μ_{it} and the wedge τ_{it}^j , but the FOC alone cannot tell us how much revenue-share variation is due to one versus the other. Since heterogeneous markups are themselves a form of distortion, they could in principle be folded into the broader category of wedges. But not all wedges are markups, and conflating them leads to misinterpretation.

Separate identification requires bringing in additional modeling assumptions and typically additional data. The choice of what to assume determines what gets identified. In quantifying misallocation, Restuccia and Rogerson (2008) and Hsieh and Klenow (2009) impose structure on markups—assuming common output elasticities γ_t^j within industries and a common demand elasticity that pins down the markup, so that all firm-specific variation in revenue shares is absorbed by exogenous distortion wedges $(1 + \tau_{it}^j)$.⁹ De Loecker and Warzynski (2012) take the opposite approach. They impose structure on the wedges (by effectively assuming they are common across firms), so all variation in revenue shares is attributed to markups. Both approaches achieve identification, but by assumption rather than by data. A third path is to render either markups or wedges, or both, a function of observables. Choi et al. (2024), for instance, use a demand model to quantify markups and then use the residual variation in revenue shares to identify input-market wedges.

These wedges can arise for three distinct reasons.

The first source of wedges arises when observed input prices are not the allocative prices that govern firms' optimization. Even in a static setting, government-induced frictions—tariffs, taxes, quotas, or size-based regulations—can drive a gap between the price a firm pays and the shadow cost it responds to. A specific form of this wedge is input-market power (Robinson, 1933). Concretely, suppose the firm is a monopsonist in the labor market. It faces upward-sloping labor supply, so it must pay a higher wage to hire more. The firm

⁹Restuccia and Rogerson (2017) survey the misallocation literature; see also Haltiwanger et al. (2018) on measurement. Peters (2020); Edmond et al. (2023); Baqaee and Farhi (2020); Baqaee et al. (2024) discuss heterogeneous markups as a source of misallocation. That said, markup heterogeneity need not imply misallocation. Bornstein and Peter (2025) show that under nonlinear pricing, markup heterogeneity across firms can be an equilibrium feature that does not reflect inefficiency.

then acts as if the wage were $W_{it}^j(1 + \epsilon_{it}^j)$, where ϵ_{it}^j is the elasticity of the input price with respect to quantity hired, or equivalently the inverse input-supply elasticity. The result is a “markdown” of the observed wage relative to the marginal revenue product. Production-based markup estimates will conflate that markdown wedge with markup μ_{it} .

Following Dobbelaere and Mairesse (2013), several papers use the divergence of markup estimates across inputs to quantify input-market power, treating the gap between $\hat{\mu}_{it}^L$ and $\hat{\mu}_{it}^M$ as informative about the markdown rather than as a nuisance.¹⁰ Recent work on labor-market monopsony makes the same point from the labor side. Wage posting, bargaining, efficiency wages, and other wage-setting institutions determine how labor-supply elasticities map into markdowns and, therefore, which FOC applies (Caldwell et al., 2026).

The second source of wedges arises when the cost-minimization problem is inherently dynamic, yet the researcher imposes a static framework. Capital is the classic case. Today’s investment affects future production costs, so the relevant shadow price of capital incorporates adjustment costs and expectations, not just today’s rental rate. As Basu and Fernald (2002) and Doraszelski and Jaumandreu (2019) show, the resulting shadow price can still be written as $W_{it}^j(1 + \tau_{it}^j)$, where τ_{it}^j depends on stocks, flows, and expectations, so the input-demand equation retains the form of eq. (9).¹¹

Adjustment costs are not limited to capital, nor are they inherently inefficient (Asker et al., 2014). Cooper et al. (2024) find labor adjustment costs are important and rising in U.S. manufacturing. Their estimates imply that treating labor as fully flexible generates substantial and rising dispersion in production-based markups, even when true markups do not vary. Even materials and other intermediate inputs may face dynamic frictions through supply-chain relationships and delivery lags (Liu and Tsyvinski, 2024; Dhyne et al., 2022; Acemoglu and Tahbaz-Salehi, 2025).

A related concern is that observed wages may not equal the allocative shadow cost even if the firm minimizes costs and there are no adjustment costs or input-market frictions. Most employment relationships are long-term. Hall (1980) argues that observed wages are installment payments on a firm’s long-term obligation to its workers, not spot prices that reflect the current marginal value of labor. If so, the observed wage may differ substantially

¹⁰Studies of labor-market power have a long history, including Manning (2003). In addition to Dobbelaere and Mairesse (2013), recent examples include Berger et al. (2022); Yeh et al. (2022); Jarosch et al. (2024); Kirov and Traina (2022); Rubens (2023).

¹¹Basu et al. (2001, p. 245) and Doraszelski and Jaumandreu (2019) consider adjustment costs in the context of the production approach to markups. Basu et al. (2001) argue that, if adjustment-cost wedges fluctuate around their steady-state value, longer-run markup trends may be more reliable even if period-by-period estimates are not. See also Berndt and Fuss (1986) and Hulten (1986).

from the shadow cost that governs the firm’s current input decision (Basu and House, 2016; Kudlyak, 2024). Similar issues can arise in firm-to-firm transactions governed by relational contracts, where observed transfer prices may reflect long-term relationship value rather than current marginal cost.¹² In both cases, using observed prices in place of allocative shadow costs introduces a time-varying wedge τ_{it}^j into eq. (9).

The third source of wedges arises when the firm’s optimization problem does not take the form of cost minimization at all. Wage bargaining is the leading example, as in search-and-matching models.¹³ The outcome of the bargaining problem can be interpreted as a wedge τ_{it}^j in eq. (9) (Doraszelski and Jaumandreu, 2019).¹⁴ For example, even if true markups are constant, changes in bargaining power or workers’ outside options can change the implied τ_{it}^j and the estimated markup $\hat{\mu}_{it}$.

In short, frictionless input markets are the exception. When measured markups diverge across inputs, as in Table 1a, the gap is itself a diagnostic. It suggests a wedge τ_{it}^j —friction or misspecification—rather than confirming market power. But a diagnostic cannot, on its own, separate the wedge from the markup. As eq. (9) makes clear, a single FOC cannot identify two unknowns. Disentangling the markup from the wedge requires bringing in either additional data that pins down the allocative input price, or additional structure that models the source of the wedge.

4.2 How (flexibly) should we model production?

A second conceptual hypothesis for the discrepancies in the garden is that the assumed production function is too restrictive. Then variation in revenue shares that truly reflects technological differences across firms instead loads onto the only other object allowed by FOC eq. (3), the markup. Unlike the wedge problem in Section 4.1, the FOC may be correctly specified but the estimated output elasticity differs from the true one.

How restrictive functional forms load variation onto markups. The ratio of FOCs across two inputs shows how a restrictive functional form can generate divergences in measured markups. Suppose labor and materials are both fully flexible and satisfy FOC eq. (3).

¹²Rosen (1985) surveys implicit contracts in labor markets. Macchiavello and Morjaria (2023) survey relational contracts between firms.

¹³Rogerson et al. (2005) survey search-and-matching models of the labor market, which are also inherently dynamic. They discuss different specifications of the bargaining process.

¹⁴Here is a simple bargaining model where the outcome yields eq. (9). The wage is set to split the match surplus; the worker’s outside option is U_t and the worker’s bargaining share is β_t . The match surplus is the difference between the worker’s marginal product and outside option: $W_{it}^j = U_t + \beta_t \left(\frac{P_{it}}{\mu_{it}} \frac{\partial F_{it}}{\partial X_{it}^j} - U_t \right)$. Rearranging yields an expression of the form eq. (9), where $(1 + \tau_{it}^j) \equiv \left[\frac{1}{\beta_t} - \frac{1 - \beta_t}{\beta_t} \cdot \frac{U_t}{W_{it}^j} \right]$.

Let $\hat{\mu}_{it}^j$ denote the markup estimated using input X^j . For each firm i , the ratio of the estimated FOCs for labor to materials exactly satisfies:

$$\frac{s_{it}^L}{s_{it}^M} \cdot \frac{\hat{\mu}_{it}^L}{\hat{\mu}_{it}^M} = \frac{\hat{\gamma}_{it}^L}{\hat{\gamma}_{it}^M}. \quad (10)$$

With Cobb-Douglas estimated at the industry level, the right-hand side is constant across firms, so $\hat{\mu}_{it}^L/\hat{\mu}_{it}^M$ must account for all variation in s_{it}^L/s_{it}^M . If firms truly share the same Cobb-Douglas function—so the true output elasticities γ^L and γ^M are identical across firms—then any cross-firm variation in $\hat{\mu}_{it}^L/\hat{\mu}_{it}^M$ reflects sampling, measurement, or estimation error. If not, the restriction forces technology differences into the markup ratio.

We can permit more variation in output elasticities in two ways: finer aggregation or a more flexible functional form. Each gives elasticities more room to explain revenue-share variation. One response is to estimate elasticities at finer industry levels. Even with Cobb-Douglas, allowing coefficients to vary across four-digit rather than two-digit industries lets elasticities explain more revenue-share variation. More-aggregated estimates risk conflating markup and technology heterogeneity. Table 1b is consistent with this mechanism. At two digits markups rise, whereas at four digits they do not, suggesting that less granular elasticities can force technological variation into measured markups (Foster et al., 2026). In practice, however, going much finer is difficult. Estimating separate production functions at the firm or product level is practically impossible with the sample sizes available in most datasets.

Doraszelski and Jaumandreu (2019), Raval (2023), and Demirer (2025) go further, arguing for flexible functional forms that relax Hicks-neutrality. Labor-augmenting technical progress can explain the markup discrepancies in both Tables 1a and 1b even when inputs are fully flexible and do not face non-markup wedges τ_{it}^j . Suppose the production function is CES with labor-augmenting technical progress, as in eq. (8). Then the ratio in eq. (10) becomes:

$$\frac{s_{it}^L}{s_{it}^M} \cdot \frac{\hat{\mu}_{it}^L}{\hat{\mu}_{it}^M} = \left(\frac{\alpha_l}{\alpha_m} \right) \cdot \left(\frac{L_{it}}{M_{it}} \right)^{\frac{\sigma-1}{\sigma}} \cdot (A_{it}^L)^{\frac{\sigma-1}{\sigma}}.$$

Relative output elasticities ($\gamma_{it}^L/\gamma_{it}^M$) now depend on relative factor intensities (L_{it}/M_{it}) and labor-augmenting productivity A_{it}^L . Even if A_{it}^L is common within a period (so $A_{it}^L = A_t^L$), this extra degree of freedom can explain variation in relative factor costs. Demirer (2025) finds allowing for labor-augmenting productivity reduces estimated markups and flattens their upward trend compared with Cobb-Douglas or Hicks-neutral CES/translog specifications.

A more flexible production function can give $\hat{\gamma}_{it}^L/\hat{\gamma}_{it}^M$ more scope to fit the data. Three

questions guide this choice:

- At what level of aggregation should industries be grouped for estimation?
- How flexible should the production function be?
- Is productivity Hicks-neutral, or input-biased (such as labor-augmenting)?

Each works through the same mechanism, letting output elasticities vary more across firms.

Flexibility, however, is not a free lunch. Pushed too far, flexibility reverses the danger: the output elasticity, not the markup, might absorb variation that it should not. Without sufficient structure—especially in small samples—it may be hard to estimate anything sensible. As we disaggregate, we may end up fitting a flexible functional form to increasingly idiosyncratic noise. This surfaces in two ways. First, the elasticity estimates can be unreliable or even nonsensical: large standard errors, sensitivity to outliers, or negative output elasticities for a sizeable chunk of firms (which then imply negative estimated markups). Foster et al. (2026) discuss some of these challenges. Second and more subtly, even when the estimates remain in a sensible range, estimation error may let the elasticity soak up variation that properly belongs to the markup.

Thus, too little and too much flexibility are both concerns. The first forces genuine technology variation onto the markup, the second loads genuine markup variation onto the output elasticity. The general point is about getting closer to the truth, not merely allocating variation between markups and output elasticities. Whichever side of the FOC eq. (3) carries the parametric structure risks compressing the true variation in that parameter, a choice that production-approach papers rarely make explicit. An industry-level Cobb-Douglas production function, for instance, fixes the output elasticity across firms, so every bit of cross-firm variation in revenue shares is forced onto the markup.

Nothing in this discussion of flexibility is special to the production function. Indeed, the same kind of concern would arise if the parametric structure sat on the demand side. To see this, read the FOC in reverse. Rather than a “production approach to markup estimation,” one could pursue a “markup approach to production estimation.” By estimating markups with the standard IO toolkit, a researcher could then back out the output elasticity as the residual that reconciles the markup with the observed revenue share.

Those markup estimates, however, do not follow from the data alone. They require both a demand system and an assumption about firm conduct, the behavioral premise that mirrors cost minimization in the production approach. Conduct fixes the mapping from demand elasticities to markups (for example, the same demand elasticities imply different markups

under Bertrand-Nash than under Cournot). Meanwhile, the functional form of the demand system governs how much markups are allowed to vary across firms—just as the production function governs how much output elasticities can vary. At one extreme, iso-elastic CES demand with monopolistic competition (Dixit and Stiglitz, 1977) delivers a single common markup, forcing all cross-firm variation in revenue shares onto output elasticities—the exact mirror of the industry-level Cobb-Douglas case. In that limit, one would rightly worry that imposing a common markup compresses genuine variation in market power. On either side of the FOC, the operative question is the same: which parameter does the assumed structure restrict too severely?

In our view, the operative concern points toward too little flexibility, not too much, in estimating output elasticities. The empirical patterns in Section 3.2 suggest this choice has bite: small relaxations of parametric restrictions reallocate substantial variation away from markups and toward elasticities. The R^2 diagnostic we develop in Section 7 provides a way to systematize and examine this concern about flexibility.

Constant returns to scale: A useful restriction with real costs. Constant returns is the most common functional form restriction in the production approach, and it has a clear upside. Under constant returns and assuming we can observe allocative input prices, output elasticities equal cost shares directly. No estimation is needed. Recall from Section 2.2 that the degree of returns to scale Γ_{it} equals the ratio of average to marginal cost, and the markup equals price over marginal cost. Together these imply:

$$\gamma_{it}^j = \Gamma_{it} c_{it}^j,$$

where $c_{it}^j = W_{it}^j X_{it}^j / \text{Cost}_{it}$ is input j 's share of total costs. With constant returns ($\Gamma_{it} = 1$), the output elasticity equals the cost share directly: $\gamma_{it}^j = c_{it}^j$. This is appealing because cost shares vary across firms and over time, giving output elasticities firm-specific variation that a single industry-level Cobb-Douglas coefficient cannot provide. That additional variation means cost shares can potentially absorb more of the cross-firm heterogeneity in revenue shares, leaving less to be loaded onto the markup.

Unfortunately, the cost-share approach replaces one set of difficult-to-test assumptions with others. Rather than assuming a restrictive—yet hard to estimate (Section 6)—production function, the researcher assumes constant returns and that allocative input prices (inclusive of adjustment or other shadow costs) are observed. Neither assumption is easy to check (though Foster et al. (2026) argue that both fail empirically). Just as a misspecified production function loads true technology variation onto markups, incorrect assumptions regarding

RTS and allocative input prices also lead to mismeasurement.

If returns to scale in fact vary across firms or over time—because of fixed costs, scale economies, or technology differences—that variation would be interpreted as markup variation. The same is true of deviations from perfect competition in input markets. If monopsony power varies across firms and over time, using cost shares will load that variation into the markup. Section 4.1 engaged with deviations from perfect competition in input markets. Deviations from constant RTS are equally plausible. Ruzic and Ho (2023) find that a secular decline in returns to scale—from increasing toward constant—helps rationalize long-run trends in factor shares and improves the measurement of misallocation, suggesting that time-varying RTS are empirically relevant. Basu et al. (2006) and Ahmad et al. (2024) find evidence consistent with approximately constant RTS at the three-digit industry level, but there may still be heterogeneity across industries and firms. McAdam et al. (2024) find in a panel of European firms that the typical four-digit industry has close to constant returns but that there is considerable heterogeneity.

4.3 Gross output versus value added

The choice of whether to use a gross-output or value-added production function is conceptually distinct from functional-form flexibility. Value added can be a useful concept. For example, GDP is the sum of value added over all producers. In addition, for macro modeling, it can be helpful to abstract from the complicated flows of intermediate inputs across firms.

We address three questions: the conceptual meaning of a value-added markup; its relationship to the gross-output markup; and whether one should estimate the value-added markup directly rather than converting from gross-output estimates. Our practical takeaway is to estimate markups using gross output wherever possible.

What is a value-added markup? Value-added and gross-output markups are related but conceptually distinct. The nominal accounting definition is uncontroversial. Nominal value added is a firm’s revenue net of its spending on intermediate inputs, $P_{it}Y_{it} - W_{it}^M M_{it}$. A markup, however, is a ratio of a price to a marginal cost, and nominal value added is neither.

To speak of a value-added markup, we need to decompose nominal value added $P_{it}^V V_{it}$ into a real quantity V_{it} and a price P_{it}^V . The real value added index V_{it} is meant to capture what the firm itself adds using capital and labor. Conceptually, it is “shoes lacking leather, made without power” (Domar, 1961, p. 716). No market trades this object directly. Hence, V_{it} is conceptually slippery as a production measure.

Still, assuming we can separate price from quantity, the value-added markup is a familiar object applied to a less familiar index. Suppose the production function is separable between intermediate inputs and value added:

$$Y_{it} = F(V_{it}, M_{it}) = F(G(K_{it}, L_{it}, A_{it}), M_{it}). \quad (11)$$

With separability, we can write the cost-minimization problem in two stages. In Stage 1, the firm minimizes the cost of primary inputs $W_{it}^L L_{it} + W_{it}^K K_{it}$ subject to producing the value-added composite $V_{it} = G(K_{it}, L_{it}, A_{it})$. In Stage 2, the firm minimizes total cost over V and M to produce gross output. The Stage-1 Lagrange multiplier, λ_{it}^V , is the marginal cost of producing a unit of real value added. Definitionally, the value-added markup is

$$\mu_{it}^V \equiv \frac{P_{it}^V}{\lambda_{it}^V}.$$

If labor is a flexible input in Stage 1, its FOC (following the logic of Section 2.1) is

$$\mu_{it}^V = \frac{\gamma_{L,it}^V}{s_{L,it}^V}, \quad (12)$$

where $\gamma_{L,it}^V \equiv G_L L_{it}/V_{it}$ is labor's output elasticity in G and $s_{L,it}^V \equiv W_{it}^L L_{it}/P_{it}^V V_{it}$ is labor's share in nominal value added. The value-added markup is, once again, the ratio of a flexible factor's output elasticity to its revenue share—just applied to the value-added sub-production function rather than to gross output.

How does the value-added markup relate to the gross-output markup? Both arise from the same firm-level cost-minimization problem, so they are related—but the relationship runs through the structure of the outer production function $F(V, M)$. Applying the same FOC logic to the Stage-2 problem for the value-added composite gives

$$\mu_{it}^V = \mu_{it} \cdot \frac{s_{it}^V}{\gamma_{V,it}}. \quad (13)$$

The value-added share of revenue is $s_{it}^V = 1 - s_{it}^M$. The elasticity of gross output with respect to V_{it} is $\gamma_{V,it} \equiv \partial \log F / \partial \log V$. Thus, the value-added markup is the gross-output markup, rescaled. Price is rescaled by s^V , the share of revenue accruing to value added; marginal cost is rescaled by γ_V , how strongly V enters production.¹⁵

Under the standard assumption that $F(V, M)$ is homogeneous of degree one in V and M ,

¹⁵The rescaling also involves Y_{it}/V_{it} in the numerator and denominator, which cancel.

the value-added markup μ_{it}^V is (weakly) larger than its gross-output counterpart μ_{it} .¹⁶ Two observations sign the gap. First, in the numerator, the price rescaling P_{it}^V/P_{it} depends on the importance of materials in revenue, per the accounting identity: $s_{it}^V = 1 - s_{it}^M$. Second, in the denominator, Euler's theorem for homogeneous functions yields $\gamma_{V,it} = 1 - \gamma_{M,it}$. Using the FOC for materials, the denominator becomes $\gamma_{V,it} = 1 - \mu_{it}s_{it}^M$. Hence, the rescaling of marginal cost is driven by the markup as well as the importance of materials in production. Substituting both relationships into (13) yields:

$$\mu_{it}^V = \mu_{it} \cdot \frac{1 - s_{it}^M}{1 - \mu_{it}s_{it}^M}. \quad (14)$$

With imperfect competition, the value-added markup μ_{it}^V rises above the markup for gross output μ_{it} because the value-added share of revenue in the numerator, $1 - s_{it}^M$, exceeds the elasticity of gross output with respect to value added in the denominator, $1 - \mu_{it}s_{it}^M$.¹⁷ For value added to have any meaning, the gross-output markup cannot be too high. As $\mu_{it}s_{it}^M$ goes to one, material inputs alone explain all of gross output, leaving nothing for value added to contribute.

The gap between value-added and gross-output markups can be large. Suppose the revenue share of intermediate inputs is 50 percent. Then a gross-output markup of 1.2 corresponds to a value-added markup of 1.5. More strikingly, the sales-weighted average gross-output markup of 1.6 in De Loecker et al. (2020) would correspond to a value-added markup of 4.

Note that one cannot simply plug such a large average gross-output markup (after converting to value added) into a representative-firm macro model. Basu (2019) points out that the sales-weighted average of $\mu = 1.6$ and the implied $\mu^V = 4$ would require implausibly large pure profits and implausibly negative productivity growth. For profits, most estimates suggest average returns to scale, Γ , are close to constant. Then eq. (6) implies pure economic profits of 38 percent of revenue; the value-added counterpart is double that, or 75 percent of GDP. For productivity growth, these markups imply output elasticities so high that, based on eq. (5), the contribution of input growth would exceed output growth. The value-added conversion simply makes the implausibility more obvious. In Section 7.3, we note there is

¹⁶Sato (1976) shows that an index of real value added is not uniquely identified and interpretable without both separability (as in equation (11)) and $F(V, M)$ being homogeneous of degree one. Note that the sub-function $V = G(K, L, A)$ need not be homogeneous of degree one. For example, suppose $Y = AK^{\gamma^K}L^{\gamma^L}M^{\gamma^M}$. We can write this as $Y = V^{1-\gamma^M}M^{\gamma^M}$ with $V = A^{1/(1-\gamma^M)}K^{\gamma^K/(1-\gamma^M)}L^{\gamma^L/(1-\gamma^M)}$.

¹⁷In representative-firm models à la Rotemberg and Woodford (1995) the same markup comparison appears as a result of aggregation and double marginalization. If every firm in the supply chain charges a markup, value-added markups cumulate along the supply chain and are higher than their gross-output counterparts.

no model-free mapping from micro markup estimates to a representative-firm number.¹⁸

Should one estimate the value-added markup directly? Our answer is a firm no. Estimating value-added markups directly requires very strong (and counterfactual) functional form assumptions. The discussion of the two questions above *assumed* there is a separable production function as in (11). But separability alone is not enough. Estimation using value added is misspecified unless the production function (11) is Leontief between value added and materials.

Basu and Fernald (1995) show that, under imperfect competition, empirical measures of value added are generally contaminated by the gross-output markup μ_{it} . To see the intuition, consider a standard empirical measure of value added that takes the difference between gross output and intermediate inputs, both valued at base-year prices. Normalizing the base-year output price, $P_{i,0}$, to one, this double-deflated value added is $V_{it}^{DD} = Y_{it} - W_{i,0}^M M_{it}$ (where $W_{i,0}^M$ is the relative price of intermediates in the base year). The fundamental problem is that we are subtracting intermediates valued with their price, not their marginal product. But with imperfect competition, $\mu_{it} > 1$, the value of the marginal product of an input exceeds its price. So some of the productive contribution of intermediates remains in value added—and the missing contribution itself depends on the gross-output markup.¹⁹

The one escape hatch is if the production function (11) is not only separable, but is Leontief between V_{it} and M_{it} . Since materials are used in fixed proportion to output, there is no substitution margin. Measured real value added tracks the true value-added function $V_{it} = G(K_{it}, L_{it}, A_{it})$ because the firm uses a fixed amount of materials per unit of output regardless of prices or markups.

Suppose the gross-output production function is not separable, let alone Leontief. What happens when we nevertheless estimate a “production function” for V_{it}^{DD} as a function of K_{it} and L_{it} and apply the direct value-added markup formula (12)? The problem is that

¹⁸De Loecker et al. (2020) do not argue for taking their sales-weighted average as a representative-firm number. They emphasize that the entire distribution is important. For example, they show that, given Jensen’s inequality, the economy-wide profit rate is much lower, and more plausible, than the one obtained by plugging the sales-weighted average markup and scale elasticity into eq. (6).

¹⁹If input prices drift far from their base-year values, V^{DD} can deviate substantially from the “true” $V = G(K, L, A)$, since firms respond to current prices, not base-year prices. V^{DD} can even be negative while nominal value added is positive and V is well behaved. Chained value added in modern national accounts avoids these problems by continuously updating the base year. Basu and Fernald (1995) compare alternative empirical measures of value added. Basu and Fernald (1997, 2002) point out that the national accounts identity says that aggregate real GDP equals the aggregate of *measured* real industry value added. With base-year prices, $\sum_i V_i^{DD} = GDP$, the sum of real final expenditure, regardless of whether V_{it}^{DD} equals the “productive” V_{it} from (11). The reason is that the “extra” productive contribution of intermediates is part of output that can be used in the economy.

the production relationship is misspecified. The share of labor payments in nominal value added, $s_{L,it}^V$, is still defined as a nominal accounting measure. But empirical estimates of the output elasticity $\gamma_{V,it}^L$ are not structural objects because real V_{it} does not, in fact, exist. Even if $\mu_{it} = \mu_{it}^V = 1$, the non-structural estimated $\hat{\gamma}_{V,it}^L$ need not equal the nominal revenue share. And if different firms face different intermediate input prices or choose different materials intensities, we will find spurious variation in estimated markups—even when no firm has market power. The problems are even worse, of course, if $\mu_{it} > 1$.

The available evidence is not supportive of the Leontief escape hatch. Jorgenson et al. (1987) and Ruzic (2026) reject separability, let alone the Leontief restrictions.

Our recommendation is therefore to estimate the markup using a gross-output specification and, when a macro-compatible object is needed, to obtain the value-added markup by converting via equation (14) rather than estimating it directly. (This requires care since, as noted above, a sales-weighted *average* gross-output markup cannot be run naively through the conversion and treated as a representative-firm parameter.) The data requirements are the same, since calculating double-deflated value added requires data on real and nominal gross output and intermediates. But direct value-added estimates make sense only with Leontief production, which is at odds with the evidence.

4.4 Takeaways from the garden of forking paths

The divergences in the garden point to important conceptual choices in implementing the production approach to markups. How should researchers respond?

First, the choice of input should be grounded in a persuasive argument that the central FOC holds. This requires arguing from institutional detail rather than default conventions. That might mean focusing on a subset of materials—such as energy inputs in industries where they are purchased on competitive spot markets; or on categories of labor, such as temporary-contract workers, whose wages are likely to reflect contemporaneous market conditions, as in van Heuvelen et al. (2021). Making the case for flexibility means convincing the reader that the observed input price is the allocative one to which the firm responds.

Second, when plausible frictions—such as input market power, adjustment costs, or regulatory constraints—threaten the link between the first-order condition and markups, we may need to add more structure to the estimation by modeling the source of the wedge.²⁰ We can use two inputs to estimate two frictions (such as markups and markdowns); or incorpo-

²⁰For example, Choi et al. (2024) model product-market power, input-market power, and input-specific distortions inside the same FOC, identifying two with model structure and inferring one as a residual.

rate dynamic optimization to account for capital or labor adjustment costs; or use auxiliary data to calibrate or instrument for these frictions.²¹ Such structure can help avoid loading non-markup distortions into the measured markup.

Third, researchers can choose a production technology that is sufficiently flexible to explain variation in revenue shares without forcing it into the markup residual. This might involve estimating elasticities at a more granular industry level, moving beyond Cobb-Douglas to CES or translog specifications, or allowing for input-augmenting technical change. These choices expand the scope for output elasticities to capture genuine technological heterogeneity. The tradeoff is that too much flexibility strains estimation. Flexible specifications might be overparameterized and fit noise in small samples.

Ultimately, the production approach’s central vulnerability—that the markup absorbs all misspecification—also guides best practice. If estimates across inputs or specifications agree, that convergence is consistent with small wedges and a well-specified production function. If they disagree, as they often do in Tables 1a and 1b, that divergence is informative. It signals frictions or misspecification that should be investigated, not papered over. A complementary approach is to focus on longer-run averages and trends. If the frictions that contaminate markup estimates are roughly stationary over time, longer-run trends may still be informative about market power even when cross-sectional comparisons are unreliable.

The conceptual challenges discussed in this section compound if the underlying data are inadequate. Even a correctly specified model can then produce misleading markup estimates, as Section 5 documents.

5 The data constraint

The production approach needs a revenue share s_{it}^j and an output elasticity γ_{it}^j . Data shortcomings can affect both. Mismeasured or misclassified costs, such as the COGS/SG&A boundary discussed below, distort the revenue share directly. At the same time, missing or noisy inputs contaminate the production-function regressions used to estimate output elasticities. These difficulties might contribute to why markup estimates diverge across studies.

This section focuses on the data issues; Section 6 takes up some of the econometric challenges they create. Microdata have long been known to fall short of theoretical ideals (Grunfeld and Griliches, 1960; Griliches and Ringstad, 1971). We first discuss how measurement error biases

²¹Examples of each approach include Dobbelaere and Mairesse (2013); Doraszelski and Jaumandreu (2019); Kirov and Traina (2022); Cooper et al. (2024).

markups. We then discuss specific data shortcomings, assess sample representativeness, and offer research recommendations based on historical parallels.

5.1 Why measurement error matters

The production markup $\hat{\mu}_{it}$ is the ratio of an estimated output elasticity $\hat{\gamma}_{it}^j$ to a measured revenue share s_{it}^j . This ratio structure means that measurement errors can distort markups in ways that are predictable in form but difficult to sign.

First, micro datasets often miss economically important inputs. For example, proprietary software, informal labor arrangements, and other intangibles often do not appear in production surveys. Similarly, data on the skill composition of the labor force or the extent of factor utilization are typically missing. To see why missing inputs matter, suppose we know a priori that returns to scale are constant. As discussed in Section 2.2, the markup then equals revenue divided by total cost: $\hat{\mu}_{it} = P_{it}Y_{it}/\text{Cost}_{it}$. If we miss inputs, we understate economic costs and overstate the markup. The bias approximately equals the input’s cost share times the measurement error.²²

If we instead estimate output elasticities econometrically, then missing inputs lead to omitted variable bias. Suppose materials are the fully flexible input and we estimate a Cobb-Douglas production function. If we omit a portion of, say, intangible capital, then all regression coefficients are potentially biased. The bias in the estimated output elasticity of materials, $\hat{\gamma}^M$, depends on the partial correlation between omitted capital and materials, conditional on included variables. Because this is a conditional correlation, the bias can be positive or negative even if the unconditional correlation is positive—so the direction is not obvious without knowing the data.

Of course, output itself may also be mismeasured. We defer output mismeasurement to Section 6, where unobserved firm-level output prices enter as the second layer of endogeneity.

Second, even when inputs are not systematically missing, they may be measured with noise. Classical measurement error in a single input attenuates its own estimated elasticity. However, the effect on the coefficients for other (accurately measured) inputs is, in general, ambiguous, depending on conditional correlations in the data. In practice, multiple inputs might be mismeasured. The net effect on output elasticities and markups depends on which inputs carry more noise and how correlated different inputs are; it cannot generally be signed without dataset-specific knowledge. For instance, suppose materials are mismeasured. On its own, that would lead to attenuation bias in $\hat{\gamma}^M$. But measurement error in labor could

²²Measuring input j as $\hat{X}^j = (1+\epsilon)X^j$ yields markup $\hat{\mu} = \mu/(1+\epsilon c^j)$ with an error of about $-\epsilon c^j$ percent.

partially or totally offset the attenuation. When noise obscures labor’s true variation, ordinary least squares (OLS) assigns some of labor’s productive contribution to materials, which tends to push $\hat{\gamma}^M$ back up.

Capital is especially likely to be measured with error. Book values, uncertain depreciation schedules, and missing intangibles all add error. Collard-Wexler and De Loecker (2016) suggest that Monte Carlo simulations can help assess the quantitative (and directional) bias that arises. In their simulations, when the variance of capital’s measurement error is 40 percent of capital’s conditional variance, capital’s coefficients can be biased downward by a factor of two. But, for estimating the output elasticity for materials, the sizeable mismeasurement of capital has a more muted effect than does measurement error in labor. Capital adjusts slowly—firms cannot instantly buy or sell machines in response to current conditions; labor responds more quickly. Capital and materials therefore move together less than do labor and materials. Hence, mismeasured capital is less able to push $\hat{\gamma}^M$ upward. In their simulations, the net bias in estimated markups from capital mismeasurement is smaller than one might fear. But it is still present and its direction depends on data specifics.

Nonetheless, the practical stakes are high. Even in this simplified setup, plausible levels of measurement error can shift markup estimates by 5–15 percent (Collard-Wexler and De Loecker, 2016). The direction depends on conditional correlations among inputs, not just on which input is mismeasured. If the variance of the measurement error is relatively constant, the bias mainly affects the level of markups. If the variance of the measurement error is changing over time, it could also influence estimated trends.

5.2 Specific measurement shortcomings in standard datasets

The previous subsection focused on the general mechanics of how missing inputs or classical measurement error could affect results. This subsection discusses five common shortcomings, roughly ordered by their practical difficulty.

Revenue vs. quantity data. Most micro datasets record revenues, not physical quantities and prices separately. In U.S. microdata, both Compustat and individual Census surveys offer predominantly revenue-based output measures. Separating price from quantity is possible only for a narrow slice of the economy—Foster et al. (2008) study just 11 manufacturing products, including coffee, ready-mixed concrete, and motor gasoline. Such data are generally unavailable at the scale needed to study broad sectors or the whole economy. That said, survey data from some other countries—Colombia and India, for instance—sometimes separate output price from quantity. And many customs datasets used in international trade

also contain unit values.

When Foster et al. (2008) do separate prices from quantities, they uncover a striking puzzle: physical productivity is *negatively* correlated with price, whereas revenue productivity is *positively* correlated with price. The resolution reveals why revenue data creates problems for markup estimation. A physically productive firm has lower costs and can profitably charge less—hence the negative price correlation. But revenue productivity mixes efficiency with pricing power. High revenue productivity could reflect genuine efficiency or else high markups.²³

Without observing prices and quantities separately, we risk conflating technical efficiency and market power. Indeed, using Colombian data that distinguish price from quantity, in tandem with external instruments, de Roux et al. (2024) find that the level of measured markups is some 30 percent lower than implied in more standard approaches. We return to this challenge in Section 6.

Missing inputs and input quality create hard-to-detect biases. Microdata rarely contain direct information on input quality, which can bias production function estimates as well as productivity comparisons across firms (Griliches, 1957; Fox and Smeets, 2011; Grieco et al., 2016). Sometimes inputs are not recorded at all. Autor et al. (2020) study labor shares in U.S. Census microdata for six large sectors; only in manufacturing can they construct labor shares of value added, because most sectoral Census microdata lack systematic information on intermediate inputs. Even in manufacturing, inputs of business services are often missing. As discussed in the previous section, these omissions are likely to bias estimated output elasticities, whether estimated via cost shares or regression.

The mapping from COGS and SG&A to a flexible input is unclear. Compustat reports total employees and book value of capital, but lacks the detailed input classifications available in Census surveys—and intermediate inputs are missing entirely.²⁴ Instead, firms classify costs under Generally Accepted Accounting Principles (GAAP) as either Cost of Goods Sold (COGS) or Selling, General, and Administrative (SG&A) expenses. COGS

²³Research from other countries reinforces these concerns. Garcia-Marin and Voigtländer (2019) show that using revenue productivity rather than physical productivity underestimates export efficiency gains in Chilean plant data. Lenzu et al. (2026) combine firm-level output prices and quantities with quasi-experimental variation in credit supply in Belgium and show that revenue-based measures underestimate the long-run elasticity of physical productivity to credit supply by half.

²⁴Compustat’s staff expense variable covers only 5–20 percent of firm-years depending on sample construction, with non-random selection toward larger firms and specific industries. For example, Brynjolfsson and Hitt (2003) use Compustat for production function estimation but impute labor costs for firms with missing staff expense data. The headcount variable is more widely available but is unaudited and misses labor quality, non-employee labor, and outsourced services.

covers costs directly attributed to production: direct materials, direct labor, and manufacturing overhead. SG&A covers everything else: executive salaries, marketing, R&D, and administrative overhead.

De Loecker et al. (2020) argue that this accounting structure offers an advantage over Census surveys. Census data, for instance, lump production and overhead workers together, suggesting that an amalgam of all workers might not be a suitable choice for a flexible input in the production approach to markups. By contrast, they argue that in choosing what expenses to report as COGS, firms themselves report on the relevant flexible inputs used in production. In this view, expenses under SG&A should explicitly be excluded from the flexible-variable categorization in the production approach. In their sample, the COGS share of revenue has declined in recent decades, while their estimated output elasticity for COGS does not change much. They therefore interpret the declining COGS revenue share as a rising markup.

But the mapping between accounting categories and economic inputs is not clear-cut. For instance, a salesperson’s commission sits in SG&A yet scales directly with sales. Because it is a selling expense, it is definitionally excluded from COGS; yet it can reasonably be viewed as part of the flexible input required to produce and sell a unit of output. Indeed, the smooth within-firm evolution of SG&A (Traina, 2018) suggests these costs behave more like routine expenses that vary with output than like the lumpy adjustments characteristic of fixed costs. This distinction is quantitatively important. As discussed in the garden section (3), whether SG&A enters the flexible-input measure shapes the estimated markup trend. With COGS alone, markups rise substantially after 1980, but once SG&A is folded into the flexible input, Traina (2018) finds that most of this increase disappears.

Capital is hard to measure. Micro datasets typically provide book values of capital—historical purchase prices adjusted by accounting depreciation rules—rather than economic values reflecting productive capacity. Book values can differ sharply from true productive input. A fully depreciated but still-productive machine carries zero book value, whereas a recently purchased but obsolete asset still carries a high book value (Becker et al., 2006).

The problem deepens with intangible assets. A growing literature documents the importance of intangible capital—brand equity, organizational capital, and other assets beyond software and R&D (Corrado et al., 2009)—but these typically appear in accounting data as operating expenses rather than capitalized assets.²⁵ Karabarbounis and Neiman (2019) document substantial growth in “factorless income”—the apparent profits remaining after subtracting

²⁵Koh et al. (2020) argue that the measured decline in the U.S. labor share is entirely accounted for by the capitalization of intellectual property products in the national accounts—illustrating how changes in accounting treatment, rather than changes in market power, can shift measured factor shares.

measured payments to labor and capital from GDP. This could reflect true market-power rents, an understated user cost of capital, or returns to unmeasured intangible capital.

Under the cost-share approach, any understatement of true costs translates directly into inflated markup estimates. Econometric corrections can help: Ližal and Galuščák (2012), Collard-Wexler and De Loecker (2016), and Kim et al. (2016) show that correcting for capital mismeasurement can nearly double estimated capital elasticities. The cost-share approach also requires a measure of the user cost of capital. Because most capital is owned rather than rented, the user cost must be imputed from depreciation rates, expected capital gains, risk premia, and tax adjustments (Hall and Jorgenson, 1967; Jorgenson et al., 1987). User-cost estimation is itself a major measurement challenge.

Measuring the cost of inflexible inputs requires prices that are rarely observed.

For inputs that cannot be freely adjusted—capital and, in the short run, certain types of labor—the relevant cost concept is the shadow price, not the market price. When a firm cannot instantly hire or fire workers, or cannot costlessly buy or sell capital, the shadow value of those inputs reflects adjustment costs and expectations, not just today’s market wage or rental rate. Using observed market prices to measure the cost of inflexible inputs therefore introduces systematic error into both the revenue share and the estimated elasticity simultaneously, that is, into the two ingredients whose ratio defines the markup. This problem cuts across all of the measurement challenges above, since capital and quasi-fixed labor are precisely the inputs most likely to be mismeasured in other dimensions as well (Fox and Smeets, 2011; Ornaghi, 2006; Kim and Traina, 2025).

5.3 Are the firms we observe representative?

The production approach can be applied at scale. Still, even perfect measurement at the firm level would not resolve the challenge that the firms we observe most comprehensively might not be representative of the economy we want to understand.

The U.S. Census of Manufactures is the gold standard for production function estimation: broad coverage, detailed input data, and separation of plants from firms (Dunne et al., 1988; Baily et al., 1992; Foster et al., 2008; Syverson, 2011). But manufacturing accounted for only 11 percent of U.S. private-industry value added in 2024 (BEA, 2026). If markups in services or finance differ systematically from those in manufacturing, studies based on Census manufacturing data are learning about a shrinking slice of the economy. In addition, manufacturing firms have relatively clear distinctions between production workers and overhead, and between raw materials and administrative expenses. But boundaries blur for service

firms. A software engineer at a technology firm: production labor or overhead? As the economy shifts toward services, accounting categories designed for manufacturing become less meaningful. Census microdata remain best-in-class in coverage and sample size, and researchers are exploiting within-sector detail—for example, Dhyne et al. (2022) use Belgian firm-to-firm transaction data that capture input linkages and directly record intermediate input prices—but the representativeness problem does not go away.

Given manufacturing’s declining share, researchers increasingly turn to Compustat, which offers broad sectoral coverage and is easy to access. Compustat covers only public firms, which account for roughly half of industry sales (Davis et al., 2007), and some 40 percent of those sales come from foreign operations, complicating the interpretation of domestic market power. Several papers (Ali et al., 2009; Keil, 2017; Decker and Williams, 2023) find that common concentration measures at the 4-digit NAICS level based on Compustat are only weakly correlated with those from the more comprehensive economic census. Sample construction choices also matter: Benkard et al. (2025b) show that unreported sample restrictions in De Loecker et al. (2020) drop 27 percent of available Compustat observations. With the full sample, they find that markup increases are more muted and concentrated in Finance and Insurance, where cost measurement is especially problematic.

Whatever the dataset, a basic diagnostic is to check whether micro aggregates match published national accounts totals. Fernald and Piga (2023) find that factor shares in the NBER-CES manufacturing industry data are wildly different from those in the national accounts. In 1997, for example, labor’s share of value added in manufacturing was only 31 percent, compared to 58 percent in the BEA-BLS KLEMS dataset for the same sector. The main difference is that industry and aggregate accounts seek to fill in known data shortcomings by combining and integrating multiple data sources—tax records, banking surveys, innovation surveys—that researchers rarely access; micro data should not be expected to aggregate to published totals without understanding that integration. Since revenue shares directly map to markups, what looks like a high markup might also be missing labor compensation or business services.

5.4 Historical lessons and recommendations

Measurement challenges are not new and are, of course, pervasive. How economists confronted earlier episodes offers lessons for the production approach to markups.

Both macroeconomics and IO were confronted with residuals that potentially conflated true economic phenomena with measurement artifacts. In macro, the puzzle was aggregate total

factor productivity (TFP). Solow (1957) found that what we now call the Solow residual explained 88 percent of growth in U.S. output per hour from 1909 to 1949. Jorgenson and Griliches (1967) hypothesized that this residual reflected measurement error in output and inputs. They issued a “manifesto” for improving measurement (Berman and Jaffe, 2024, p. 579).²⁶ They applied index-number theory to construct better measures: Divisia output indexes, hedonic price adjustments, theory-based indexes of labor composition, and user-cost weights for heterogeneous capital. More than half a century of research in national accounting circles has refined these measures at an industry and aggregate level. Such data are not just the sum over firm-level datasets, because they combine multiple data sources, including statistically representative surveys. This is a reason why the growth-rate implementation of production-approach markups has mostly used industry data (Section 2.2).

IO faced a parallel crisis. The “structure-conduct-performance” (SCP) paradigm (Bain, 1951; Bresnahan, 1989; Schmalensee, 1989) tried to infer market power from correlations between industry concentration and accounting profit rates. It failed on two fronts. The first is causal ambiguity. Even with perfect data, the correlation between concentration and profits is uninterpretable without additional economic structure (Demsetz, 1973). The second is measurement. Accounting profitability at best proxies a price–average-cost wedge, not the markup of price over marginal cost (Fisher and McGowan, 1983). The same capital measurement problems that bedevil the production approach—risk premia, inflation, depreciation rules, missing intangibles—made accounting profitability an unreliable proxy for true economic returns. The IO response was to abandon broad cross-industry regressions in favor of deep, structurally disciplined studies of specific industries, using institutional knowledge to pin down the assumptions that aggregate approaches cannot test (Berry et al., 2019).

The production approach to markups tries to be broad like the macro tradition and agnostic about the structure of demand like early IO—and in doing so it has to face the measurement challenges of both. Like the SCP paradigm, it is sometimes forced to use accounting costs and book values. Like early aggregate productivity accounting, it must take a stand on input measurement without the cross-checking that comes from integrating multiple data sources. Substantive progress may require both better data—the Jorgenson-Griliches response—and more disciplined structural assumptions—the IO response.

Four practical recommendations follow from this history. First, the Jorgenson-Griliches agenda of improving underlying data remains as relevant as in 1967. Better measurement of

²⁶Measurement progress since the 1960s owes much to Jorgenson and Griliches, whose contributions extend well beyond their joint work. Griliches emphasized microdata, while Jorgenson focused on macrodata and industry aggregates (Berman and Jaffe, 2024; Fernald, 2024). Despite measurement advances, TFP still accounted for roughly half of labor-productivity growth from 1948–2024 (Fernald, 2014).

inputs—especially capital services, intangible investment, firm-level prices, and the boundary between production and overhead labor—would strengthen the foundation for production-based markup estimation (Griliches and Mairesse, 1998; Lev, 2001). Second, because the direction of bias from classical measurement error is ambiguous (Section 5.1), simulation can help assess its importance in a given application. Collard-Wexler and De Loecker (2016), for example, use dataset-specific simulations to bound measurement bias for estimated markups. Third, a simple check is to compare micro aggregates with published national accounts totals—large discrepancies are a warning sign that something is being missed, as Fernald and Piga (2023) illustrate for manufacturing. Fourth, when measurement problems cannot be resolved, be explicit about the resulting biases and whether they can plausibly explain patterns attributed to market power. These issues also feed directly into the econometric challenges of estimating output elasticities, which Section 6 takes up next.

6 Practice and pitfalls in estimating output elasticities

Production-based markups are only as reliable as the output elasticities that underpin them. The measurement challenges documented in Section 5, particularly unobserved output prices, feed directly into this estimation problem. The challenge is that most tools for estimating those elasticities were built for a world of perfect competition, and using them to measure departures from perfect competition creates problems that run deeper than is commonly appreciated. The applied economist has a rich toolbox, but its tools force tradeoffs. Some are restrictive for an IO economist who wants to remain agnostic about the structure of production and competition. Others are overly general for a macroeconomist who wants to calibrate a specific model.

We organize the discussion in four parts. First, we show how moving from physical output to revenue data adds unobserved prices to the regression residual and creates a new source of bias. Second, we show what goes wrong if those extra terms are ignored—including a striking result that naive estimation with revenue data can be entirely uninformative about markups. Third, we evaluate the main econometric tools the literature has developed, asking how each fares once we take the richer residual seriously. Fourth, we offer an assessment of what the toolkit can and cannot deliver.

6.1 The problem: Two layers of endogeneity

The fundamental challenge in estimating output elasticities is endogeneity: firms choose their inputs in response to forces the econometrician cannot observe, so input choices are

correlated with the regression residual. This problem has two layers in realistic settings.

Layer 1: Physical productivity. The cleanest statement of the problem starts from the physical production function. For concreteness, take the Cobb-Douglas production function eq. (7) with three inputs and a firm-specific productivity term a_{it} :

$$y_{it} = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it} + a_{it}, \quad (15)$$

where all variables are in logs and γ^X denotes the output elasticity of input X . The residual here is just productivity a_{it} . The endogeneity problem is the classic “transmission bias” (Marschak and Andrews, 1944). That is, a firm that draws a high productivity shock may want to use more materials, so m_{it} and a_{it} move together. OLS attributes that comovement to technology and overstates $\hat{\gamma}^M$.²⁷ Because the production approach calculates markups as $\hat{\mu}_{it} = \hat{\gamma}^M / s_{it}^M$, overstating $\hat{\gamma}^M$ directly overstates measured markups. This first layer is well understood, and the econometric toolkit was largely built to address it.

Layer 2: Unobserved output prices. Few datasets contain firm-level output quantities. Suppose we observe firm revenue $r_{it} = p_{it} + y_{it}$, where p_{it} is the firm’s (log) output price. Adding p_{it} to both sides of eq. (15) gives the revenue production function that is estimated:

$$r_{it} = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it} + \underbrace{a_{it} + p_{it}}_{\text{residual}}. \quad (16)$$

The residual now contains the firm’s output price as well as its productivity. Output prices are correlated with input choices for a different reason than productivity. Holding productivity and demand fixed, firms with more market power charge higher prices, restrict output, and therefore use fewer inputs. Hence, output prices and input quantities are negatively correlated (Klette and Griliches, 1996), biasing output elasticities *downward*—opposite from Layer 1. The two biases need not cancel. The revenue-based residual $a_{it} + p_{it}$ conflates technology and demand-side forces in a way that standard tools, designed for eq. (15), were not built to handle. De Loecker et al. (2016) confirm this is not a theoretical curiosity. In Indian manufacturing data where firm-level prices are observed, revenue-based coefficients are biased estimates of output elasticities.

An additional complication arises when input quantities are also unobserved. Many datasets record only input expenditures $e_{it}^M = w_{it}^M + m_{it}$, so firm-specific input prices w_{it}^M enter the

²⁷Henceforth, we assume the bias is positive. It need not be. Suppose firms have market power, prices are sticky, and all firms receive the same productivity improvement. Marginal cost falls but firms do not adjust prices. The rising markup dampens input demand, leaving the overall OLS bias ambiguous.

residual as well.²⁸ We focus primarily on the first two layers—productivity and output prices—which have received the most attention in practice.

A unifying look at input demand. To see why all three components cause endogeneity, it helps to look at what determines a firm’s materials choice. Under Cobb-Douglas production, the first-order condition for materials gives their conditional demand:

$$m_{it} = \frac{1}{1 - \gamma^M} (a_{it} + p_{it} + \log \gamma^M - w_{it}^M - \log \mu_{it} + \gamma^K k_{it} + \gamma^L l_{it}). \quad (17)$$

Every component of the composite residual—productivity a_{it} , output price p_{it} , and input price w_{it}^M —drives materials demand. Hence, the flexibility that makes materials a good candidate for the key FOC ensures that materials respond to forces the econometrician commonly cannot observe.

6.2 What goes wrong if you ignore the extra layers

Ignoring the richer residual has serious consequences, even for sophisticated estimation strategies. Even setting aside output and input prices, OLS applied to eq. (15) overstates $\hat{\gamma}^M$ because productive firms use more materials. Since markups are computed as $\hat{\mu}_{it} = \hat{\gamma}^M / s_{it}^M$, the overstatement flows directly into estimated markups. This “transmission bias” is the reason the literature moved away from OLS toward more sophisticated approaches. But moving to revenue data in eq. (16) adds a second source of bias that can partially offset or further amplify the first, in ways that are difficult to sign ex ante.

The most striking consequence of ignoring Layer 2 is a result due to DeSouza (2009) and Bond et al. (2021). Under their maintained assumptions, if output elasticities are estimated from revenue data without correcting for omitted prices, the production-approach markup formula is tautologically uninformative. With revenue data, OLS recovers not the output elasticity γ^M but a revenue elasticity that already embeds the markup, $\hat{\gamma}^{\text{revenue},M} = \gamma^M / \mu$. Plugging this into the markup formula then gives:

$$\hat{\mu}^{\text{revenue}} = \frac{\hat{\gamma}^{\text{revenue},M}}{s^M} = \frac{\gamma^M / \mu}{\gamma^M / \mu} = 1,$$

an estimated markup of one, regardless of the markup’s true value.

One response to missing firm-level prices is to deflate revenues by an industry price index

²⁸The composite residual becomes $\varepsilon_{it} = a_{it} + p_{it} - \gamma^M w_{it}^M$. If firms with market power in input markets face lower prices and purchase more, input expenditures are correlated with the input-price term—a source of endogeneity discussed further in Section 4.1.

p_t . Hashemi et al. (2022) show that this “fix” resolves the problem only under perfect competition—where markups equal one and the exercise is unnecessary. To see the logic, the price deviation in the residual becomes $\tilde{p}_{it} = p_{it} - p_t$. Under perfect competition, the deflation works because firms charge the same quality-adjusted price, so $\tilde{p}_{it}$ is zero and, thus, uncorrelated with inputs. But under imperfect competition—the case of interest—firm-level prices reflect market power and are correlated with input choices. Klette and Griliches (1996) show this biases elasticity estimates downward, and De Loecker et al. (2016) show the bias is empirically meaningful. Revenue productivity may still be a useful object in some settings (Foster et al., 2008), but it is not physical productivity.

Both layers of the residual create a fundamental circularity: estimating output elasticities requires controlling for markups (because markups affect input demand through eq. (17)), but markups are what we are trying to measure using those elasticities. This means that any estimation strategy that ignores markups is potentially misspecified. The circularity also explains why firm-level price data are so valuable. These data allow researchers to separate the productivity and price components of the residual directly, breaking the loop. Such data are increasingly available in manufacturing and trade datasets for homogeneous goods, though services and differentiated products remain challenging.

6.3 The toolkit: What each approach does and does not handle

The literature has developed a range of tools to address endogeneity. Each was designed primarily for Layer 1—a world where the residual is just productivity.

Fixed effects remove time-invariant components of the residual, but leave the time-varying components that matter most for markup estimation. Introduced by Mundlak (1961) for agricultural production, where productivity largely reflects time-invariant characteristics like soil fertility, firm fixed effects absorb the permanent part of productivity, output prices, and input prices. But time-varying productivity shocks, demand fluctuations, and input-price variation all remain in the residual and remain correlated with input choices.

Unfortunately, fixed effects exacerbate measurement error in capital. Because within-firm variation in capital is dominated by investment and depreciation, both of which are measured with error, demeaning amplifies that noise. Griliches and Mairesse (1998) document that panel methods applied to micro data have produced disappointing results, with unreasonably low capital coefficients and implausible returns to scale. Becker et al. (2006) show in an in-depth study of capital measurement that different ways of measuring capital which ought to be equivalent—such as perpetual inventory methods or inferring capital from capital-

producing sectors—lead to meaningfully different production function estimates. Measurement concerns were the primary reason fixed effects were largely abandoned for production function estimation (De Loecker and Warzynski, 2012).

Instrumental variables are the most conceptually transparent solution, but the bar for a valid instrument is very high in the richer-residual world. A valid instrument z_{it} must be correlated with the input choice but uncorrelated with *all* components of the composite residual—productivity and output prices at minimum, and input prices when expenditures proxy for quantities. The input demand eq. (17) clarifies what could work in principle.

Any variable that shifts input demand without entering the production function could be a valid instrument. Firm-level input prices w_{it}^M are a natural candidate when available (Doraszelski and Jaumandreu, 2013). That said, if we observe input expenditures, not input quantities, the same price enters the residual mechanically. In that case, input-price variation cannot adequately instrument expenditure unless we can recover quantities or impose additional structure. Firm-level input prices are also often correlated with local demand conditions that affect output prices, undermining validity (Gandhi et al., 2020).²⁹ Aggregate demand instruments—government spending, oil price shocks, monetary policy—have been used in industry time-series approaches (Section 2.2). But they often have little explanatory power at a firm level, and they raise validity concerns of their own. Demand shocks that move output prices or induce entry and exit change the prices and markups embedded in the revenue-based residual. Despite these limitations, instrument relevance is testable, and IV deserves serious consideration when strong instruments are available.

Control functions are the current workhorse of the production function literature, largely because they avoid the need for external instruments. However, their key assumptions are individually questionable under imperfect competition and must hold simultaneously. The approach, developed by Olley and Pakes (1996) and extended by Levinsohn and Petrin (2003), uses observable firm decisions as proxy variables for unobserved productivity. Olley and Pakes (1996) originally used investment as the proxy. But investment is often lumpy and infrequent. Hence, Levinsohn and Petrin (2003) proposed intermediate inputs, which adjust more smoothly and respond more directly to productivity shocks. The idea is that if a firm’s investment or materials choice responds monotonically to productivity, we can

²⁹Even if a firm takes input prices as given, general equilibrium effects can create problems. If all firms in a market receive a positive productivity shock, the market price for their output may fall, creating a correlation between the shock a_{it} and the output price p_{it} that contaminates the residual.

invert that relationship to express unobserved productivity as a function of observables:

$$a_{it} = h^{-1}(k_{it}, l_{it}; m_{it}).$$

Under these assumptions, substituting this into the production function controls for the endogenous productivity term. In practice, the approach proceeds in three steps:

1. *First stage:* Regress output on inputs and the control variable (typically with a flexible polynomial) to estimate the combined production and control function $\phi_t(k_{it}, l_{it}, m_{it})$. The residual from this regression is interpreted as measurement error. The fitted value $\hat{y}_{it}$ provides output cleansed of classical measurement error. Individual production function parameters are not yet identified, since the materials coefficient cannot be separated from the part of $h^{-1}(\cdot)$ that depends on materials.
2. *Second stage:* For candidate parameters $(\gamma^K, \gamma^L, \gamma^M)$, back out productivity as $\hat{a}_{it}(\gamma) = \hat{y}_{it} - \gamma^K k_{it} - \gamma^L l_{it} - \gamma^M m_{it}$.
3. *Third stage:* Assume productivity is Markovian. Hence, we can regress $\hat{a}_{it}$ on $\hat{a}_{i,t-1}$ to decompose it into a predictable component and an innovation ξ_{it} . Because ξ_{it} is orthogonal to inputs chosen before the innovation is realized, lagged inputs form the basis of moment conditions for generalized method of moments (GMM): $\mathbb{E}[\xi_{it}(\gamma) \cdot z_{i,t-1}] = 0$. With revenue data, however, this step recovers revenue productivity $a_{it} + p_{it}$, not physical productivity. The Markov assumption must hold for the composite—a strong requirement when demand and technology respond to different forces.

Akerberg et al. (2015) refine the approach by assuming labor is chosen before materials and moving to a value-added specification. Labor then keeps variation of its own, and the materials elasticity is not needed.³⁰ But identification remains fragile. When materials optimally adjust to productivity, they become collinear with it, leaving no independent variation to identify how inputs transform into output. A second flexible input fares no better, since it is pinned down by the same state variables. An input escapes only if chosen before productivity is fully known, as labor is here. Gandhi et al. (2020) show the same point nonparametrically for gross-output settings, where the flexible input’s elasticity is exactly the object markup inference requires. Their remedy backs the elasticity out of the input’s revenue share—which presumes perfect competition. Flynn et al. (2019) show that, with markups, lagged flexible instruments can generate spurious skewness in the estimated

³⁰The value-added specification requires that materials enter in fixed proportions so the flexible primary input’s elasticity is identified. But as Section 4.3 shows, the Leontief fixed-proportions assumption is rejected by the data, leaving measured real value added contaminated by the markup.

markup distribution; imposing constant returns to scale resolves this identification failure.

Three assumptions are load-bearing—and each is strained by imperfect competition. First, what is known as the *scalar unobservable* assumption requires productivity to be the *only* unobserved driver of the proxy variable. But in the richer-residual world, input choices also respond to output prices and markups, so the proxy can reflect multiple unobservables and not be invertible. For instance, Akerberg et al. (2007) show that when there are two latent state variables—productivity and an unobserved markup determinant—two independent controls are needed, not one. Second, *monotonicity* requires the proxy to be strictly increasing in productivity, but when more productive firms charge lower prices (the Klette-Griliches logic), revenue productivity $a_{it} + p_{it}$ may not be monotonic in a_{it} . The monotonicity assumption can also fail for macroeconomic reasons: with sticky prices or inelastic short-run demand, firms may initially *reduce* flexible input use following a positive productivity shock and only expand later, as documented by Galí (1999) and Basu et al. (2006). Doraszelski and Li (2025) show that invertibility is testable and propose modifications when it fails. Third, the *Markov* assumption on productivity must now hold for the composite residual $a_{it} + p_{it}$, which requires productivity and prices to evolve in lockstep—implausible when demand responds to different forces than technology.

Proposed remedies tend to reintroduce the structural assumptions the production approach was designed to avoid. Doraszelski and Jaumandreu (2019) show imperfect competition renders the control function uninvertible when input choices depend on unobserved markup determinants. De Loecker et al. (2020) and De Ridder et al. (2026) tackle this issue by adding market shares as a second control variable. But this requires additional assumptions about the competitive environment. Kirov et al. (2026) instead tackle the issue by including observable price controls directly in the first-order conditions (see also Kasahara and Sugita, 2020; Akerberg and De Loecker, 2024). These strategies capture price variation that drives input choices, but their validity rests on assumptions about firm behavior—precisely the structural detail the production approach aims to avoid.

These problems concern whether control functions succeed on their own terms. But there is a more basic gap. Even when the proxy assumptions hold, the method addresses Layer 1 (unobserved productivity) without resolving Layer 2 (unobserved output prices). In our view, that gap, not the proxy conditions, is the binding constraint on using control functions for markup estimation. The DeSouza-Bond et al. uninformativeness problem remains unless firm-level prices or an explicit demand model supplement the control function. A practical recommendation is to report whether the proxy addresses output-price variation, input-price variation, or both, and conduct sensitivity checks—such as the invertibility test of Doraszelski

and Li (2025)—to assess whether the scalar unobservable assumption is plausible in their setting.

Dynamic panel methods replace the scalar unobservable assumption with time-series assumptions, but those assumptions must now hold for the full composite residual. These methods difference the production function to remove fixed effects and use lagged inputs as instruments for current input changes, relying on the assumption that lagged inputs are uncorrelated with productivity innovations (Arellano and Bond, 1991; Blundell and Bond, 1998, 2000). System GMM combines differenced and level equations to improve efficiency. With revenue data, however, the error term becomes $\Delta a_{it} + \Delta p_{it}$, so the moment conditions require both productivity and prices to follow similar time-series processes—a demanding and implausible requirement when demand and technology respond to different forces. Bond and Söderbom (2005) also note that lagged instruments only have identifying power in the presence of frictions such as adjustment costs, but adjustment costs violate the static first-order condition that underpins the markup calculation itself. Brand (2019) proposes treating observed output as a noisy signal of productivity, allowing unobservables to evolve nonlinearly, and using lagged output as an instrument when productivity is persistent and measurement error is not. A valuable validation would use datasets that separate output prices from quantities to test whether physical and revenue productivity have time-series properties consistent with the approach—those assumptions underpin the estimated elasticities and hence any markups inferred from them. These refinements are promising, but the fundamental tension between the time-series assumptions and the composite residual remains.

Structural demand modeling is the most direct solution to the omitted-price problem, but it sacrifices the production approach’s model-free appeal. If we are willing to specify a demand system, we can derive estimating equations in terms of observed revenue rather than unobserved quantities. For example, Klette and Griliches (1996) propose combining the production function with an isoelastic CES demand curve:

$$y_{it} = y_t - \eta(p_{it} - p_t),$$

where η is the (absolute value of the) price elasticity of demand and y_t , p_t are industry (not firm) quantity and price indexes. Substituting into the production function yields a revenue-based estimating equation whose coefficients are combinations of output elasticities and the demand elasticity:

$$r_{it} - p_t = \frac{\eta - 1}{\eta} (\gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it}) + \frac{1}{\eta} y_t + \frac{\eta - 1}{\eta} a_{it}.$$

Output elasticities can then be separated from demand elasticities and estimated using control function methods. De Loecker (2011a) extends this approach to multiproduct firms; Ruzic and Ho (2023) and Choi et al. (2024) extend it to models with heterogeneous markups and oligopolistic competition. The benefit is transparency. Structural assumptions are explicit rather than buried in econometric corrections, and a fuller model supports richer counterfactual analysis. The cost is that markups become functions of the assumed demand and conduct model—returning, in effect, to traditional IO methods. If demand and competition must be specified to estimate output elasticities, the production approach’s central appeal—avoiding those specifications—is substantially weakened.

These methods face the same fundamental tension. The FOC that delivers $\mu = \gamma/s$ assumes a flexibly adjusted input, but identifying γ often relies on frictions or auxiliary structure that contradicts that FOC.

6.4 Recommendations for applied work

Stated plainly, no existing method fully resolves the endogeneity problem. The production approach’s promise of estimating markups without specifying market structure remains, in its most ambitious form, unfulfilled. What follows is practical guidance given that reality.

First, firm-level price data are the cleanest fix when available. Collecting firm-specific output and input prices can eliminate the output-price bias directly, without overburdening the econometrics. Manufacturing and trade datasets increasingly contain quantity information for homogeneous goods, allowing researchers to construct firm-level prices. De Ridder et al. (2026) provide the most systematic assessment using French manufacturing data, comparing revenue-based and quantity-based markup estimates. Their simulations—calibrated to a specific oligopoly model (Cournot competition)—yield a 0.94 correlation between revenue-based and true log markups. The French data are less reassuring. For the level of the markup the revenue- and quantity-based estimates correlate only about 0.3, and even for dispersion and changes over time the correlation, at roughly 0.6 to 0.8, falls short of the simulation benchmark. Where the level is the object of interest, revenue data are a poor proxy, and firm-level quantity data should be used whenever available. We return to the gap between the simulation and the data below. Services and differentiated products remain challenging, but where quantity data exist, they should be used. For differentiated products, combining product-level data with production records—as in De Ridder et al. (2026)—can approximate firm-level prices even when true quantities are unavailable.

Second, instrumental variables deserve serious consideration when strong instruments are

available. IV is the most assumption-transparent method in the toolkit. The exclusion restriction is explicit, and overidentifying restrictions can be tested when multiple instruments are available—unlike the scalar unobservable and monotonicity assumptions underlying control functions. When firm-level input prices are available, they provide a natural candidate instrument that can address both layers of endogeneity simultaneously. The main caveat is input-market power. If firms with monopsony power in labor or materials markets face systematically lower input prices, input prices are correlated with the input-price term in the residual, violating the exclusion restriction.

Third, control functions remain the workhorse, but research needs to be explicit about what they leave unresolved. The approach is useful because it requires fewer external instruments than IV, and its practical implementation is well-developed (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Akerberg et al., 2015). But in settings with meaningful price dispersion or market power, the scalar unobservable assumption may be violated, the monotonicity assumption may be strained, and the Markov assumption on the composite residual becomes difficult to defend. Researchers using control functions should report sensitivity to these assumptions and, where possible, test invertibility following Doraszelski and Li (2025).

Fourth, the applied economist should be honest about the gap between the toolkit and the problem. Each method makes different compromises between identification strength and the assumptions underlying markup inference. Fixed effects and deflation work only under restrictive conditions. IV needs variation uncorrelated with all three residual components—a high bar. Control functions assume productivity is the sole unobservable, failing when demand affects input choices. Dynamic panels rely on implausible joint time-series assumptions for the composite residual. Structural demand modeling addresses these issues by imposing explicit structure, but sacrifices the model-free appeal of the production approach.

The right choice depends on the research question, the institutional setting, and the data available. What should not vary is the obligation to be transparent: state clearly which sources of bias the chosen method addresses, and which it does not. Acknowledging these limits honestly is the foundation; Section 7 builds on it with concrete recommendations for making production-based markup estimation more reliable and more transparent.

7 A call to arms

We began this review with a unifying FOC linking revenue shares, output elasticities, and markups. We end by using that same FOC to launch three methodological calls to arms. First, a transparency decomposition to quantify how much revenue-share variation reflects

markups versus output elasticities. Second, stress-testing estimates of market power. Third, more work mapping firm-level markup heterogeneity into macroeconomic models.

7.1 Transparency: An R-squared for revenue shares

A recurring theme in this review is how much of the observed variation in revenue shares, across firms and over time, reflects demand-side forces like markups? How much instead reflects technological differences in output elasticities?

Implementing the production approach frequently imposes strong parametric restrictions on output elasticities. As a result, variation in revenue shares gets disproportionately loaded onto markups. To see why, take logs of FOC eq. (3) for input X^j :

$$\ln \hat{s}_{it}^j = \ln \hat{\gamma}_{it}^j - \ln \hat{\mu}_{it}.$$

This identity says that every firm-level deviation in a revenue share must be accounted for by either the output elasticity or the markup. The variance decomposition makes this explicit:

$$\text{Var}(\ln \hat{s}_{it}^j) = \text{Var}(\ln \hat{\gamma}_{it}^j) + \text{Var}(\ln \hat{\mu}_{it}) - 2 \text{Cov}(\ln \hat{\gamma}_{it}^j, \ln \hat{\mu}_{it}).$$

When estimation restricts how much output elasticities can vary—compressing $\text{Var}(\ln \hat{\gamma}_{it}^j)$ relative to $\text{Var}(\ln \gamma_{it}^j)$ —the estimated markup inherits the leftover variation by construction. That is, parametric restrictions on output elasticities mechanically raise the dispersion of estimated markups and their explanatory power for revenue shares.

Consider this stark Cobb-Douglas example. Suppose we estimate a single output elasticity $\hat{\gamma}^j$ shared by all firms in a given period. Then $\text{Var}(\ln \hat{\gamma}_{it}^j) = 0$ and $\text{Cov}(\ln \hat{\gamma}_{it}^j, \ln \hat{\mu}_{it}) = 0$ by construction, and the identity collapses to:

$$\text{Var}(\ln \hat{s}_{it}^j) = \text{Var}(\ln \hat{\mu}_{it}).$$

All revenue-share variation is left to the estimated markup by construction. Regressing $\ln \hat{s}_{it}^j$ on $\ln \hat{\mu}_{it}$ yields an R^2 of exactly one because the parametric restriction on output elasticities forces all variation into the estimated markup. With Cobb-Douglas, it is impossible to distinguish a world where markups truly explain all variation from one where a restrictive functional form simply leaves no room for output elasticities to explain any.

Rather than requiring a full variance decomposition, the R^2 from a simple regression offers

a practical transparency check. Consider:

$$\ln \hat{s}_{it}^j = \alpha + \beta \ln \hat{\gamma}_{it}^j + \varepsilon_{it},$$

with α the constant, β the slope coefficient and ε_{it} the error term. The R^2 of this regression measures the share of revenue-share variation linearly explained by the estimated output elasticity. The remaining $1 - R^2$ is the fraction of revenue-share variation not linearly explained by the estimated output elasticity. Researchers can also run this within dimensions of interest (year, industry, or industry-year) by including fixed effects; the within R^2 then quantifies relative explanatory power conditional on those controls.

Table 2 applies this diagnostic to Compustat data using replication code from existing papers. The Output Elasticity columns report the R^2 from the above regression; the residual columns report $1 - R^2$. All rows use a Compustat sample based on the De Loecker et al. (2020) protocol and use COGS as the expenditure proxy for the flexible input.

Table 2: An R^2 Diagnostic for Revenue-Share Variation

	Full Sample		Industry-Year	
	Output Elasticity	Unexplained Residual	Output Elasticity	Unexplained Residual
De Loecker et al. (2020)				
<i>Replication File</i>				
Cobb-Douglas, Industry-Year	0.04	0.96	0.00	1.00
<i>Cost Shares under Constant RTS</i>				
Cost Share, 2-digit NAICS	0.00	1.00	0.00	1.00
Cost Share, 4-digit NAICS	0.01	0.99	0.00	1.00
De Ridder et al. (2026)				
<i>Markup Toolbox: Translog, 2nd order</i>				
No interactions of inputs, Firm-Year	0.01	0.99	0.01	0.99
Interactions of inputs, Firm-Year	0.11	0.89	0.07	0.93

Note: All rows use the same Compustat sample ($N = 190,388$, 1980–2016), cleaned following the De Loecker et al. (2020) protocol with support restrictions from De Ridder et al. (2026). Cost of Goods Sold (COGS) is the expenditure proxy for the flexible input throughout; SG&A is excluded to match the De Loecker et al. (2020) specification. We report the R^2 from regressions of form $\ln \hat{s}_{it}^j = \alpha + \beta \ln \hat{\gamma}_{it}^j + \varepsilon_{it}$ in the Output Elasticity columns: the share of revenue-share variation linearly explained by the estimated output elasticity. The residual columns report $1 - R^2$, the share of revenue-share variation not linearly explained by estimated output elasticities and therefore mechanically left to the markup component in the FOC. Cost-share rows proxy output elasticities with median industry-year COGS cost shares under constant RTS.

Start with the Cobb-Douglas benchmark from De Loecker et al. (2020), which estimates sector-year output elasticities. Adding industry-year fixed effects to the regression sets the output elasticity’s explanatory power to exactly zero by construction: a single elasticity applies to every firm in a given industry-year. Hence, all within-industry variation in revenue shares is left to markups. More striking is how little changes in the full sample. Time-varying elasticities can absorb some revenue-share variation, but only 4 percent of it.

Cost-share proxies under constant RTS (Section 4.2) tell a starker story. Because median industry-year cost shares assign one elasticity per industry-year cell, they have zero within-cell variation by construction. In the full sample, cost shares explain at most 1 percent of revenue-share variation; the rest is left to the markup term by construction.

A natural response is to estimate elasticities more flexibly. Translog production functions allow estimated coefficients to interact with firm-specific inputs, producing output elasticities that vary across firms and time. The bottom panel of Table 2 reports results using the De Ridder et al. (2026) toolbox. Even here, the pattern persists. A translog without cross-input interactions explains about 1 percent of revenue-share variation with output elasticities, whether in the full sample or within industry-year. Adding cross-input interactions raises the elasticity share to 11 percent in the full sample and 7 percent within industry-year—the highest of any specification, but most revenue-share variation is still left to markups.

At face value, the translog results seem consistent with output elasticities not varying much. But identifying translog terms requires enough similar firms to estimate how input ratios affect output. Compustat has few firms in narrow industries, and its inputs—Cost of Goods Sold and book capital—are coarse. Whether the low R^2 for elasticities reflects limited elasticity variation or a data limitation is hard to say with these data alone.

The table illustrates a broader point. The more that estimation restricts output elasticities, the more revenue-share variation is mechanically left to markups. The R^2 does not reveal how much elasticity variation is truly in the data. It reveals only how a given specification divides revenue-share variation between elasticities and markups. That division is what readers need. It shows how much of the estimated markup variation is built in by the specification. Reporting it should be routine. The R^2 makes doing so essentially free.³¹

³¹Because $\ln s = \ln \gamma - \ln \mu$ is an identity, the full variance decomposition includes a covariance term. For the specifications in Table 2, this term is small enough that the simple regression R^2 closely tracks the variance shares. The advantage of the regression is that it requires no additional computation beyond what applied researchers already produce.

7.2 Stress-testing market power

A useful step is to stress-test production-based markup estimates. Three strategies are especially promising: comparing production-based markups against demand-based markups common in industrial organization; using simulated data—where the truth is known—to evaluate estimator performance; and exploiting natural experiments to test whether markup estimates detect real shifts in competitive conditions.

First, the industrial organization literature has long estimated markups from the demand side. In principle, both demand and production approaches measure the same object—the ratio of price to marginal cost—but with different data, assumptions, and methods. Comparing them in settings where both can be applied offers a useful check. When both approaches agree, we gain confidence that the empirical signal is robust. When they diverge, the comparison can pinpoint what each method is and is not capturing. Such cross-method comparisons remain rare.

The demand approach derives markups from a model of consumer behavior and firm pricing, typically applied to a narrowly defined industry; the classic example is Berry et al. (1995)’s study of the automobile market. Researchers specify a structural demand system, recover price elasticities from observed market shares and prices, and use a conduct model (such as Bertrand competition) to back out markups. The approach enables detailed welfare and counterfactual analysis but requires rich market-level data, functional-form assumptions about preferences and competition, and can be sensitive to instrument selection.

Do demand and production approaches to markups agree in practice? Evidence is mixed. Grieco et al. (2024) study the U.S. auto industry from 1980 to 2018. Demand-based estimates show markups declining substantially, while production-based estimates from De Loecker et al. (2020) show markups that are lower but rising. The authors trace much of the divergence to demand-side improvements in product quality and variety that production-based methods may not fully capture, suggestive of what each approach actually measures. A more encouraging comparison comes from De Loecker and Scott (2025)’s study of the U.S. beer industry, where both methods deliver broadly similar average markups, at least for certain years and specifications. Döpper et al. (2025) estimate demand-based markups for more than 100 consumer product categories. They find evidence of rising markups for 2006–2019 that are consistent with De Loecker et al. (2020) in direction though not in magnitude.

Second, a complementary strategy is to use simulated data—where the true markups are known—to evaluate how well production-based estimators perform. De Ridder et al. (2026) conduct Monte Carlo simulations of this kind. Their results suggest that, in their spe-

cific model environment, using revenue rather than quantity data leads to biased markup estimates; they argue that trends and dispersion are better recovered than levels. As we discussed in Section 6, their French data partly bear this out, with much higher revenue-versus-quantity correlations for trends and dispersion than for levels. Two points temper the conclusion, however. Their estimates hold the production-function parameters fixed over the sample, so the trend result presumes the time-stable technology that Section 4.2 questions. And because the production-approach markup is an estimated output elasticity divided by the observed revenue share, some of the trend correlation may reflect the revenue share itself; the natural benchmark, the correlation the revenue share alone delivers, is not reported. Controlled simulation exercises nonetheless help map out the conditions under which production-based estimation can be trusted.

Third, natural experiments that generate plausibly independent shifts in competitive conditions can help test whether markup estimates detect the expected response. Miller and Weinberg (2017) use the MillerCoors joint venture as such a test, showing that demand-based markup estimates correctly pick up the resulting price increases. Carrillo et al. (2026) take a different approach. They examine quasi-random variation in public procurement contracts for construction services in Ecuador, which allows them to identify features of the marginal product distribution, test for misallocation, and estimate welfare losses. Hughes and Majerovitz (2025) go further, using a randomized experiment providing grants to Sri Lankan microenterprises to estimate the dispersion of marginal products—and hence the costs of misallocation—without imposing any production function assumptions. These settings not only help assess production-based methods but, when combined with structural estimation, can generate insights about firm behavior under imperfect competition that reduced-form approaches alone cannot deliver.

Although stress testing should be standard, not every paper needs to do everything. The production approach already demands expertise across multiple fields, and research benefits from specialization. Some researchers will push the frontier of measurement; others will assess robustness through demand modeling, experimental design, or simulation. What we advocate is a collective effort. Studies with rich data should compare multiple approaches; papers introducing new methods should perform extensive and realistic simulations; empirical applications should seek quasi-experimental checks. Through this division of labor, the literature can better measure market power and develop a clearer understanding of when and why methods diverge.

7.3 Macro implications of micro markups

Despite the conceptual and empirical challenges, firm-level markup estimates can inform our understanding of productivity, welfare, and resilience to shocks. We suggest three research questions to guide the agenda for bridging micro measurement and macro modeling: What economic forces underlie measured markups? How much firm-level heterogeneity is needed to explain macroeconomic outcomes? And how do markups interact with the economy’s network of input-output linkages?

First, knowing that markups exist is not enough. For policy analysis, we also need to know why firms charge them. Markups can arise from different sources, including consumer-demand elasticities, strategic pricing in oligopoly markets, or search frictions (Menzio, 2026). The production approach, by design, recovers equilibrium markups without taking a stand on what generates them. This makes it a powerful tool for documenting patterns in the data, but leaves us without guidance on how to embed markups into structural models where their cause determines the welfare and growth implications.

Concretely, future research can help quantify the prevalence of “good” versus “bad” markups. Markups that reflect pure economic profits invite antitrust scrutiny, whereas those that merely cover fixed costs or reward innovation do not, so conflating the two can misdirect policy. Bad markups reflect entry barriers or inefficient conduct that depresses output and growth (Bresnahan, 1989; Berry et al., 2019; Covarrubias et al., 2020). Good markups are the return to innovation in Schumpeterian models (Klette and Kortum, 2004; Peters, 2020). Consistent with markups balancing incentive effects against efficiency costs, Aghion et al. (2005) find an inverted-U relationship between competition and innovation (see Gilbert 2006 for a review). The difficulty is that production-markup estimates alone cannot tell the two apart, since the first-order condition recovers the equilibrium markup whatever its source. Distinguishing them requires auxiliary evidence, such as entry and exit, R&D, or the productivity of high-markup firms. The same facts admit both readings. Autor et al. (2020) find evidence consistent with superstar firms whose high markups reflect efficient scale and innovation. Aghion et al. (2023) argue that comparable markups can stem from process inefficiencies and R&D misallocation, that is, “bad” rents. Quantifying how prevalent each type is would help researchers model the source of markups in macroeconomic models.³²

Second, how much heterogeneity matters for macro outcomes is a central open question. In other words, do we need the full distribution of firm-level markups to draw reliable macro conclusions, or does a representative firm get us close enough? On the household side,

³²Aghion et al. (2023) and De Ridder (2024) argue that good rents can morph into bad rents as technological winners erect barriers to entry, ultimately reducing innovation and growth.

macroeconomists debate whether simplified two-agent New Keynesian (TANK) models capture enough heterogeneity, or whether richer heterogeneous-agent New Keynesian (HANK) models are required despite being harder to work with (Galí et al., 2007; Kaplan et al., 2018). An analogous question arises for firms. Tractability pushes modelers toward parsimony; but ignoring important dimensions of heterogeneity risks distorting key counterfactuals.

Economic theory has not yet resolved how to aggregate firm-level markups into representative macro models. Rotemberg and Woodford (1995, 1999) assume there is a representative firm that uses value added and intermediates in fixed proportion, which gives rise to a value-added aggregate production function with a value-added markup as in equation (14). But as discussed in Section 4.3, this is not the world we live in. Basu and Fernald (1997) argue that the right way to aggregate firm-level elasticities and markups depends on model specifics; for example, they provide a stylized example where the answer depends on whether marginal inputs are allocated proportionally to average inputs over the business cycle. In the context of profit shares, Hasenzagl and Pérez (2026) find that the same firm-level markup distribution can imply different macro outcomes depending on model specifics.³³

Empirically, many aspects of heterogeneity shape outcomes. Revenue shares vary widely across firms but not randomly. Large, expanding firms tend to have low labor shares (Autor et al., 2020; Kehrig and Vincent, 2021). Van Reenen (2018) documents growing dispersion across firms in sales, productivity, and wages; he finds that increases in the measured aggregate markup reflect reallocation toward larger, more productive firms. Workhorse models feature firm heterogeneity in productivity (Ericson and Pakes, 1995; Hopenhayn, 1992; Melitz, 2003) and demand elasticities (Atkeson and Burstein, 2008). In contrast to the empirical evidence, however, they typically assume common production technologies within industries. How one aggregates also matters. Edmond et al. (2023) show sales-weighted markups rise much more steeply than cost-weighted ones, with the gap reflecting allocative inefficiency from dispersion.

Third, research can help evaluate how the economy’s network structure shapes the welfare consequences of markups. A long tradition of network analysis—from Leontief (1941) to work reviewed by Carvalho and Tahbaz-Salehi (2019) and Baqaee and Rubbo (2023)—shows that intermediate input flows can amplify or attenuate distortions, so the location and propagation of markups through the network are as important as their average level.

All else equal, upstream markups are more costly to the economy than downstream ones because they distort production directly rather than mainly redistributing income. A markup

³³Growth accounting has long emphasized that microeconomic heterogeneity in markups and returns to scale affects macro aggregates (e.g., Basu and Fernald (1997, 2002)).

on goods sold to final consumers may redistribute income from consumers to firms, with limited real effects, especially if labor is supplied inelastically. In contrast, a markup on intermediate inputs sold to other producers acts as a tax on production, distorting input choices and depressing aggregate productivity. It violates the Diamond and Mirrlees (1971) production-efficiency theorem and pushes the economy inside its production-possibilities frontier (Basu, 1995; Basu and Fernald, 2002; Bigio and La’O, 2020; Baqaee and Farhi, 2020).

Whether markups actually generate misallocation also depends on how firms contract with one another. When firms transact on spot markets, markups compound along supply chains through “double marginalization,” magnifying inefficiencies at each stage. But when upstream and downstream firms bargain directly, as in many long-term supplier relationships, markups need not generate misallocation even when they appear large. The macro consequences of markup heterogeneity therefore cannot be read off from markup estimates alone; they depend on network position and the institutional arrangements through which firms trade. Future research should clarify when markups primarily redistribute income, when they distort production, and how to incorporate these distinctions into tractable macro models.

7.4 Takeaways for practice

The preceding discussion points to three research priorities.

First, researchers should report the R^2 diagnostic for revenue-share variation. If estimated output elasticities explain little of that variation, that could mean the specification leaves too little room for technology.

Second, if possible, production-based markups should be tested against outside evidence—demand-based estimates, simulations with known truth, or quasi-experimental shifts in competitive conditions. Such comparisons remain rare. Expanding them is the most promising near-term path to putting the literature on firmer empirical footing.

Third, aggregating from firm to macro requires understanding when markup heterogeneity reflects misallocation versus efficient dispersion, and how network position shapes the welfare consequences of markups. This is the least resolved of the three and the most important for macroeconomists who want to use markup estimates in general-equilibrium models.

Progress on all three—transparency, testing, and aggregation—requires confronting the fundamental theme of this review: production-based markups are residuals, and residuals are only as informative as the models, data, and econometrics that generate them.

8 Conclusion

The production approach to markups is both powerful and fragile. Its power comes from its minimal structure. With cost minimization and a flexible input, all we need is the input’s share in revenue and its output elasticity to back out a markup. Its fragility comes from the same source. Because markups are residuals, whatever the model, data, or econometrics fail to capture loads onto the markup. Researchers make a series of choices, each with risks.

For markup estimation, Hall (1986, 1988, 1990) marked a shift from “craft to mass production” (Basu, 2024). Craft methods in IO focus on careful studies of narrow industries with rich structural detail. But macroeconomists and trade economists need scale. Hall’s production approach and the large subsequent literature allow such scale.

But mass production requires more than a factory blueprint. For a factory to produce, the firm needs local institutional knowledge to acquire needed permits, reliably source inputs, and much besides. Likewise, production-based markups require defensible input choices, data that match the theoretical objects, and credible elasticity estimates. This review has identified challenges that need to be addressed.

Production-based markup estimates should not be taken at face value. The disagreements across studies are not minor. Using the same data, different inputs imply opposite trends, different production functions yield markedly different levels, and different estimation methods produce divergent results. These disagreements reflect non-markup frictions, specification choices, data limitations, and econometric pitfalls—and they signal that the production approach’s assumptions are difficult to satisfy in practice. More systematic attention to the checks developed in the preceding sections would sharpen the evidence on market power and reduce the risk that conclusions rest on artifacts of specification or measurement. That requires transparency, testing, robustness, and better data.

What should economists do with the evidence? Despite uncertainties, we can press ahead on both theory and measurement. On the theory side, there is no reason to wait. Markups shape resource allocation and welfare. They interact with misallocation, innovation, network structure, and cyclical dynamics. They also shape how shocks propagate across firms and industries. But we also need evidence on the structures that actually drive markups to guide theory. The way forward is to use the approach honestly: report what the residual absorbs, test estimates against outside evidence, and treat specification choices as consequential rather than interchangeable. Measurement leaves us with residuals. The challenge is to combine institutional knowledge and economic theory to discipline what those residuals mean for growth, welfare, and policy.

References

- Acemoglu, Daron and Alireza Tahbaz-Salehi**, “The Macroeconomics of Supply Chain Disruptions,” *Review of Economic Studies*, 2025, *92*, 656–695.
- Akerberg, Daniel and Jan De Loecker**, “Production Function Identification Under Imperfect Competition,” 2024. CEPR Discussion Paper No. 19640.
- , **C. Lanier Benkard, Steven Berry, and Ariel Pakes**, “Econometric Tools for Analyzing Market Outcomes,” in “Handbook of Econometrics,” Vol. 6, Elsevier, 2007, pp. 4171–4276.
- , **Kevin Caves, and Garth Frazer**, “Identification Properties of Recent Production Function Estimators,” *Econometrica*, 2015, *83*, 2411–51.
- Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li**, “A Theory of Falling Growth and Rising Rents,” *Review of Economic Studies*, 2023, *90*, 2675–2702.
- , **Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt**, “Competition and Innovation: An Inverted-U Relationship,” *Quarterly Journal of Economics*, 2005, *120*, 701–728.
- Ahmad, Mumtaz, John G. Fernald, and Hashmat Khan**, “Returns to Scale in U.S. Production, Redux,” 2024. Working Paper.
- Ali, Ashiq, Sandy Klasa, and Eric Yeung**, “The Limitations of Industry Concentration Measures Constructed with Compustat Data: Implications for Finance Research,” *Review of Financial Studies*, 2009, *22*, 3839–3871.
- Arellano, Manuel and Stephen Bond**, “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations,” *Review of Economic Studies*, 1991, *58*, 277–297.
- Asker, John, Allan Collard-Wexler, and Jan De Loecker**, “Dynamic Inputs and Resource (Mis)Allocation,” *Journal of Political Economy*, 2014, *122*, 1013–1063.
- Atkeson, Andrew and Ariel Burstein**, “Pricing-to-Market, Trade Costs, and International Relative Prices,” *American Economic Review*, 2008, *98*, 1998–2031.
- Autor, David, David Dorn, Lawrence Katz, Christina Patterson, and John Van Reenen**, “The Fall of the Labor Share and the Rise of Superstar Firms,” *Quarterly Journal of Economics*, 2020, *135*, 645–709.
- Baily, Martin, Charles R. Hulten, and David Campbell**, “Productivity Dynamics in Manufacturing Plants,” *Brookings Papers on Economic Activity: Microeconomics*, 1992, *1992*, 187–267.
- Bain, Joe S.**, “Relation of Profit Rate to Industry Concentration: American Manufacturing, 1936–1940,” *Quarterly Journal of Economics*, 1951, *65*, 293–324.

- Baqae, David Rezza and Elisa Rubbo**, “Micro Propagation and Macro Aggregation,” *Annual Review of Economics*, 2023, 15, 91–123.
- **and Emmanuel Farhi**, “Productivity and Misallocation in General Equilibrium,” *Quarterly Journal of Economics*, 2020, 135, 105–163.
- , – , **and Kunal Sangani**, “The Darwinian Returns to Scale,” *Review of Economic Studies*, 2024, 91, 1373–1405.
- Barkai, Simcha**, “Declining Labor and Capital Shares,” *Journal of Finance*, 2020, 75, 2421–2463.
- Basu, Susanto**, “Intermediate Goods and Business Cycles: Implications for Productivity and Welfare,” *American Economic Review*, 1995, 85, 512–531.
- , “Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence,” *Journal of Economic Perspectives*, 2019, 33, 3–22.
- , “Basu Remarks on Bob Hall,” 2024. Unpublished slides.
- **and Christopher L. House**, “Allocative and Remitted Wages: New Facts and Challenges for Keynesian Models,” in “Handbook of Macroeconomics,” Vol. 2, Elsevier, 2016, pp. 297–354.
- **and John G. Fernald**, “Are Apparent Productive Spillovers a Figment of Specification Error?,” *Journal of Monetary Economics*, 1995, 36, 165–188.
- **and –** , “Returns to Scale in U.S. Production: Estimates and Implications,” *Journal of Political Economy*, 1997, 105, 249–283.
- **and –** , “Why Is Productivity Procyclical? Why Do We Care?,” in “New Developments in Productivity Analysis,” University of Chicago Press, 2001, pp. 225–302.
- **and –** , “Aggregate Productivity and Aggregate Technology,” *European Economic Review*, 2002, 46, 963–991.
- , – , **and Matthew D. Shapiro**, “Productivity Growth in the 1990s: Technology, Utilization, or Adjustment?,” *Carnegie-Rochester Conference Series on Public Policy*, 2001, 55, 117–165.
- , – , **and Miles S. Kimball**, “Are Technology Improvements Contractionary?,” *American Economic Review*, 2006, 96, 1418–1448.
- BEA**, “GDP by Industry,” <https://www.bea.gov/data/gdp/gdp-industry> 2026. April 9, 2026.
- Becker, Randy A., John C. Haltiwanger, Ron S. Jarmin, Shawn D. Klimek, and Daniel J. Wilson**, “Micro and Macro Data Integration: The Case of Capital,” in “A New Architecture for the US National Accounts,” University of Chicago Press, 2006, pp. 541–610.

- Benkard, C. Lanier, Nathan H. Miller, and Ali Yurukoglu**, “Rejoinder to “The Rise of Market Power and the Macroeconomic Implications: Reply to Benkard, Miller, and Yurukoglu (2025)”,” 2025. Working Paper.
- , – , and – , “The Rise of Market Power and the Macroeconomic Implications: Comment,” 2025. NBER Working Paper No. 34513.
- Berger, David, Kyle Herkenhoff, and Simon Mongey**, “Labor Market Power,” *American Economic Review*, 2022, 112, 1147–1193.
- Berman, Eli and Adam B. Jaffe**, “Zvi Griliches (1930–1999),” in “The Palgrave Companion to Harvard Economics,” Palgrave Macmillan, 2024, pp. 573–596.
- Berndt, Ernst R. and Melvyn A. Fuss**, “Productivity Measurement with Adjustments for Variations in Capacity Utilization and Other Forms of Temporary Equilibrium,” *Journal of Econometrics*, 1986, 33, 7–29.
- Berry, Steven, James Levinsohn, and Ariel Pakes**, “Automobile Prices in Market Equilibrium,” *Econometrica*, 1995, 63, 841–890.
- , **Martin Gaynor**, and **Fiona Scott Morton**, “Do Increasing Markups Matter? Lessons from Empirical Industrial Organization,” *Journal of Economic Perspectives*, 2019, 33, 44–68.
- Bigio, Saki and Jennifer La’O**, “Distortions in Production Networks,” *Quarterly Journal of Economics*, 2020, 135, 2187–2253.
- Blundell, Richard and Stephen Bond**, “Initial Conditions and Moment Restrictions in Dynamic Panel Data Models,” *Journal of Econometrics*, 1998, 87, 115–143.
- and – , “GMM Estimation with Persistent Panel Data: An Application to Production Functions,” *Econometric Reviews*, 2000, 19, 321–40.
- Bond, Stephen and Måns Söderbom**, “Adjustment Costs and the Identification of Cobb-Douglas Production Functions,” 2005. Institute for Fiscal Studies Working Paper No. W05/04.
- , **Arshia Hashemi**, **Greg Kaplan**, and **Piotr Zoch**, “Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data,” *Journal of Monetary Economics*, 2021, 121, 1–14.
- Borges, Jorge Luis**, “The Garden of Forking Paths,” in “Ficciones,” Editorial Sur, 1941.
- Bornstein, Gideon and Alessandra Peter**, “Nonlinear Pricing and Misallocation,” *American Economic Review*, 2025, 115, 3853–3908.
- Brand, James**, “Estimating Markups Under Imperfect Competition,” 2019. Working Paper.
- Bresnahan, Timothy F.**, “Empirical Studies of Industries with Market Power,” in “Handbook of Industrial Organization,” Vol. 2, Elsevier, 1989, pp. 1011–1057.

- Brynjolfsson, Erik and Lorin M. Hitt**, “Computing Productivity: Firm-Level Evidence,” *Review of Economics and Statistics*, 2003, *85*, 793–808.
- Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf**, “Does the United States Have a Productivity Slowdown or a Measurement Problem?,” *Brookings Papers on Economic Activity*, 2016, *2016*, 109–182.
- Caldwell, Sydnee, Arindrajit Dube, and Suresh Naidu**, “Monopsony Makes It Big,” *Journal of Economic Literature*, 2026. Forthcoming.
- Carrillo, Paul, Dave Donaldson, Dina Deborah Pomeranz, and Monica Singhal**, “Misallocation in Firm Production: A Nonparametric Analysis Using Procurement Lotteries,” 2026. IZA Discussion Paper No. 18612.
- Carvalho, Vasco M. and Alireza Tahbaz-Salehi**, “Production Networks: A Primer,” *Annual Review of Economics*, 2019, *11*, 635–663.
- Choi, Jaedo, Andrei A. Levchenko, Dimitrije Ruzic, and Younghun Shim**, “Superstars or Supervillains? Large Firms in the South Korean Growth Miracle,” 2024. NBER Working Paper No. 32648.
- Collard-Wexler, Allan and Jan De Loecker**, “Production Function Estimation and Capital Measurement Error,” 2016. NBER Working Paper.
- Cooper, Russell W., John C. Haltiwanger, and Jonathan Willis**, “Declining Responsiveness at the Establishment Level: Sources and Productivity Implications,” 2024. Federal Reserve Bank of Atlanta Working Paper No. 2024-3a.
- Corrado, Carol, Charles R. Hulten, and Daniel Sichel**, “Intangible Capital and U.S. Economic Growth,” *Review of Income and Wealth*, 2009, *55*, 661–685.
- Covarrubias, Matias, Germán Gutiérrez, and Thomas Philippon**, “From Good to Bad Concentration? US Industries over the Past 30 Years,” *NBER Macroeconomics Annual*, 2020, *34*, 1–46.
- Crouzet, Nicolas and Janice Eberly**, “Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles,” 2019. NBER Working Paper No. 25869.
- Davis, Steven J., John C. Haltiwanger, Ron S. Jarmin, and Javier Miranda**, “Volatility and Dispersion in Business Growth Rates: Publicly Traded versus Privately Held Firms,” in “NBER Macroeconomics Annual 2006, Volume 21,” MIT Press, 2007, pp. 107–180.
- De Loecker, Jan**, “Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity,” *Econometrica*, 2011, *79*, 1407–1451.
- , “Recovering Markups from Production Data,” *International Journal of Industrial Organization*, 2011, *29*, 350–355.
- and **Frederic Warzynski**, “Markups and Firm-Level Export Status,” *American Economic Review*, 2012, *102*, 2437–71.

- **and Paul T. Scott**, “Markup Estimation Using Production and Demand Data: An Application to the US Brewing Industry,” *Review of Economic Studies*, 2025. Forthcoming.
- **, Jan Eeckhout, and Gabriel Unger**, “The Rise of Market Power and the Macroeconomic Implications,” *Quarterly Journal of Economics*, 2020, *135*, 561–644.
- **, – , and –**, “The Rise of Market Power and the Macroeconomic Implications: Reply to Benkard, Miller, and Yurukoglu (2025),” 2025. Working Paper.
- **, Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik**, “Prices, Markups, and Trade Reform,” *Econometrica*, 2016, *84*, 445–510.
- De Ridder, Maarten**, “Market Power and Innovation in the Intangible Economy,” *American Economic Review*, 2024, *114*, 199–251.
- **, Basile Grassi, and Giovanni Morzenti**, “The Hitchhiker’s Guide to Markup Estimation: Assessing Estimates From Financial Data,” *Econometrica*, 2026, *94*, 137–168.
- de Roux, Nicolás, Marcela Eslava, Santiago Franco, and Eric Verhoogen**, “Estimating Production Functions in Differentiated-Product Industries with Quantity Information and External Instruments,” 2024. NBER Working Paper No. 28323 (October 2024 update).
- Decker, Ryan A. and Jacob Williams**, “A Note on Industry Concentration Measurement,” 2023. FEDS Notes No. 2023-02-03.
- Demirer, Mert**, “Production Function Estimation with Factor-Augmenting Technology: An Application to Markups,” 2025. Working Paper.
- Demsetz, Harold**, “Industry Structure, Market Rivalry, and Public Policy,” *Journal of Law and Economics*, 1973, *16*, 1–9.
- DeSouza, Sergio A.**, “Estimating Mark-ups from Plant-Level Data,” *The Journal of Industrial Economics*, 2009, *57*, 353–363.
- Dhyne, Emmanuel, Ayumu Ken Kikkawa, and Glenn Magerman**, “Imperfect Competition in Firm-to-Firm Trade,” *Journal of the European Economic Association*, 2022, *20*, 1933–1970.
- Diamond, Peter A. and James A. Mirrlees**, “Optimal Taxation and Public Production I: Production Efficiency,” *American Economic Review*, 1971, *61*, 8–27.
- Dixit, Avinash K. and Joseph E. Stiglitz**, “Monopolistic Competition and Optimum Product Diversity,” *American Economic Review*, 1977, *67*, 297–308.
- Dobbelaere, Sabien and Jacques Mairesse**, “Panel Data Estimates of the Production Function and Product and Labor Market Imperfections,” *Journal of Applied Econometrics*, 2013, *28*, 1–46.
- Domar, Evsey D.**, “On the Measurement of Technological Change,” *The Economic Journal*, 1961, *71*, 709–729.

- Döpfer, Hendrik, Alexander MacKay, Nathan H. Miller, and Joel Stiebale**, “Rising Markups and the Role of Consumer Preferences,” *Journal of Political Economy*, 2025, *133*, 2462–2505.
- Doraszelski, Ulrich and Jordi Jaumandreu**, “R&D and Productivity: Estimating Endogenous Productivity,” *Review of Economic Studies*, 2013, *80*, 1338–1383.
- **and —**, “Using Cost Minimization to Estimate Markups,” 2019. CEPR Discussion Paper No. 14114.
- **and Lixiong Li**, “Production Function Estimation without Invertibility: Imperfectly Competitive Environments and Demand Shocks,” 2025. NBER Working Paper No. 33939.
- Dunne, Timothy, Mark J. Roberts, and Larry Samuelson**, “Patterns of Firm Entry and Exit in US Manufacturing Industries,” *The RAND Journal of Economics*, 1988, *19*, 495–515.
- Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu**, “How Costly Are Markups?,” *Journal of Political Economy*, 2023, *131*, 1619–1675.
- Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin**, “The Decline of the U.S. Labor Share,” *Brookings Papers on Economic Activity*, 2013, *2013*, 1–63.
- Ericson, Richard and Ariel Pakes**, “Markov-Perfect Industry Dynamics: A Framework for Empirical Work,” *Review of Economic Studies*, 1995, *62*, 53–82.
- Feenstra, Robert C. and David E. Weinstein**, “Globalization, Markups, and US Welfare,” *Journal of Political Economy*, 2017, *125*, 1040–1074.
- Fernald, John G.**, “A Quarterly, Utilization-Adjusted Series on Total Factor Productivity,” 2014. Federal Reserve Bank of San Francisco Working Paper No. 2012-19.
- , “Dale W. Jorgenson (1933–2022),” in “The Palgrave Companion to Harvard Economics,” Palgrave Macmillan, 2024, pp. 597–631.
- **and Eugenio Piga**, “Comment on: Bottlenecks: Sectoral Imbalances and the US Productivity Slowdown,” in “NBER Macroeconomics Annual 2023, Volume 38,” University of Chicago Press, 2023, pp. 208–221.
- , **Robert Inklaar, and Dimitrije Ruzic**, “The Productivity Slowdown in Advanced Economies: Common Shocks or Common Trends?,” *Review of Income and Wealth*, 2025, *71*, e12690.
- Fisher, Franklin and John McGowan**, “On the Misuse of Accounting Rates of Return to Infer Monopoly Profits,” *American Economic Review*, 1983, *73*, 82–97.
- Flynn, Zach, James Traina, and Amit Gandhi**, “Measuring Markups with Production Data,” 2019. SSRN Working Paper.
- Foster, Lucia S., John C. Haltiwanger, and Chad Syverson**, “Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?,” *American Economic Review*, 2008, *98*, 394–425.

- , – , and **Cody Tuttle**, “Rising Markups or Changing Technology?,” 2026. Manuscript, University of Texas.
- Fox, Jeremy and Valérie Smeets**, “Does Input Quality Drive Measured Differences in Firm Productivity?,” *International Economic Review*, 2011, 52, 961–989.
- Galí, Jordi**, “Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?,” *American Economic Review*, 1999, 89, 249–271.
- , **J. David López-Salido, and Javier Vallés**, “Understanding the Effects of Government Spending on Consumption,” *Journal of the European Economic Association*, 2007, 5, 227–270.
- Gandhi, Amit, Salvador Navarro, and David A. Rivers**, “How Heterogeneous Is Productivity? A Comparison of Gross Output and Value Added,” 2017. University of Western Ontario Working Paper.
- , – , and – , “On the Identification of Gross Output Production Functions,” *Journal of Political Economy*, 2020, 128, 2973–3016.
- Garcia-Marin, Alvaro and Nico Voigtländer**, “Exporting and Plant-Level Efficiency Gains: It’s in the Measure,” *Journal of Political Economy*, 2019, 127, 1777–1825.
- Gelman, Andrew and Eric Loken**, “The Statistical Crisis in Science,” *American Scientist*, 2014, 102, 460–465.
- Gilbert, Richard**, “Looking for Mr. Schumpeter: Where Are We in the Competition–Innovation Debate?,” *Innovation Policy and the Economy*, 2006, 6, 159–215.
- Grieco, Paul, Charles Murry, and Ali Yurukoglu**, “The Evolution of Market Power in the US Automobile Industry,” *Quarterly Journal of Economics*, 2024, 139, 1201–1253.
- , **Shengyu Li, and Hongsong Zhang**, “Production Function Estimation with Unobserved Input Price Dispersion,” *International Economic Review*, 2016, 57, 665–690.
- Griliches, Zvi**, “Specification Bias in Estimates of Production Functions,” *Journal of Farm Economics*, 1957, 39, 8–20.
- and **Jacques Mairesse**, “Production Functions: The Search for Identification,” in “Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium,” Cambridge University Press, 1998, pp. 169–203.
- and **Vidar Ringstad**, *Economies of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data*, North-Holland Publishing Company, 1971.
- Grunfeld, Yehuda and Zvi Griliches**, “Is Aggregation Necessarily Bad?,” *Review of Economics and Statistics*, 1960, 42, 1–13.
- Gutiérrez, Germán and Thomas Philippon**, “Investmentless Growth: An Empirical Investigation,” *Brookings Papers on Economic Activity*, 2017, 2017, 89–190.

- Hall, Robert E.**, “Employment Fluctuations and Wage Rigidity,” *Brookings Papers on Economic Activity*, 1980, *11*, 91–142.
- , “Market Structure and Macroeconomic Fluctuations,” *Brookings Papers on Economic Activity*, 1986, *2*, 285–322.
- , “The Relation Between Price and Marginal Cost in US Industry,” *Journal of Political Economy*, 1988, *96*, 921–47.
- , “Invariance Properties of Solow’s Productivity Residual,” in “Growth/ Productivity/Unemployment: Essays to Celebrate Robert Solow’s Birthday,” MIT Press, 1990, pp. 71–112.
- **and Dale W. Jorgenson**, “Tax Policy and Investment Behavior,” *American Economic Review*, 1967, *57*, 391–414.
- Haltiwanger, John C., Robert Kulick, and Chad Syverson**, “Misallocation Measures: The Distortion That Ate the Residual,” 2018. NBER Working Paper No. 24199.
- Hasenzagl, Thomas and Luis Pérez**, “The Micro-Aggregated Profit Share,” 2026. Working Paper.
- Hashemi, Arshia, Ivan Kirov, and James Traina**, “The Production Approach to Markup Estimation Often Measures Input Distortions,” *Economics Letters*, 2022, *217*, 110673.
- Hopenhayn, Hugo A.**, “Entry, Exit, and Firm Dynamics in Long Run Equilibrium,” *Econometrica*, 1992, *60*, 1127–50.
- Hsieh, Chang-Tai and Peter J. Klenow**, “Misallocation and Manufacturing TFP in China and India,” *Quarterly Journal of Economics*, 2009, *124*, 1403–48.
- Hughes, David and Jeremy Majerovitz**, “Measuring Misallocation with Experiments,” 2025. Working Paper.
- Hulten, Charles R.**, “Productivity Change, Capacity Utilization, and the Sources of Efficiency Growth,” *Journal of Econometrics*, 1986, *33*, 31–50.
- Jarosch, Gregor, Jan Sebastian Nimczik, and Isaac Sorkin**, “Granular Search, Market Structure, and Wages,” *Review of Economic Studies*, 2024, *91*, 3569–3607.
- Jorgenson, Dale W. and Zvi Griliches**, “The Explanation of Productivity Change,” *Review of Economic Studies*, 1967, *34*, 249–280.
- , **Frank M. Gollop, and Barbara M. Fraumeni**, *Productivity and US Economic Growth*, Harvard University Press, 1987.
- Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante**, “Monetary Policy According to HANK,” *American Economic Review*, 2018, *108*, 697–743.
- Karabarbounis, Loukas and Brent Neiman**, “The Global Decline of the Labor Share,” *Quarterly Journal of Economics*, 2014, *129*, 61–103.

- **and** –, “Accounting for Factorless Income,” *NBER Macroeconomics Annual*, 2019, 33, 167–228.
- Kasahara, Hiroyuki and Yoichi Sugita**, “Nonparametric Identification of Production Function, Total Factor Productivity, and Markup from Revenue Data,” 2020. arXiv Working Paper.
- Kehrig, Matthias and Nicolas Vincent**, “The Micro-Level Anatomy of the Labor Share Decline,” *Quarterly Journal of Economics*, 2021, 136, 1031–1087.
- Keil, Jan**, “The Trouble with Approximating Industry Concentration from Compustat,” *Journal of Corporate Finance*, 2017, 45, 467–479.
- Kim, Ginha and James Traina**, “Markups, Production Technology, and the Cost of Capital in Financial Markets,” 2025. New York University Working Paper.
- Kim, Kyoo Il, Amil Petrin, and Suyong Song**, “Estimating Production Functions with Control Functions When Capital Is Measured with Error,” *Journal of Econometrics*, 2016, 190, 267–279.
- Kirov, Ivan and James Traina**, “Labor Market Power and Technological Change in US Manufacturing,” 2022. Working Paper.
- , **Paolo Mengano, and James Traina**, “Measuring Markups with Revenue Data,” *International Journal of Industrial Organization*, 2026, 105, 103263.
- Klette, Tor Jakob**, “Market Power, Scale Economies and Productivity: Estimates from a Panel of Establishment Data,” *The Journal of Industrial Economics*, 1999, 47, 451–476.
- **and Samuel Kortum**, “Innovating Firms and Aggregate Innovation,” *Journal of Political Economy*, 2004, 112, 986–1018.
- **and Zvi Griliches**, “The Inconsistency of Common Scale Estimators When Output Prices Are Unobserved and Endogenous,” *Journal of Applied Econometrics*, 1996, 11, 343–61.
- Koh, Dongya, Raül Santaaulàlia-Llopis, and Yu Zheng**, “Labor Share Decline and Intellectual Property Products Capital,” *Econometrica*, 2020, 88, 2609–2628.
- Kudlyak, Marianna**, “How Cyclical Is the User Cost of Labor?,” *Journal of Economic Perspectives*, 2024, 38, 159–80.
- Lenzu, Simone, David A. Rivers, Joris Tielens, and Shi Hu**, “Financial Shocks, Productivity, and Prices,” 2026. Federal Reserve Bank of New York Staff Report No. 1193.
- Leontief, Wassily**, *The Structure of American Economy, 1919–1929: An Empirical Application of Equilibrium Analysis*, Harvard University Press, 1941.
- Lev, Baruch**, *Intangibles: Management, Measurement, and Reporting*, Brookings Institution Press, 2001.

- Levinsohn, James and Amil Petrin**, “Estimating Production Functions Using Inputs to Control for Unobservables,” *Review of Economic Studies*, 2003, 70, 317–41.
- Liu, Ernest and Aleh Tsyvinski**, “A Dynamic Model of Input–Output Networks,” *Review of Economic Studies*, 2024, 91, 3608–3644.
- Ližal, Lubomír and Kamil Galuščák**, “The Impact of Capital Measurement Error Correction on Firm-Level Production Function Estimation,” 2012. William Davidson Institute Working Paper.
- Macchiavello, Rocco and Ameet Morjaria**, “Relational Contracts: Recent Empirical Advancements and Open Questions,” *Journal of Institutional and Theoretical Economics*, 2023, 179, 673–700.
- Manning, Alan**, *Monopsony in Motion: Imperfect Competition in Labor Markets*, Princeton University Press, 2003.
- Marschak, Jacob and William H. Andrews**, “Random Simultaneous Equations and the Theory of Production,” *Econometrica*, 1944, 12, 143–205.
- McAdam, Peter, Philipp Meinen, Chris Papageorgiou, and Patrick Schulte**, “Returns to Scale: New Evidence from Administrative Firm-Level Data,” 2024. Deutsche Bundesbank Discussion Paper.
- Melitz, Marc**, “The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity,” *Econometrica*, 2003, 71, 1695–725.
- Menzio, Guido**, “Markups: A Search-Theoretic Perspective,” *Journal of Political Economy*, 2026. Forthcoming.
- Miller, Nathan H. and Matthew C. Weinberg**, “Understanding the Price Effects of the MillerCoors Joint Venture,” *Econometrica*, 2017, 85, 1763–1791.
- Mundlak, Yair**, “Empirical Production Function Free of Management Bias,” *Journal of Farm Economics*, 1961, 43, 44–56.
- Olley, G. Steven and Ariel Pakes**, “The Dynamics of Productivity in the Telecommunications Equipment Industry,” *Econometrica*, 1996, 64, 1263–97.
- Ornaghi, Carmine**, “Assessing the Effects of Measurement Errors on the Estimation of Production Functions,” *Journal of Applied Econometrics*, 2006, 21, 879–891.
- Peters, Michael**, “Heterogeneous Markups, Growth, and Endogenous Misallocation,” *Econometrica*, 2020, 88, 2037–2073.
- Raval, Devesh**, “Testing the Production Approach to Markup Estimation,” *Review of Economic Studies*, 2023, 90, 2592–2611.
- Restuccia, Diego and Richard Rogerson**, “Policy Distortions and Aggregate Productivity with Heterogeneous Establishments,” *Review of Economic Dynamics*, 2008, 11, 707–20.
- and —, “The Causes and Costs of Misallocation,” *Journal of Economic Perspectives*, 2017, 31, 151–74.

- Robinson, Joan**, *The Economics of Imperfect Competition*, Macmillan, 1933.
- Rogerson, Richard, Robert Shimer, and Randall Wright**, “Search-Theoretic Models of the Labor Market: A Survey,” *Journal of Economic Literature*, 2005, *43*, 959–988.
- Romer, Paul**, “Endogenous Technological Change,” *Journal of Political Economy*, 1990, *98*, S71–S102.
- Rosen, Sherwin**, “Implicit Contracts: A Survey,” *Journal of Economic Literature*, 1985, *23*, 1144–1175.
- Rotemberg, Julio J. and Michael Woodford**, “Dynamic General Equilibrium Models with Imperfectly Competitive Product Markets,” in “Frontiers of Business Cycle Research,” Princeton University Press, 1995, pp. 243–293.
- **and** —, “The Cyclical Behavior of Prices and Costs,” in “Handbook of Macroeconomics,” Vol. 1, North-Holland, 1999, pp. 1051–1135.
- Rubens, Michael**, “Market Structure, Oligopsony Power, and Productivity,” *American Economic Review*, 2023, *113*, 2382–2410.
- Ruzic, Dimitrije**, “The Factor Bias of External Inputs: Implications for Substitution between Capital and Labor,” 2026. CEPR Discussion Paper No. 21591.
- **and Sui-Jade Ho**, “Returns to Scale, Productivity, Measurement, and Trends in U.S. Manufacturing Misallocation,” *Review of Economics and Statistics*, 2023, *105*, 1287–1303.
- Sato, Kazuo**, “The Meaning and Measurement of the Real Value Added Index,” *Review of Economics and Statistics*, 1976, *58*, 434–442.
- Schmalensee, Richard**, “Inter-Industry Studies of Structure and Performance,” in “Handbook of Industrial Organization,” Vol. 2, Elsevier, 1989, pp. 951–1009.
- Solow, Robert M.**, “Technical Change and the Aggregate Production Function,” *Review of Economics and Statistics*, 1957, *39*, 312–20.
- Syverson, Chad**, “What Determines Productivity?,” *Journal of Economic Literature*, 2011, *49*, 326–65.
- , “Challenges to Mismeasurement Explanations for the US Productivity Slowdown,” *Journal of Economic Perspectives*, Spring 2017, *31* (2), 165–186.
- , “Markups and Markdowns,” *Annual Review of Economics*, 2025, *17*, 57–76.
- Traina, James**, “Is Aggregate Market Power Increasing? Production Trends Using Financial Statements,” 2018. Working Paper.
- van Heuvelen, Gerrit Hugo, Leon Bettendorf, and Gerdien Meijerink**, “Markups in a Dual Labour Market: The Case of the Netherlands,” *International Journal of Industrial Organization*, 2021, *77*, 102762.

Van Reenen, John, “Increasing Differences Between Firms: Market Power and the Macroeconomy,” in “Changing Market Structures and Implications for Monetary Policy,” Federal Reserve Bank of Kansas City, 2018.

Varian, Hal R., *Microeconomic Analysis*, W. W. Norton, 1984.

Yeh, Chen, Claudia Macaluso, and Brad Hershbein, “Monopsony in the US Labor Market,” *American Economic Review*, 2022, *112*, 2099–2138.