FEDERAL RESERVE BANK OF SAN FRANCISCO

WORKING PAPER SERIES

Micro and Macro Perspectives on Production-Based Markups

John Fernald
INSEAD
Federal Reserve Bank of San Francisco, Economist Emeritus

Amit Gandhi
Airbnb

Dimitrije Ruzic
INSEAD

James Traina
NYU Stern Abu Dhabi

September 2025

Working Paper 2025-20

https://doi.org/10.24148/wp2025-20

Suggested citation:

Fernald, John, Amit Gandhi, Dimitrije Ruzic, and James Traina. "Micro and Macro Perspectives on Production-Based Markups." Federal Reserve Bank of San Francisco Working Paper 2025-20. https://doi.org/10.24148/wp2025-20

Micro and Macro Perspectives on Production-Based Markups

John Fernald
Amit Gandhi
Dimitrije Ruzic
James Traina *

August 2025

Abstract

We review the "production approach" to estimating markups—the ratio of price to marginal cost. Paired with increasingly rich microdata and advances in production-function estimation, the method enables scalable analysis of markups across firms, industries, and time. We survey what economists need to know about the production approach, emphasizing both its promise and its fragility. Conceptually, empirically, and econometrically, the production-based markup is a residual—absorbing model misspecification, data limitations, and unobserved frictions. These challenges help explain why empirical results often diverge, including on whether markups have risen sharply in recent decades. We outline practical guidance for researchers and highlight directions for future work: improving transparency in reporting, validating production-based markups against demand-based and quasi-experimental estimates, and integrating firm-level heterogeneity into macroeconomic models. The production approach is not a finished product, but it remains a uniquely powerful tool for studying market power and its implications for productivity, welfare, and macroeconomic dynamics.

JEL Codes: D24, D43, E22, E23, L11, L16, O33, O47
Keywords: Production-based markups, market power, production-function estimation, firm heterogeneity, micro-to-macro linkages

*Fernald (john.fernald@insead.edu): INSEAD and FRBSF Economist Emeritus. Gandhi (amitgandhi@gmail.com): Airbnb. Ruzic (dimitrije.ruzic@insead.edu): INSEAD. Traina (james.traina@nyu.edu): NYU Stern Abu Dhabi. We thank David Romer; seminar participants at INSEAD, the San Francisco Fed, and the Richmond Fed; and several excellent referees for helpful comments and feedback. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of San Francisco or the Board of Governors of the Federal Reserve System.

Macroeconomics is microeconomics plus aggregation.

Basu and Fernald (1997, p.264), paraphrasing Franklin Fisher

We review the "production approach" to estimating markups of price over marginal cost. Intuitively, markups are about pricing; pricing relates to marginal cost, and marginal cost reflects the production constraints facing firms. The increased availability of firm-level microdata and new econometric methods makes it easier to estimate production functions and infer markups. We survey what macroeconomists need to know about the microeconometric production approach. Our goal is to bridge micro and macro perspectives, highlighting challenges in moving from firm-level analyses to economy-wide conclusions.

Hall (1986, 1988, 1990) introduced the production approach. He showed that cost minimization implies the markup can be written as the ratio of the output elasticity of a flexible input to its revenue share (the share of that input's costs in revenue). Markups can be read directly from producer behavior. There is no need to specify a demand system or market structure. This makes the approach attractive relative to typical methods from industrial organization (IO) that are tailored to narrow industries. The production approach scales to large panels of firms across the economy and over time.

De Loecker and Warzynski (2012) showed how to apply the production approach to firm-level data using modern methods for estimating production functions. Interest in market power has since soared, in part because sharply rising markups might provide 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]

Modern macroeconomics rests on imperfect competition. In endogenous growth models, firms charge a markup 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 data, applied at scale, allow macroeconomists to incorporate rich heterogeneity, strengthening the microeconomic foundations of macro analysis.[2]

[1]References 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. (2006).

[2]For example, theory suggests that both high and heterogeneous markups can have significant welfare costs. Edmond et al. (2023) find welfare costs rise sharply as average markups rise. Baqaee and Farhi (2020) estimate that "eliminating the misallocation resulting from the large and dispersed markups...would raise aggregate TFP by about 15%."

However, the empirical evidence for rising markups is far from settled. Some production-approach specifications (e.g., De Loecker et al. 2020) show sharp increases in recent decades. Others do not (e.g., Foster et al. 2024; Traina 2018; Demirer 2025). Even if markups have risen, the production approach is silent on their structural drivers. Are they "bad markups," reflecting barriers to entry? Or "good markups," reflecting returns to innovation?

To put the empirical disagreements in context, our review emphasizes that the production approach to markups is an exercise in naming a residual. From a cost-minimizing first-order condition, the markup is the wedge between the output elasticity of a flexible input and its revenue share. It absorbs whatever the model, data, or econometrics fail to capture. We estimate the production function and infer market power as a residual. Differences in (mis)measurement and (mis)specifications across papers—even those using the same data—can therefore yield different residuals, different production markups.

Even if we resolve all the empirical disagreements, a key issue for macro is how to interpret the micro estimates. The implications of market power depend on its cause. The strength of the production approach is that it requires limited structure. That strength is also its weakness. We need the structure for welfare analysis. Modern IO methods rely on structural models tailored to competition and demand within narrowly defined markets. These models offer detailed insights into firm behavior and allow interpretation of equilibrium markups in a highly context-dependent way.

The production approach offers breadth: scalable estimates of markups across industries and firms over time. IO offers depth: detailed structural insights in narrow markets. Its narrow focus limits its usefulness to macroeconomists, who need economy-wide coverage. Macroeconomists must combine both perspectives, complementing markup estimates with institutional and structural knowledge, to build models of market power that match reality. We return to these issues at the end of the paper.

We begin this review by describing the production approach to markups in section 1. We highlight how few assumptions are needed—mainly a flexible input and cost minimization. The cost-minimizing FOC is quite simple: For a flexible input and a firm that takes input prices as given, the markup is its output elasticity relative to its revenue share.

Unfortunately, the devil is in the implementation. As we document in Section 2, the empirical evidence is mixed. The fundamental challenge is that implementing the elegant production approach requires substantive choices—each of which layers on different auxiliary assumptions, with limited guidance. The "garden of forking paths" in Section 2 documents the sensitivity across two (of multiple) sets of choices. One regards which input is taken to be

flexible—mainly labor or materials. Another is how much variation we allow across firms and over time in output elasticities. Several studies use the same data and estimation technique to show that these choices matter. The results are stark: Some choices imply sharply rising markups; others do not.

The remainder of the review explores conceptual (Section 3), data (Section 4), and econometric (Section 5) hypotheses for why empirical results with micro data are so sensitive. We also discuss directions forward. The underlying issue is that, because production markups are residuals, any misspecification or mismeasurement is swept into the estimated markup.

Section 3 discusses two conceptual hypotheses for the dissonant results of the garden. Either the key FOC is misspecified and needs an extra wedge for one or both inputs; or the production function is misspecified and does not allow sufficient flexibility in the estimated output elasticities. Most obviously, suppose we assume (as much of the literature does) that the production technology is Cobb-Douglas and pool firms at an aggregated level to estimate this production technology. Then, by construction, we do not allow variation in output elasticities. To satisfy the FOC, we impose that all observed variation in revenue shares reflects markups and none reflects non-markup frictions (e.g., adjustment costs, input markdowns, or regulatory barriers) or variable output elasticities. This tension between markups and technology is a recurring theme throughout our review.

Having theoretical hypotheses does not prove they fully explain the dissonant results. In principle, we could also have the wrong revenue share in the numerator. Or, even if we have the correct production function, estimation bias might be more severe for one input than another. Section 4 discusses data constraints. Data rarely match the theoretical objects assumed by the production approach. Inputs and outputs are measured with error—and sometimes, as with many intangibles, not measured at all. Measurement shortcomings can distort revenue shares or bias econometric estimation of output elasticities.

Section 5 discusses econometric pitfalls and how research has tried to address them. Inputs are endogenous to unobserved technology shocks, and we usually lack firm-specific output and input prices. Considerable progress has been made in recent decades, e.g., control functions and dynamic panel methods. Most of these new methods were originally developed assuming perfect data—including firm-specific output prices—and perfect competition. Progress has been made at relaxing these restrictions, but not yet enough. Each method introduces its own assumptions and fragilities. The econometrician is trying to recover primitives from behavior jointly shaped by technology, demand, and frictions. This simultaneity means even the best estimator rests on structural choices that are not innocuous.

In each of Sections 3 to 5 we offer both practical guidance and avenues for future research. Section 6 then concludes with three calls to arms. All three highlight the tension between market power and technology in the production approach to markups. First, we call for transparency and systematic reporting of how much revenue-share variation is explained by markups versus output elasticities. Second, we call for validation of production-based markups using 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.

In sum, the production approach is both powerful and fragile. Its potential is unmatched for estimating markups at scale. With continued refinement, it can become an even more reliable tool for understanding firms, markets, and the aggregate economy.

1 The production approach to markup estimation

Production functions and markups are inextricably linked through factor demand. Firms with market power restrict input use to lower output and raise prices. We first present the key cost-minimizing first-order condition (FOC) for optimal input use, then describe the role this FOC plays in growth accounting, where the production approach to markup estimation originated. Finally, we discuss how the FOC again takes center stage in the new literature on production-based microdata markups.

1.1 The unifying first-order condition

Consider the following production function F for firm i in time t:

$$Y_{it} = F(K_{it}, L_{it}, M_{it}, A_{it}). \tag{1}$$

Y_{it} is gross output; for a multi-product firm, it is an index over its various outputs. We separate inputs into capital, K_{it}, labor, L_{it}, and intermediate inputs, M_{it}. Each input is an index over heterogeneous types of labor, capital, or intermediates.[3] A_{it} denotes productivity.

Suppose some $X_{it} \in \{K_{it}, L_{it}, M_{it}\}$ is fully flexible. It is chosen in the current period and affects only current-period costs, unlike dynamic inputs with adjustment costs. The firm takes the input price W_{it}^X as given, ruling out input-market power (e.g., monopsony à la

[3]The separability assumptions in (1) impose that elasticities of substitution between elements of K_{it} and L_{it} are equal, which rules out possibilities like computers complementing high-skilled workers but substituting for low-skilled workers or energy complementing capital but substituting for labor (Berndt and Wood, 1975).

Robinson 1933). The firm minimizes costs:

$$\begin{aligned} \min_{X_{it}} \quad & W_{it}^X X_{it} \\ \text{s.t.} \quad & Y_{it} = F(K_{it}, L_{it}, M_{it}, A_{it}). \end{aligned} \tag{2}$$

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

$$W_{it}^X = \lambda_{it} \frac{\partial F(K_{it}, L_{it}, M_{it}, A_{it})}{\partial X_{it}}. \tag{3}$$

The Lagrange multiplier λ_{it} is the cost of relaxing the production constraint by one unit, i.e., marginal cost. This FOC links a factor's price to its marginal product and the firm's marginal cost. To interpret this condition, we define three terms:

- $\mu_{it} \equiv \frac{P_{it}}{\lambda_{it}}$ is the firm's markup of price over marginal cost,
- $\gamma_{it}^X \equiv \frac{\partial F}{\partial X_{it}} \frac{X_{it}}{Y_{it}}$ is the elasticity of output with respect to input X_{it}, and
- $s_{it}^X \equiv \frac{W_{it}^X X_{it}}{P_{it} Y_{it}}$ is the share of the factor's costs in revenue.

With these definitions, the FOC implies the flexible factor's output elasticity, γ_{it}^X, equals a markup μ_{it} over the factor's revenue share s_{it}^X:

$$\gamma_{it}^X = \mu_{it} s_{it}^X. \tag{4}$$

FOC (4) is the key equation in the production approach to markups. We next trace out this approach's origins in growth accounting and then return to its implementation in microdata.

1.2 Growth accounting and the production approach

In a series of papers, Bob Hall pioneered the production approach to measuring markups in time-series data (Hall, 1986, 1988, 1990). Growth accounting was one of the first applications of the approach. Markups affect the interpretation of common productivity measures used in both macro and micro.

Following Solow (1957), we differentiate the production function (1) logarithmically to quantify the contributions of inputs versus productivity to growth. Define $j \equiv \log J$ as variable

J in log-levels, so $\Delta j \equiv \Delta \log J$ is its log-growth rate. The change in output is:[4]

$$\Delta y_{it} = \gamma_{it}^K \Delta k_{it} + \gamma_{it}^L \Delta l_{it} + \gamma_{it}^M \Delta m_{it} + \Delta a_{it}. \tag{5}$$

Hall's insight was to substitute the FOC (4) into (5)—without imposing that $\mu = 1$:

$$\begin{aligned} \Delta y_{it} &= \mu_{it} \left(s_{it}^K \Delta k_{it} + s_{it}^L \Delta l_{it} + s_{it}^M \Delta m_{it} \right) + \Delta a_{it} \\ &\equiv \mu_{it} \Delta x_{it} + \Delta a_{it}. \end{aligned} \tag{6}$$

In this "Hall equation," $\Delta x_{it} \equiv s_{it}^K \Delta k_{it} + s_{it}^L \Delta l_{it} + s_{it}^M \Delta m_{it}$ is revenue-share-weighted input growth. Classic growth accounting (Solow, 1957; Jorgenson and Griliches, 1967) imposes $\mu_{it} = 1$. Equation (6) then defines the log-growth rate of TFP, Δtfp, as an index number:

$$\Delta tfp_{it} = \Delta y_{it} - \Delta x_{it} = \Delta a_{it}. \tag{7}$$

No estimation is required. If μ_{it} is indeed one, and if measured factor shares and quantity growth rates are correct, then this TFP residual equals true productivity growth, Δa_{it}.

With markups, however, Δtfp does not fully account for the productive contribution of inputs. Markups create a wedge between output elasticities and factor shares. Some of the productive contribution of input changes bleeds into the TFP residual:

$$\Delta tfp_{it} = (\mu_{it} - 1)\Delta x_{it} + \Delta a_{it}. \tag{8}$$

Hall argued this additional term could explain why aggregate TFP is procyclical, rising in booms and falling in recessions.

This work also highlights how economic profits relate to markups and returns to scale. The scale elasticity is $\gamma_{it} \equiv \gamma_{it}^K + \gamma_{it}^L + \gamma_{it}^M$: If all inputs change by some proportion d, then output changes by $\gamma_{it} \cdot d$. Constant returns means $\gamma_{it} = 1$. Suppose FOC (4) holds for all inputs.[5]

[4] We have normalized the elasticity of output with respect to technology, $\frac{\partial F(K_{it}, L_{it}, M_{it}, A_{it})}{\partial A_{it}} \frac{A}{F}$, to one. Any variation in this elasticity from non-factor-neutral technology will be subsumed into Δa_{it}. Solow operated in continuous time; using log-changes is the discrete-time approximation. The differentiation provides a local approximation to a general production function, so the output elasticities γ_{it}^X in general vary over time for all inputs X. With Cobb-Douglas, (5) is exact and $\gamma_{it}^X = \gamma_i^X$ are time-invariant.

[5] The FOC might need to be modified for some inputs, such as those with adjustment costs or other frictions. As Section 3.2 discusses, the correct "shadow price" of input X might be $W_{it}^X(1 + \tau_{it}^X)$. The full shadow price then needs to be included in the definition of economic costs below, but the logic remains.

Economic profits arise from higher markups μ_{it} or lower scale elasticities γ_{it}:

$$\begin{aligned}\gamma_{it} &= \mu_{it}\left(s_{it}^{K} + s_{it}^{L} + s_{it}^{M}\right) \\ &= \mu_{it}\left(\frac{\text{Cost}_{it}}{\text{Revenue}}\right) \\ \gamma_{it} &= \mu_{it}(1 - s_{\Pi_{it}}),\end{aligned} \tag{9}$$

where $s_{\Pi_{it}} \equiv (P_{it}Y_{it} - \text{Cost}_{it})/P_{it}Y_{it}$ is the rate of pure economic profit and $\text{Cost}_{it} \equiv W_{it}^{K}K_{it} + W_{it}^{L}L_{it} + W_{it}^{M}M_{it}$ is total economic cost—the sum of all actual and shadow expenditures on inputs.[6] Economic profits therefore arise when markups exceed the scale elasticity ($\mu_{it} > \gamma_{it}$). Whether profits result from markups or from scale elasticities is important in macro models because, as Basu (2019) notes, welfare depends on markups not profits.

Given this close relationship, knowing either markups or scale elasticities is sufficient to disentangle the relationship between growth in output, inputs, and productivity. The Hall equation (6) shows the relationship using markups μ_{it} and revenue-weighted input growth Δx_{it}. To see the relationship with returns to scale, we combine the FOC (4) and the middle equality in (9) to note an equivalence between an input's markup-adjusted revenue share $\mu_{it}s_{it}^{X}$ and its scale-adjusted cost share $\gamma_{it}c_{it}^{X}$, where $c_{it}^{X} \equiv W_{it}^{X}X_{it}/\text{Cost}_{it}$:

$$\gamma_{it}^{X} = \mu_{it}s_{it}^{X} = \gamma_{it}c_{it}^{X}. \tag{10}$$

Substituting this equality into equation (6), we see output growth Δy_{it} linked to cost-share-weighted input growth Δx_{it}^{cost} using the scale elasticity γ_{it}:

$$\Delta y_{it} = \gamma_{it}\Delta x_{it}^{\text{cost}} + \Delta a_{it}. \tag{11}$$

For growth accounting, we can measure productivity growth Δa_{it} with a markup μ_{it} and equation (6) or with a scale elasticity γ_{it} and equation (11). If we estimate either μ_{it} or γ_{it}, we can infer the other as the residual, ensuring that equation (9) holds.

Hall introduced this production approach to markups by estimating equation (6)—or, equivalently, equation (11)—in industry time-series data. Input growth is presumably endogenous

[6]Noting that $\mu_{it} = P_{it}/\lambda_{it}$, where λ_{it} is marginal cost, we can use the middle equality in (9) to show that γ_{it} is the ratio of average to marginal cost. Increasing returns can take multiple forms. For example, suppose labor is the only input and that $Y_{it} = (L_{it} - \bar{L}_{it})^{\delta_{it}^{L}} - \Phi_{it}$. Φ_{it} is a fixed cost of production defined in units of output and $\bar{L}_{it}$ is overhead labor required for any level of production. Then $\gamma_{it} = \gamma_{it}^{L} = \delta_{it}^{L}\left(1 + \frac{\Phi_{it}}{Y}\right)\left(1 + \frac{\bar{L_{it}}}{L_{it} - \bar{L_{it}}}\right)$. Increasing returns can arise from $\delta_{it}^{L} > 1$ (decreasing marginal cost); from overhead labor, $\bar{L}_{it} > 0$; or if there are general fixed costs of production ($\Phi_{it} > 0$). The macro implications of decreasing marginal cost can differ from those of fixed costs or overhead factors.

to productivity change (the "transmission problem" of Marschak and Andrews, 1944). Hall and the large literature that followed proposed using aggregate demand instruments, such as military spending or identified monetary-policy innovations. These demand-side instrumental variables are correlated with input movements across many industries but should not be correlated with an industry's productivity shock Δa_{it}—making them valid instruments. Subsequent work followed the same production approach in industry data, sometimes with additional controls (e.g., for factor utilization or effects that are external to an industry).[7]

In his initial work, Hall (1986, 1988, 1990) found evidence of extremely large markups (typically in the range of two to four) using industry value-added data. Subsequent papers with industry data generally found much smaller evidence of widespread markups using gross-output data (even when converted to a value-added basis). For example, Basu and Fernald (1997) estimate that the typical firm has close to constant returns to scale. Given their estimate that the rate of economic profit s^{Π} averaged about 3 percent in data from 1959 to 1989, their estimates imply that markups were typically only modestly above one.

A challenge with industry results is that industries are not decision-makers for whom the cost-minimization problem (2) naturally applies. There can be reallocation effects within industries (Basu and Fernald, 1997). Nevertheless, the Hall approach laid the groundwork for the new literature on production-based markups in microdata, to which we turn next.

1.3 The new literature on production-based markups

The production approach to markup estimation has expanded rapidly with the increasing availability of firm microdata.[8] The first-differenced, time-series approach of equations (6) or (11) is poorly suited to these data (De Loecker, 2011b). First-differencing the production function exacerbates measurement error in panel data (Griliches and Mairesse, 1998). Aggregate-demand instruments that work with industry time series have low power in short panels and limited (if any) firm-level cross-sectional variation.

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. Conceptually, FOC (4) shows that markup estimation requires three steps:

- correctly selecting a flexible input X_{it},
- measuring its revenue share s_{it}^X, and
- specifying a production function and estimating the output elasticity γ_{it}^X.

[7] E.g., Caballero and Lyons (1992); Burnside et al. (1995); Basu and Fernald (1995), Basu et al. (2006).
[8] Klette (1999) was an early application of the Hall approach to firm-level data.

For flexible input X_{it}, the markup μ_{it} can be inferred as variation in a firm's revenue share for that input, s_{it}^X, that is not explained by the output elasticity γ_{it}^X: $\mu_{it} = \frac{\gamma_{it}^X}{s_{it}^X}$.

The first two steps are related. We need a fully flexible input so that the input's revenue share can potentially be observed in the data. Revenue shares for quasi-fixed inputs depend on (unobservable) shadow prices, which makes it difficult to observe the relevant input cost.

For the third step, though we need the output elasticity for only one flexible input, estimation generally requires all inputs to avoid omitted-variable bias. For example, with a time-invariant Cobb-Douglas production function in labor, materials, and capital:

$$y_{it} = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it} + a_{it}. \tag{12}$$

Omitting any input in estimation—even a quasi-fixed one like capital—risks biasing the estimated output elasticity (γ^L or γ^M) that we want to use to measure markups.

The simplicity of the three-step production approach makes it conceptually appealing relative to demand-side alternatives that can require extensive customization for each product or industry. For example, many IO studies have followed the approach of Berry et al. (1995). That paper focuses on a narrow product category, automobiles. It explicitly models consumer demand for different product attributes, as well as the nature of strategic interactions among firms and firms' profit-maximization problems. This approach requires detailed product-level data as well as a large number of structural assumptions regarding consumer demand, strategic interaction, and the nature of profit maximization.

The production approach relies on cost minimization, making it broadly applicable across products and industries. Cost minimization is a general principle: regardless of how complex or dynamic the profit-maximization problem may be, or even if firms are not strictly maximizing profits, they typically aim to produce at the lowest possible cost. For example, firms may face constraints on adjusting prices each period or operate in complex competitive environments, yet still choose the least-cost method to produce what they sell.

At the same time, the production approach to markups—like TFP measurement—is about naming a residual. The markup is the residual that ensures FOC (4) holds given a revenue share s_{it}^X and an estimated output elasticity γ_{it}^X. Any misspecification—of the FOC, the production function, or the econometrics of estimating the production function—or mismeasurement of the factor shares gets pushed into the residual markup. We return regularly to this concern about mismeasurement and misspecification. As we highlight, each choice in implementing the FOC involves additional sets of assumptions.

2 The garden of forking paths

Estimating markups from the production approach appears straightforward at first glance. The central first-order condition, equation (4), offers a simple prescription: For a flexible input, the markup equals the ratio of its output elasticity to its revenue share. Translating this equation into empirical practice requires a series of decisions. A body of evidence shows that these decisions can substantially influence the results. 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—two decisions that every researcher must make when applying the production approach: (1) which input to use, and (2) how flexibly to specify technology. These forks are not exhaustive but they are illustrative: Each has been studied empirically using variants of the same data, and each has been shown to materially affect the estimated markups.

The primary goal of this section is to provide a high-level overview of these decision points, using existing studies that offer direct comparisons. We focus on cases where researchers apply multiple choices within the same empirical setting, allowing clean contrasts. For the most part, we defer interpretation of discrepancies until later sections. A secondary goal is to highlight uncertainty about estimates of the level and trend in production-based markups. It is easy to find plausible specifications in which markups are relatively constant.

The rest of the review then digs deeper into the sensitivity to these two decisions and examines additional decisions and forks for which the literature has not yet provided evidence.

2.1 First fork: Which input to use for the production approach?

Having chosen the production approach, the researcher faces a deceptively simple question: Which input should be used to infer the markup?

In principle, any fully flexible input should yield the same markup. In practice, markup estimates based on different inputs often diverge in both levels and trends. We highlight this first fork because, in many applications of the production approach, the input choice is unexamined: An input is simply asserted as the relevant flexible input.

Table 1a presents results from three papers that compare markup estimates from different inputs using the same dataset. These choices lead to different conclusions about market power and motivate our subsequent deeper dive into theory, measurement, and estimation.

Table 1a: Garden of Forking Paths

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 (1978–1996)	Decrease of 20%	Increase of 15%
Colombia (1978–1996)	Decrease of 30%	Increase of 10%
India (1998–2014)	Decrease of 40%	No change
Indonesia (1991–2000)	Decrease of 10%	Increase of 5%
Doraszelski & Jaumandreu (2019)		
	Labor	Materials
Spain (1990-2012)	Exporters charge higher markups	Exporters charge smaller markups
Raval (2023)	Energy	Non-Energy (Raw) Materials
Chile (1978–1996)	Increase of 20%	Increase of 15%
Colombia (1978–1996)	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	Increase: 1.15 to 1.17

Sources: Figures 2 and 5 from Raval (2023), Table 1 from Doraszelski & Jaumandreu (2019), and Figure 2 from Traina (2018). Percentages rounded to nearest 5%.

Raval (2023) uses manufacturing census data from five countries to show that labor- and materials-based markup estimates are not just noisy proxies for the same object—they imply opposite trends. For example, U.S. labor-based markups nearly double between 1970 and 2010, while materials-based markups fall in half. In Chile and Colombia, labor-based markups fall while materials-based markups rise. Within all five countries, labor and materials markups are negatively correlated across firms, both in levels and trends. These patterns persist across various specifications and estimation methods.

This choice of input also matters for purely cross-sectional comparisons: Doraszelski and Jaumandreu (2019) compare exporters and non-exporters in Spain, and find that conclusions about which group charges higher markups depend on the input used in the production

approach. Their headline result is striking: exporters appear to charge higher markups when the production approach is applied to labor, but appear to charge lower markups if the production approach is applied to materials.

Even when materials are split into subsets, markup trends differ sharply by subset. As Table 1a shows with additional results from Raval (2023), markup estimates based on energy versus non-energy (raw) materials yield divergent conclusions. In Chile both series rise, but in Colombia estimated markups based on energy fall 70% while markups based on non-energy materials rise 20%. In India and Indonesia, the markup trends switch signs. As it is not obvious how to rank the relative flexibility of energy and non-energy raw materials, these patterns are important to note because input flexibility is often assumed, not demonstrated. For instance, implementations of the production approach often use materials—rather than, say, labor—on the asserted grounds that materials are more flexible. Hence it is important to note that even within materials, markup estimates can diverge across subsets.

While the preceding studies leverage detailed census-type manufacturing data that cleanly separate inputs into categories like capital, labor, and materials, many applied papers estimating production-based markups rely on financial statement data, which offer less granular classifications. A widely used source is Compustat, which reports standardized accounting data for publicly traded U.S. firms. Within Compustat, inputs are not classified by production function categories but by accounting definitions. The most relevant measures are Cost of Goods Sold (COGS) and Selling, General, and Administrative Expenses (SGA), which are sometimes combined into Operating Expenses (OPEX). COGS generally reflects expenditures directly tied to production; SGA captures administrative expenses such as advertising, management salaries, and office costs. OPEX aggregates both.

The final row of Table 1a highlights the sensitivity of markup trends to input definitions in financial data. Traina (2018) finds that omitting SGA expenses—and thus focusing only on COGS—leads to higher markup estimates. Using Compustat data from 1950 to 2016, the markup series based solely on COGS implies a substantial rise in markups, from 1.19 in 1950 to 1.45 in 2016. The series based on OPEX suggests that markups were lower and nearly flat, rising from 1.15 to 1.17. Including SGA wipes out most of the rise.

All the papers referenced in this section reinforce the same lesson: the choice of inputs shapes substantive conclusions about markups. While the logic of the production approach suggests that any flexible input should yield the same markup, different inputs—whether labor, materials, energy, or accounting aggregates—often imply divergent trends. In practice, input flexibility may vary across sectors, time horizons, or even firm size. Nearly any input is

likely subject to some frictions, such as adjustment costs, long-term contracts, or institutional constraints. Before formalizing how we might think about these concerns, we highlight one more important choice in the garden of forking paths.

2.2 Second fork: How flexibly to model output elasticities?

Having chosen the production approach and a flexible input, the researcher confronts a second fork: how to estimate that input's output elasticity? The choice of a production function and the granularity of estimation determines how much output elasticities can vary across firms and over time. For example, the Cobb-Douglas production function (12) restricts output elasticities $\widehat{\gamma}^X$ to be common across firms and time. If this specification is too restrictive, genuine variation in output elasticities will be forced into estimated markups.

Given this concern, researchers often allow more flexibility in the production technology. One approach is to let Cobb-Douglas elasticities vary across time or more granular firm groups (e.g., narrower industries). Another is to move beyond Cobb-Douglas. A CES production function allows for non-unitary substitution between inputs. Even more general, a translog is a second-order approximation to any arbitrary production technology.[9] Researchers can also relax assumptions on productivity. For instance, one generalization is to allow productivity to be a vector, $A_{it} = \left(A^H_{it}, A^L_{it}\right)$, where A^H_{it} is Hicks-neutral and A^L_{it} is labor augmenting productivity. The general production function becomes $Y_{it} = A^H_{it} F(K_{it}, A^L_{it} L_{it}, M_{it})$. A CES implementation would parametrize $F(\cdot)$ as a CES function so that

$$Y_{it} = A^H_{it} \left[(1 - \alpha_l - \alpha_m) K_{it}^{\frac{\sigma-1}{\sigma}} + \alpha_l \left(A^L_{it} L_{it} \right)^{\frac{\sigma-1}{\sigma}} + \alpha_m M_{it}^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}. \tag{13}$$

With this technology, output elasticities depend on input intensity as well as A^L_{it}. We defer a more complete discussion of flexible production functions to Section 3.2.

A researcher at this fork must specify a production technology and also decide how to estimate it, which we discuss in Section 5. We can avoid estimation if the production function has constant returns to scale, ($\gamma_{it} = 1$). Then output elasticities are given by cost shares: $\gamma^X_{it} = c^X_{it}$ (as discussed around equation (11)). If ($\gamma_{it} \neq 1$), however, then variation in returns to scale will be misattributed to changes in markups, an artifact of the estimation method rather than a reflection of actual market power.

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

[9] With two inputs, the baseline translog with Hicks-Neutral technology is $y_{it} = \alpha_0 + \alpha_K k_{it} + \alpha_L l_{it} + \frac{1}{2}\beta_{KK}(k_{it})^2 + \beta_{KL} k_{it} l_{it} + \frac{1}{2}\beta_{LL}(l_{it})^2 + a_{it}$, so it includes squared and cross terms. Here, input intensity informs variation in production technology. For instance, the output elasticity for labor is $\gamma^L_{it} = \alpha_L + \beta_{KL} k_{it} + \beta_{LL} l_{it}$.

central challenge: the output elasticity is not observed but inferred. Because markups are the residual that makes the FOC hold, any slip in measuring elasticities or specifying technology shows up one-for-one in the markup.

Table 1b: Garden of Forking Paths

Second Fork: how to estimate the output elasticity for a given input?

De Loecker, Eeckhout & Unger (2024)			
Compustat, 1955-2016			
	Cost Share	Cobb-Douglas	Overhead
	Increase: 1.35 to 1.75	Increase: 1.25 to 1.61	Increase: 1 to 1.32
Foster, Haltiwanger & Tuttle (2024)			
U.S. Manuf., 1977-2012			
	Cost Share	Cobb-Douglas	Translog
4-digit industry	Increase: 1.25 to 1.5	Stable: 1.25 to 1.25	Decrease: 1 to 0.99
2-digit industry	Increase: 1.4 to 1.8	Increase: 1.35 to 1.48	Increase: 1.3 to 1.6
Demirer (2025)		Cobb-Douglas	Labor-augmenting
US (1961–2018)		Increase: 1.30 to 1.50	Increase: 1.25 to 1.30
Chile (1979–1996)		Increase: 1.35 to 1.40	Decrease: 1.26 to 1.22
Colombia (1978–1991)		Stable: 1.40 to 1.40	Decrease: 1.30 to 1.28
India (1998–2014)		Stable: 1.31 to 1.31	Increase: 1.20 to 1.29
Turkey (1983–2000)		Increase: 1.25 to 1.31	Decrease: 1.20 to 1.10

Sources: Figures 7A, 1 and 8A from De Loecker, Eeckhout & Unger (2024), Table 2 from Foster, Haltiwanger & Tuttle (2024), and Figures 6 and OA-5 from Demirer (2025). De Loecker, Eeckhout & Unger (2024) use cost of goods sold as the flexible input. Foster, Haltiwanger & Tuttle (2024) use materials. Demirer (2025) uses the combination of labor and materials.

De Loecker et al. (2020) find that markups in the U.S. corporate sector rose from 1955 to 2016 under a range of approaches. They focus on COGS as the variable input; their benchmark specification assumes a Cobb-Douglas production function with elasticities that vary across two-digit NAICS industries and over time.[10] Under this specification, estimated markups increase from 1.25 in 1980 to 1.61 in 2016. When the output elasticity is held fixed at its historical cost-share average (the first column; the cost-share average is 0.85), the resulting markup series follows a similar trajectory—suggesting that rising markups are not merely an artifact of the elasticity rising over time. Even when they extend the production function

[10]For example, manufacturing is divided into three groups, NAICS 31, 32, and 33, so that, say, steel refineries and auto plants are assumed to share the same Cobb-Douglas technology.

to include a fixed (overhead) input, the upward trend in markups remains.[11]

In contrast to the consistent markup rise in De Loecker et al. (2020), Foster et al. (2024) find that alternative methods of estimating output elasticities can flip the story. Using U.S. Census of Manufacturing plant data (1977-2012), with materials as the flexible input, they compare markup trends across two dimensions in Table 1b: (i) estimation methods (cost-share proxy, Cobb-Douglas, and translog); and (ii) aggregation (coarser 2-digit industries, as in Compustat, versus finer 4-digit industries). With cost-share proxies, markup levels and trends are stable across both aggregation levels. Once elasticities are estimated, robustness vanishes. Under Cobb-Douglas, 2-digit industries show markups rising ($1.35 \rightarrow 1.48$), while 4-digit industries show no change (steady at 1.25). Under translog, the split is even sharper: a sizeable markup rise ($1.3 \rightarrow 1.6$) at the 2-digit level versus stability ($1.0 \rightarrow 0.99$) at the 4-digit level. Since the underlying plant-level revenue shares are identical, the divergences reflect how estimation methods reassign trends between technology (output elasticities) and demand (markups).

Demirer (2025) also highlights the role of technology versus demand. He finds that markup trends can reverse direction depending on whether production is modeled with Hicks-neutral or labor-augmenting technology. Across countries, Cobb-Douglas with Hicks-neutral technology consistently produces higher markup levels and stronger upward trends than the labor-augmenting alternative. Specifically, in U.S. manufacturing, Cobb-Douglas implies a markup rise from 1.30 to 1.50 between 1961 and 2018, whereas labor augmenting implies a modest increase from 1.25 to 1.30. In Chile and Turkey, the trends flip—rising under Cobb-Douglas but falling under labor-augmenting CES. These results echo findings from other studies in this section: subtle changes to functional form or elasticity assumptions can have first-order consequences for inferred market power.

Taken together, divergent trends across methods mean that apparently rising markups could signal changing technology or misspecification, not market power. The divergence highlights a central tension at the heart of the production approach: output elasticities are not directly observed but must be estimated, and markups are residuals pinned down by the FOC. Restrictive functional forms, coarse aggregation, or ignoring factor-augmenting technology all translate directly into biased markup estimates. The divergence in trends across elasticity estimation strategies is not just a technical nuisance; it fundamentally complicates the interpretation of rising markups.

[11] This specification, which they call PF2, includes SGA as a productive input along with COGS and capital. It is similar to the Traina (2018) OPEX specification from Table 1a, except that they do not combine COGS and SGA. Unlike Traina, they still find a sharply rising markup.

3 Conceptual rationales for the garden

All the studies reviewed in the previous section reinforce the same central lesson: necessary implementation choices—about inputs, production functions, and estimation—can decisively shape substantive conclusions about markups. These often divergent markups are rooted in conceptual tensions that remain unresolved in the empirical literature.

This section highlights two theoretical hypotheses that can help reconcile the discrepancies from the garden of forking paths: (1) the central first-order condition could be missing non-markup frictions and (2) the estimated production technology could be insufficiently flexible. These two hypotheses can, in principle, generate the empirical divergences, and they both have support in the literature. They are not the only hypotheses, since there could also be sources of bias arising from inadequacies in the data (Section 4) or from econometric (not conceptual) issues in estimating output elasticities (Section 5).

3.1 Do we have the right first-order condition?

The first conceptual hypothesis directly addresses the garden's puzzle that different inputs can imply markedly different markup levels and trends. The key FOC assumes a static cost-minimization problem in which the firm takes input prices as given. If these assumptions fail, an unobserved wedge τ_{it}^X emerges between the output elasticity γ_{it}^X and the observed revenue share s_{it}^X:[12]

$$\gamma_{it}^X = \mu_{it} s_{it}^X \left(1 + \tau_{it}^X\right). \tag{14}$$

Suppose we correctly estimate γ_{it}^X and observe s_{it}^X. The production approach would then measure the markup $\widehat{\mu}_{it}^X$ as the residual, $\mu_{it}\left(1 + \tau_{it}^X\right)$, that makes the FOC (14) hold. This inference would conflate the true markup μ_{it} with the input-specific non-markup wedge τ_{it}^X. Different inputs might have different (unobserved) non-markup wedges and therefore imply different levels and trends for measured markups $\widehat{\mu}_{it}^X$, as reported in Table 1a.

These non-markup, input-specific wedges τ_{it}^X are often interesting in their own right. They can arise for at least three reasons. First, even if cost minimization is static, the observed market price W_{it}^X may not be the allocative price $(1 + \tau_{it}^X)W_{it}^X$ that the firm responds to when minimizing costs. Second, the cost-minimization problem might be dynamic, with τ_{it}^X capturing the misspecification from imposing a static framework. Third, τ_{it}^X may reflect other forces—such as bargaining or search—that break the tight link between input prices and input demand in the FOC.

[12] We follow Doraszelski and Jaumandreu (2019) in writing all deviations from FOC (4) in this form.

First, one important case arises when observed input costs s_{it}^X are not the allocative costs $(1+\tau_{it}^X)s_{it}^X$ that firms consider in optimization. A large literature, following Restuccia and Rogerson (2008) and Hsieh and Klenow (2009), argues that frictions—often government induced—distort input allocation and reduce aggregate productivity. Some such frictions are reflected in input prices (e.g., cross-firm differences in tariffs or taxes) and are thus embedded in s_{it}^X, leaving production-based markups unbiased. Others, such as quotas or regulatory restrictions on firm size, limit input use without being priced into s_{it}^X; in these cases, the allocative cost that rationalizes why the firm uses too little of an input includes an unobserved shadow tax τ_{it}^X. When such wedges differ across inputs, measured markups $\widehat{\mu}_{it}^X$ will diverge in levels and (possibly in) trends, as in Table 1a.

A specific form of static, non-markup wedge comes from market power in input markets (Robinson, 1933). If the firm faces an upward-sloping input supply curve, hiring more raises the wage it must pay. Defining $\epsilon_{it}^X \equiv \frac{\partial W_{it}^X(X_{it})}{\partial X_{it}} \frac{X_{it}}{W_{it}^X}$ as the elasticity of the wage with respect to employment, input-market power ($\epsilon_{it}^X > 0$) means the firm acts "as if" the wage were $W_{it}^X(1+\epsilon_{it}^X)$. The result is a "markdown" of the observed wage relative to the marginal revenue product of the input. Following Dobbelaere and Mairesse (2013), several papers compare the residuals $\widehat{\mu}_{it}^X$ across inputs to quantify input-market power.[13]

The misallocation literature reinforces that one FOC cannot separately identify two unknowns. In that literature, FOC (14) is often used to measure unobserved distortions rather than production-based markups. Following Hsieh and Klenow (2009), researchers typically assume that output elasticities γ_t^X and markups μ_t are common across groups of firms. Any cross-firm differences in revenue shares s_{it}^X are then attributed to differences in firm-specific distortions $(1+\tau_{it}^X)$. Without additional data or modeling, one cannot say how much of a wedge comes from heterogeneous markups versus other frictions. Since heterogeneous markups themselves cause misallocation, they could be folded into the broader category of distortions driving wedges in marginal products.[14] Even so, while a markup is one type of distortion, not all distortions are markups.

A second possibility is that the cost-minimization problem is inherently dynamic, yet the researcher imposes a static framework. One dynamic problem is costly adjustment of inputs: today's choices—such as investment in a fixed factor—affect future production costs. As Basu and Fernald (2002) and Doraszelski and Jaumandreu (2019) note, the resulting "shadow"

[13]Studies of labor-market power have a long history (Manning, 2003, e.g.,). Recent examples include Dobbelaere and Mairesse (2013); Berger et al. (2022); Yeh et al. (2022); Jarosch et al. (2024); Kirov and Traina (2022); Rubens (2023).

[14]Restuccia and Rogerson (2017) survey the misallocation literature. Peters (2020); Edmond et al. (2023); Baqaee and Farhi (2020); Baqaee et al. (2024) discuss misallocation arising from markups.

input price can still be expressed as $W_{it}^X(1+\tau_{it}^X)$, where W_{it}^X is the frictionless user cost of the fixed input. In general, this wedge τ_{it}^X depends on stocks, flows, and expectations. The resulting factor-demand equation retains the form of (14).[15]

Adjustment-cost wedges and misallocation wedges both produce (14), but they differ in efficiency implications: adjustment costs are not inherently inefficient (Asker et al., 2014). Capital is the classic example, but other inputs may face adjustment costs as well. Cooper et al. (2024) find that labor adjustment costs are important—and rising—in U.S. manufacturing. Ignoring these costs and treating labor as fully flexible yields "substantial and rising dispersion in production-based markups without any variation in actual markups" (p.21). Even materials and other intermediate inputs may face dynamic frictions, as in Liu and Tsyvinski (2024), with supply-chain disruptions and delivery lags offering further evidence (Dhyne et al., 2022; Acemoglu and Tahbaz-Salehi, 2025).

Another dynamic concern is that τ_{it}^X may reflect mismeasurement of the allocative wage rather than a friction *per se*. Most employment relationships are long-term. From an implicit-contracts perspective, Hall (1980) argues that wages "should be viewed as an installment payment on the firm's long-term obligation to the worker" (p.101). The allocative shadow cost may differ substantially from the observed installment payment (Basu and House, 2016; Kudlyak, 2024). Related issues arise in the literature on relational contracts in repeated firm-to-firm transactions.[16] Using observed revenue shares instead of allocative ones effectively introduces a time-varying τ_{it}^X.

Finally, the firm's optimization problem may not take the form of (2) at all, as with wage-bargaining and rent-sharing.[17] Suppose the wage is set to split the surplus from a match, where the worker's outside option is U_t. The total surplus equals the marginal revenue product, $\left(\frac{P_{it}}{\mu_{it}}\frac{\partial F_{it}}{\partial X_{it}}\right)$ minus the outside option. The worker's share is β_t so the wage is

$$W_{it}^X = U_t + \beta_t\left(\frac{P_i t}{\mu_{it}}\frac{\partial F_{it}}{\partial X_{it}} - U_t\right). \tag{15}$$

Rearranging with $\beta_t \neq 0$ yields

$$\gamma_{it}^X = \mu_{it} s_{it}^X\left(1+\tau_{it}^X\right), \text{ where } \left(1+\tau_{it}^X\right) \equiv \left[\frac{1}{\beta_t} - \left(\frac{1-\beta_t}{\beta_t}\right)\left(\frac{U}{W}\right)\right]. \tag{16}$$

[15]Berndt and Fuss (1986) and Hulten (1986) discuss adjustment costs in growth accounting. Basu et al. (2001, p.245) and Doraszelski and Jaumandreu (2019) consider them in the production approach to markups.

[16]Rosen (1985) surveys implicit contracts in labor markets. Macchiavello and Morjaria (2023) survey relational contracts between firms. One form of restriction is quantity constraints, which create an unobserved shadow price analogous to that in the misallocation literature.

[17]Rogerson et al. (2005) survey search-and-matching models of the labor market. These models are also inherently dynamic, with different specifications of the bargaining process.

Here, $(1 + \tau_{it}^X)$ is endogenous, but the equation still takes the form of (14). Changes in bargaining power or outside options will then appear as changes in the estimated markup.

This discussion of the FOC underscores the need to choose inputs that plausibly satisfy the production approach's assumptions. Deviations from frictionless input markets are likely widespread. Differences in estimated markups across inputs—such as in Tables 1a and 1b—may therefore reflect differences in unobserved wedges τ_{it}^X. That divergence is itself a signal of frictions or misspecification, not necessarily of markups. The policy implications differ: is the wedge a markup, a markdown, a regulation, or a misspecification? The answer matters for how we read the evidence—and for what we do about it.

3.2 Are we modeling production in a sufficiently flexible way?

A second conceptual hypothesis for the input discrepancies in the garden is that the assumed production function is too restrictive. Consider an estimated output elasticity $\gamma_{it}^X = \gamma^X$ that is constant across firms and over time.[18] Since the markup is inferred as a residual from the FOC, $\hat{\mu}_{it} = \frac{\hat{\gamma}^X}{s_{it}^X}$, the inferred markups $\hat{\mu}_{it}$ would capture 100% of both the cross-sectional and the time-series variation in the flexible factor's revenue share s_{it}^X. A falling labor share, for instance, would immediately imply higher markups because changes in production technology are ruled out by assumption.

A production function that is too restrictive can also help rationalize why different inputs might imply different levels and trends of markups, as in Table 1a. Suppose labor and materials are both fully flexible and satisfy the cost-minimizing condition (4). Let $\hat{\mu}_{it}^X$ denote the markup estimated using input X. For each firm i, the ratio of the estimated FOCs for labor to materials exactly satisfies:

$$\frac{s_{it}^L}{s_{it}^M} \frac{\hat{\mu}_{it}^L}{\hat{\mu}_{it}^M} = \frac{\hat{\gamma}_{it}^L}{\hat{\gamma}_{it}^M}. \tag{17}$$

With Cobb-Douglas estimated at the industry level, the right-hand side is constant across firms, so $\hat{\mu}_{it}^L/\hat{\mu}_{it}^M$ soaks up all variation in s_{it}^L/s_{it}^M. If firms truly share the same Cobb-Douglas function, that variation in $\hat{\mu}_{it}^L/\hat{\mu}_{it}^M$ is measurement error. If not, the restriction forces technology differences into the markup ratio.

We could permit more variability in output elasticities through two main approaches: estimating at a finer level of aggregation or using a more flexible functional form. Each offers different degrees of freedom for elasticities to explain variation in revenue shares.

[18] E.g., many comparisons in Tables 1a impose a common Cobb-Douglas function for all firms in an industry.

One simple fix is to estimate elasticities at finer industry levels. Even with a Cobb-Douglas form, allowing coefficients to vary across four-digit industries rather than pooling at two digits lets elasticities explain more of the observed variation in factor shares. Coarser estimates risk conflating markup and technology heterogeneity. Table 1b shows exactly this: at two digits markups rise, at four digits they don't—evidence that coarser elasticities force technology variation into measured markups (Foster et al., 2024).

Doraszelski and Jaumandreu (2019), Raval (2023), and Demirer (2025) go further, arguing that the resolution is to use a more flexible functional form and to relax Hicks-neutrality. Labor-augmenting technical progress can potentially resolve the markup discrepancies in both Tables 1a and 1b—even when inputs are fully flexible and do not face non-markup wedges τ_{it}^X. Suppose the production function is CES with labor-augmenting technical progress, as in (13). Then the ratio (17) becomes:

$$\frac{s_{it}^L}{s_{it}^M} \cdot \frac{\hat{\mu}_{it}^L}{\hat{\mu}_{it}^M} = \left(\frac{\alpha_l}{\alpha_m}\right) \cdot \left(\frac{L_{it}}{M_{it}}\right)^{\frac{\sigma-1}{\sigma}} \cdot (A_{it}^L)^{\frac{\sigma-1}{\sigma}} \tag{18}$$

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 across firms within a period (i.e., $A_{it}^L = A_t^L$), this extra degree of freedom can explain variation in relative factor costs instead of forcing it into markup differences. Demirer (2025) finds that allowing for labor-augmenting productivity reduces estimated markup levels and flattens their upward trend compared with Cobb-Douglas or Hicks-neutral CES/translog specifications.[19]

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 (e.g., labor-augmenting)?

Each can matter through the same mechanism: allowing output elasticities to vary more across firms. When elasticities vary, differences in revenue shares can be attributed to technology rather than to markups.

[19]The labor-augmenting view rests on difficult-to-test assumptions about the nature of productivity. Since at least Solow (1957) and Sato (1967), it has been recognized that technological bias is difficult to distinguish from differential patterns of substitution.

3.3 Takeaways from the garden of forking paths

The garden of forking paths illustrates that disparate markup trends need not be mere statistical noise—they can reflect genuine conceptual choices. The forks matter, and researchers face several options for navigating them.

First, the choice of input should be grounded in a persuasive case that the central FOC is likely to hold. This requires arguing from institutional detail rather than default conventions. In some contexts, 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 or seasonal workers, whose wages are more likely to reflect contemporaneous market conditions than long-term contracts (e.g., van Heuvelen et al. 2021). Making the case for flexibility is ultimately about convincing the reader that the observed input price is the allocative one to which the firm responds when minimizing costs.

Second, when plausible frictions—such as input market power, adjustment costs, or regulatory constraints—threaten the link between the first-order condition and markups, the researcher may need to add more structure to the estimation by modeling the source of the wedge.[20] One can use two inputs to estimate two frictions (e.g., markups and markdowns); or incorporate 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 so it is not mistaken for market power.

Ultimately, these three strategies—carefully justifying the chosen input, explicitly modeling relevant frictions, and modeling production more flexibly—are ways of navigating the garden's forks with transparency and discipline.

[20] E.g., 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.

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

4 The data constraint

Implementing the production approach runs into two practical problems beyond the conceptual forks discussed in Sections 2 and 3: data limitations (this section) and econometric identification (Section 5). We need good data and sound econometrics to estimate production functions and output elasticities. Estimating production functions in micro data has always been hard (Griliches and Mairesse, 1998)—and it still is, even with new datasets and techniques. These difficulties may help explain why markup estimates diverge across studies.

The production approach requires two ingredients: a revenue share s_{it}^X and an output elasticity γ_{it}^X. The revenue share might seem straightforward but isn't always. For the FOC to hold, data should be at the level where firms optimize—establishments or firms. Microdata have long fallen short of theoretical ideals (Grunfeld and Griliches, 1960). Researchers face a tradeoff: sound microfoundations or high-quality measurement—they can rarely have both.

We start with simple simulations of how bad data can lead to biased markup estimates—through mismeasured revenue shares, cost shares, or output elasticities. We then draw lessons from both the Jorgenson-Griliches tradition of aggregate data and the industrial organization experience with firm-level analysis. Finally, we examine specific challenges in modern microdata when it comes to markup estimation.

4.1 The quantitative importance of measurement

Why do data problems matter for markup estimation? The production markup $\hat{\mu}_{it}$ is the ratio of an estimated output elasticity $\hat{\gamma}_{it}^X$ to a measured revenue share s_{it}^X. This ratio structure means measurement errors can distort markups in predictable ways. Consider two fundamental measurement challenges: systematic undermeasurement of inputs and classical measurement error.

First, micro datasets consistently miss some inputs. Proprietary software, informal labor arrangements, and other intangibles often don't appear in production surveys. When estimating markups, mismeasurement of factor shares is a first-order problem—it directly affects the denominator of the markup ratio. The effect on the markup estimate depends on how we estimate the output elasticity. For now, suppose we know that returns to scale are constant. As discussed in Section 1.2, with constant returns the markup is simply revenue divided by total cost: $\hat{\mu}_{it} = P_{it}Y_{it}/\text{Cost}_{it}$. If revenue exceeds total costs, then the firm earns profits from charging a markup (see equation (9)).[22] If we miss any inputs, we'll understate economic

[22]With constant returns, it is equivalent to focus on a single input. E.g., with materials $\gamma_{it}^M = c_{it}^M$, where c_{it}^M is materials' share of total costs. The measured markup is $\hat{\mu}_{it} = \hat{c}_{it}^M/\hat{s}_{it}^M$, but since both numerator and

costs and overstate the markup.

A simple example illustrates the consequences for markups. Consider a representative firm with true markup $\mu = 1.2$ and cost shares $c^M = 0.5$, $c^L = 0.3$, $c^K = 0.2$. What happens when we undermeasure output or each input by 10%?

Table 3: Impact of 10% Systematic Undermeasurement on Markup Estimates

Missing Input	Measured Markup	Bias
None (truth)	1.20	0%
Output	1.08	-10%
Materials	1.26	+5%
Labor	1.24	+3%
Capital	1.22	+2%

Note: Measured markup with cost shares is given by Revenue/Cost. Undermeasuring output directly affects revenue. Underestimating inputs affects measured costs.

Missing output decreases markups proportionally. Missing any input increases them. The bias approximately equals the input's cost share times the measurement error.[23]

If we estimate output elasticities through regression rather than cost shares, we face a standard omitted variable problem. All regression coefficients can be biased. Assuming we estimate a Cobb-Douglas function (12) and use materials to estimate the markup, we need its coefficient γ^M. Omitting some portion of, say, intangible capital biases $\hat{\gamma}^M$. With OLS, the bias depends on the partial correlation of the omitted K with M, given other included variables. Because this is a conditional correlation, this bias can be positive or negative even if the unconditional correlation of the omitted K and M is positive.

Second, micro datasets may not only omit relevant variables but also provide noisy measures of the variables we do observe. Since production functions have multiple regressors (inputs), classical measurement error attenuates the noisy regressor's own coefficient; measurement error in other, correlated regressors can push its estimate up.

To see the mechanics clearly, consider the case where only materials is measured with classical error: $\tilde{m} = m + u_m$ (where $m = \ln M$ and u_m is mean-zero, independent), while labor and capital are measured precisely. The OLS estimate of the materials coefficient—assuming the inputs are exogenous and $\mathbb{E}[\varepsilon_i | M_i, L_i, K_i] = 0$—converges to:

$$\text{plim}(\hat{\gamma}^M) = \gamma^M \cdot \frac{\text{Var}(m|\ell, k)}{\text{Var}(m|\ell, k) + \sigma^2_{u_m}} \tag{19}$$

denominator include $W^M_{it} M_{it}$, they cancel, yielding $\hat{\mu}_{it} = P_{it} Y_{it} / \text{Cost}_{it}$.

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

where $\sigma^2_{u_m}$ is the measurement error variance and $\ell = \ln L$, $k = \ln K$. For illustration, suppose measurement error variance equals 10% of the conditional variance of log materials—then the attenuation factor is $1/(1 + 0.1) = 0.91$, so the materials elasticity gets biased downward by about 9%. Since markups are calculated as $\mu = \gamma^M / s^M$, this 9% understatement of the elasticity translates directly into a 9% understatement of the markup (assuming the revenue share is measured without error).

When multiple inputs are mismeasured, the story becomes more complex. Measurement error in labor can partially offset materials' attenuation if the inputs are positively correlated, as OLS incorrectly attributes some of labor's productive contribution to materials. When noise obscures labor's true variation, OLS sees materials varying when labor should be varying, and assigns that productive variation to materials instead. With typical production correlations, this upward push from labor's measurement error could partially or even fully offset materials' own attenuation.

Capital measurement presents special challenges. Book values, depreciation schedules, and missing intangibles mean capital likely has the largest measurement error of any input. Collard-Wexler and De Loecker (2016) find through Monte Carlo under their calibration that when measurement error variance equals 40% of capital's conditional variance, capital coefficients can be biased downward by a factor of two. For the materials coefficient, capital's measurement error creates indirect effects that depend on how strongly materials correlates with capital. Since capital adjusts slowly while materials responds to current conditions, capital's correlation with flexible inputs like materials may be weaker than the materials-labor correlation. The net effect on markups depends on these offsetting forces: direct attenuation from materials' measurement error versus indirect effects from mismeasured labor and capital. Without knowing the actual measurement error variances and input correlations in a dataset, the direction of bias remains ambiguous.

These calculations likely understate the full complexity. With non-classical measurement error, omitted variables, or more sophisticated estimation methods like GMM, the biases become harder to predict and could go in either direction. Even in this simplified example, plausible measurement error levels can generate substantial markup biases—a 10% measurement error in materials alone creates nearly 10% downward bias in markups, while realistic combinations of measurement error across all inputs could shift markup estimates by 5–15% under reasonable calibrations. Given that many markup studies find changes of similar magnitudes over time or across industries, measurement error may drive some of the variation we attribute to real economic forces.

4.2 Lessons from economic measurement

Both macroeconomics and industrial organization faced a parallel challenge in the mid-20th century: interpreting residuals that mixed true economic phenomena with measurement artifacts. In macroeconomics, the puzzle was aggregate TFP. In industrial organization, it was profit rates and their relationship to market power. Both fields eventually concluded that improving data quality and measurement was a high-return investment, but they diverged sharply in their approaches.

The macroeconomic challenge emerged when Solow (1957) found capital per hour explained only 12 percent of U.S. output-per-hour growth from 1909 to 1949. Jorgenson and Griliches (1967) hypothesized that the residual 88 percent (TFP) reflected measurement error in output and inputs. Jorgenson and Griliches presents a "manifesto" (Berman and Jaffe, 2024, p.579) for improving measurement.[24]

Their main approach was to apply index-number implications of neoclassical production theory. They introduced Divisia (chained) measures of aggregate output and argued for hedonic adjustments to prices. They weighted hours of workers with different education or experience by relative wages. Similarly, they weighted different types of machines and structures by user-costs (implicit rental rates), capturing differences in marginal products.

Many of these ideas have subsequently been implemented in industry and aggregate data. At an industry or aggregate level, researchers can use survey methods (such as quality-adjusted price indices) that would be challenging to implement at a firm level, and they can combine multiple datasets. Measurement error at a micro level might cancel out when aggregated (Grunfeld and Griliches, 1960). Because of the easier availability and higher quality of industry data, the early implementations of the production approach to markups used industry data (Section 1.2).

Industrial organization had its own measurement crisis. Following Bain (1951), researchers spent a quarter-century documenting correlations between industry concentration and accounting profit rates (Bresnahan, 1989; Schmalensee, 1989). The goal of this "structure-conduct-performance" (SCP) paradigm was to infer market power. Did concentrated industries earn supernormal returns through anticompetitive conduct?

[24]Measurement 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, including industry aggregates (Berman and Jaffe, 2024; Fernald, 2024). Despite these advances, their hypothesis that TFP is merely measurement error falls short: from 1948–2024, TFP still accounts for roughly half of labor-productivity growth (Fernald, 2014).

The SCP paradigm failed on two fronts. First, conceptually, it suffered from causal ambiguity. A positive correlation between concentration and profits could reflect collusion (the SCP story) or, as Demsetz (1973) argues, it could reflect efficiency: productive firms naturally grow large and earn higher returns. Even with perfect data, the correlation is uninterpretable without additional economic structure. Second, and more directly relevant for production-markup estimation, the paradigm relied on accounting profitability measures, which poorly capture economic returns (Fisher and McGowan, 1983). At best, accounting data provides markups of price over average variable cost, not the economic marginal cost needed for proper inference about market power. Capital measurement is especially problematic: risk premia, inflation, depreciation rules, and the treatment of intangibles all drive wedges between accounting profitability and true economic returns. These aren't just historical concerns—the same measurement problems persist in the production approach to markups, notably if one uses cost shares. One needs all inputs and their economic costs.

Both fields responded to measurement challenges, though with divergent strategies. Macroeconomists and national accountants applied sophisticated surveys and statistical- and index-number methods to improve industry and aggregate data. The aim was partly descriptive—to better track the economy—and partly analytical: shrink the TFP residual and better estimate production functions (e.g., Jorgenson et al. (1987)). IO economists took a different path: they sought granular data within specific industries where prices and quantities could be observed separately, and developed structural models to identify the underlying economic forces. This approach allowed researchers to understand industry-specific accounting practices, institutional details, and competitive dynamics that affect how data should be interpreted—knowledge lost in aggregate approaches. Rather than inferring market power from aggregate patterns, modern IO uses detailed institutional knowledge about market structure, contracts, and competitive dynamics to discipline the analysis. The goal was to learn rigorously from carefully chosen case studies.

This history shapes how IO economists view production-based markups (Berry et al., 2019). The production approach avoids SCP's problematic causal claims—it doesn't regress profits on concentration. However, when revenue shares are measured and output elasticities estimated using book values of capital and accounting measures of costs, they inherit the measurement issues that undercut earlier attempts to infer economic fundamentals from accounting data. The same problems—depreciation schedules, missing intangibles, inflation adjustments, the gap between average and marginal costs—that prevented SCP from accurately measuring economic profitability now affect the production approach's estimates of output elasticities. An open question is the degree to which variation in markup estimates reflects these longstanding measurement artifacts.

4.3 Measurement challenges in microdata

With this backdrop—that measurement error complicates markup estimation—micro-level datasets are increasingly available. They still fall far short of the theoretical ideal. These datasets typically collect variables for accounting purposes rather than production estimation, leading to key limitations recognized since at least Griliches and Ringstad (1971).

Revenue vs. quantity data: Most micro datasets contain information on revenues, not quantities and (quality-adjusted) prices. In U.S. microdata, both Compustat and individual U.S. Censuses of different sectors offer disproportionately revenue-based measures of output. Separating output price from quantity is possible, but only for a small share of the economy. For instance, Foster et al. (2008) study 11 products from the manufacturing census, including coffee, ready-mixed concrete, and motor gasoline.[25] Such data are generally unavailable at a scale that would facilitate studying broad sectors or the economy as a whole.

When Foster et al. (2008) can separate prices from quantities in their 11 products, they uncover a striking puzzle: physical productivity is inversely correlated with price, yet revenue productivity is positively correlated with price. How can the same underlying efficiency generate opposite correlations? The answer reveals why revenue data confounds production function estimation. A physically productive firm has lower costs and can profitably charge less—hence the negative price correlation. Revenue productivity (nominal output, $P_{it}Y_{it}$, per unit of input) mixes efficiency with pricing power. High revenue productivity might reflect genuine efficiency or simply high markups. Without observing prices and quantities separately, we risk conflating operational excellence and market power. We return to this challenge in Section 5.[26]

Input quality and coverage: Microdata rarely have direct information on input quality. This biases estimates of production functions and productivity across firms (Fox and Smeets, 2011; Grieco et al., 2016). Sometimes inputs themselves aren't even recorded. For instance, 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, as most sectoral Census microdata don't contain systematic information on intermediates. Even in manufacturing, the data miss inputs of business services.

[25] Survey data from other countries—for instance Colombia and India—sometimes separate output price from quantity. And many customs datasets used in international trade have unit values.

[26] Research from other countries reinforces these concerns. For example, Lenzu et al. (2023) combine firm-level output prices and quantities with quasi-experimental variation in credit supply for Belgian firms to show that financial shocks have different effects on physical and revenue productivity. Revenue-based measures underestimate the long-run elasticity of physical productivity to credit supply by almost half.

Capital measurement: Micro datasets often provide only book values of capital—historical purchase prices adjusted by accounting depreciation rules—rather than economic values reflecting productive capacity. Book values can diverge dramatically from productive input: a fully depreciated but still-productive machine shows zero book value; an obsolete asset purchased recently maintains high book value.

This problem intensifies with intangible assets. A growing literature focuses on measuring intangible capital beyond just software and R&D, including assets such as brand equity and organizational capital (Corrado et al., 2009). These intangibles typically appear in accounting data as operating expenses—R&D, marketing, management consulting. Economically, they create durable productive assets that should be capitalized. Karabarbounis and Neiman (2019) show that what they call "factorless income" (the apparent profits after subtracting measured payments to labor and capital from GDP) has grown substantially. This could reflect true profits from market power (as the cost-share approach would imply). Or it could reflect that our discount rate was too low. Or it could represent returns to unmeasured intangible capital. In terms of the latter, expensing rather than capitalizing intangibles overstates the apparent profit rate: costs appear lower than they truly are because we miss the implicit rental cost of intangible capital. Under the cost-share approach, this inflated profit rate translates directly into inflated markup estimates.

4.4 Representativeness and coverage issues

Even if we could measure everything perfectly at the firm level, would it tell us about the aggregate economy? For macroeconomists, the ideal micro dataset would offer not just high-quality measurement but also broad coverage across the economy.

The U.S. Census of Manufactures has been the primary fuel for macroeconomists (Dunne et al., 1988; Baily et al., 1992; Foster et al., 2008; Syverson, 2011). Manufacturing is now less representative, accounting for only about 12 percent of U.S. private-industry value added (circa 2020, BEA). If manufacturing markups differ systematically from those in services or finance, we're learning about a shrinking slice of the economy.

This sectoral shift matters for measurement too. Manufacturing firms have relatively clear distinctions between production workers and overhead, between raw materials and administrative expenses. Service firms blur these boundaries. A software engineer at Google—production labor or overhead? As the economy shifts toward services, accounting categories designed for manufacturing become less economically meaningful. Still, U.S. Census microdata remains best-in-class internationally, and researchers are beginning to exploit within-sector detail such as transaction data from Europe that capture firm-to-firm input linkages.

Given manufacturing's declining share, researchers increasingly turn to Compustat, which has broad sectoral coverage and is easy to access. De Loecker et al. (2020) show how to use Compustat's accounting measures for markup estimation. Compustat covers only public firms, which account for only half of industry sales as reported in the national accounts (circa 2015). Some 40 percent of sales in Compustat today come from foreign operations (2020s). Several papers (Ali et al., 2008; Keil, 2017; Decker and Williams, 2023) find that common concentration measures at the 4-digit NAICS level are weakly correlated with those based on the more comprehensive economic census.

Compustat's accounting detail offers both advantages and challenges for markup estimation. Under Generally Accepted Accounting Principles (GAAP), firms classify costs as either Cost of Goods Sold (COGS) or Selling, General, and Administrative (SG&A) expenses. COGS includes direct materials, direct labor, and manufacturing overhead—costs directly attributed to producing goods. SG&A captures expenses not directly tied to production: executive salaries, marketing, research and development, and administrative overhead. De Loecker et al. (2020) leverage this distinction by treating COGS as the flexible input and excluding SG&A entirely from the production function.

These accounting rules exist for inventory valuation and tax purposes, not economic analysis. Traina (2018) shows this classification matters enormously: including SG&A with COGS as the flexible input eliminates most of the rise in estimated markups since 1980. Consider a firm with COGS of \$70 million and SG&A of \$20 million generating \$100 million in revenue. The flexible input share is either 0.70 (COGS only) or 0.90 (COGS+SG&A)—implying markups could be 43% or 11% for the same output elasticity, assuming $\gamma = 1$ for this composite flexible input. The COGS share of total costs has declined substantially over recent decades. Is this rising market power or accounting reclassification? Without understanding firms' specific accounting choices and how they've evolved over time, we cannot tell. The important research question that motivated Jorgenson and Griliches (1967) remains, especially in microdata: To what degree does input mismeasurement drive our results?

Bridging micro and macro Whatever the dataset, researchers must take a stand on fundamental measurement issues: defining industries over which production function coefficients are constant, measuring capital (Hall and Jorgenson, 1967; Becker et al., 2006), determining the correct measure of labor (Fox and Smeets, 2011), handling multi-product firms (De Loecker, 2011a), and choosing between plant- or firm-level analysis.

For any dataset, researchers should report how the data, once aggregated, compare with published totals as standard descriptive statistics. Fernald and Piga (2023) find that factor

shares are wildly different in the U.S. Census of Manufacturers from those in the national accounts. In 1977, labor's share of value added in manufacturing plants was only 31 percent, as opposed to 58 percent in the manufacturing sector of the national-accounts-based BEA-BLS KLEMS dataset. Such large discrepancies in factor shares translate directly into different markup estimates—what looks like a high markup might just be missing labor compensation or business services. Even when inputs are measured correctly in nominal terms, using industry-wide deflators instead of firm-specific prices can bias elasticities downward, particularly for labor (Ornaghi, 2006).

The gap between micro data and National Accounts reflects more than just coverage. National Accounts integrate multiple data sources—tax records for proprietors' income, banking surveys for financial services, innovation surveys for R&D—that researchers rarely access when working with micro data. These auxiliary datasets help split proprietors' income between labor and capital compensation, allocate business services across industries, and capture intangible investments. Without understanding this multi-source integration process, researchers might incorrectly expect their micro data to aggregate to published totals, when in fact the National Accounts rest on a broader empirical foundation guided by accounting identities and economic theory.

What to do about measurement challenges Given these pervasive measurement issues, three approaches stand out. First, the Jorgenson-Griliches agenda of improving underlying data remains as relevant today as in 1967. Better measurement of inputs—especially capital services, intangible investments, and the boundary between production and overhead labor—would strengthen the foundation for production-based markup estimation. As Griliches and Mairesse (1998) emphasized, we particularly need better firm-level price data. Lev (2001) documents how traditional accounting overlooks intangible capital, from R&D and software to organizational capital and brand value.

Second, researchers should systematically consider how specific measurement errors might shape their results. If markups differ across industries, could this reflect differences in measurement quality rather than competitive conditions? Testing alternative input definitions and accounting treatments helps rule out measurement-driven explanations.

Third, econometric methods can partially address measurement error. Several papers show that correcting for capital mismeasurement can nearly double estimated capital elasticities (Ližal and Galuščák, 2012; Collard-Wexler and De Loecker, 2016; Kim et al., 2016). While these methods cannot fully overcome poor data, they can help separate noise from true economic signals. The key insight is that measurement challenges are not mere nuisances—they fundamentally shape what we can learn about markups using the production approach.

5 Practice and pitfalls in estimating output elasticities

This section builds on the discussion of data constraints and the garden of forking paths by focusing on the estimation of output elasticities. Even after identifying a flexible input, the method used to estimate its elasticity can substantially affect both the elasticity itself and the markup inferred from the production-based FOC (4), as Table 1b illustrated. A fundamental tension complicates this task: many of the tools for estimating production functions were developed not only for a world of perfect data, but also of perfect competition. Yet our goal is to measure departures from unity—markups that arise under imperfect competition. We organize our discussion of these estimation challenges in three steps.

First, we highlight **endogeneity as the central concern.** Input choices typically respond to unobserved productivity shocks, biasing naive OLS estimates of output elasticities. To make this point transparently, we focus on a benchmark case in which all firms share a canonical Cobb-Douglas production function. This parametric setting allows us to isolate key econometric challenges and show how data limitations exacerbate them. The canonical "transmission bias" is just the first layer of a deeper problem. Most datasets record only revenue—not separate prices and quantities—shifting endogeneity concerns from unobserved productivity to unobserved "revenue productivity" shaped by demand forces. These limitations compound the endogeneity problems.

Second, we outline and provide a critical **tour of the econometric tools** developed to address these forms of endogeneity. We discuss instrumental variables, control function approaches, and dynamic panel methods. Each addresses different aspects of the identification challenge, but each carries its own set of strong assumptions and potential pitfalls. A key recurring concern is that, because of their historical origin for competitive settings, estimators of output elasticities themselves often depend on markups (i.e., the level of true markup affects the estimated output elasticity), creating a circularity problem that is difficult to resolve: we need to know markups to estimate output elasticities, which we then use to infer markups from FOC (4).

Third, we return to the practical question of **how flexibly to specify the production technology** itself. Assumptions like constant returns to scale (CRS) or value-added production simplify estimation and, in some cases, improve identification. Yet these restrictions are strong: CRS imposes that marginal and average costs are equal, and value-added models effectively rely on a Leontief structure that is rarely realistic at the firm level. Such assumptions risk conflating technological heterogeneity with variation in markups. At the same time, relaxing these restrictions introduces substantial complexity and new estimation

challenges. The key takeaway is that there is no one-size-fits-all approach—careful, transparent modeling choices, combined with robustness checks and recent diagnostic tools, can help balance tractability with economic realism.

The applied economist has a rich toolbox for addressing these challenges. However, the tools involve tradeoffs. Some tools are restrictive for an industrial organization economist who wants to remain agnostic about the structure of production and competition. Others are overly general for a macroeconomist who wants to develop or calibrate fully specified structural models. In both cases, a world of imperfect competition calls for enriching the traditional toolkit, and the first step is a clear-eyed understanding of its limitations.

5.1 The anatomy of endogeneity

Many of the econometric challenges to estimating output elasticities—and the potential tools for overcoming those challenges—can be seen in the conditional demand for a flexible input. To be concrete, consider the Cobb-Douglas production function (12). Assuming materials, m_{it}, are flexibly chosen, the first-order condition (3) defines their conditional demand as:

$$m_{it} = \frac{1}{1-\gamma^M}\left(a_{it} + p_{it} + \log\gamma^M - w_{it}^M - \log\mu_{it} + \gamma^K k_{it} + \gamma^L l_{it}\right), \tag{20}$$

with a_{it} the firm's (log) productivity, p_{it} the (log) output price, w_{it}^M the price of materials, k_{it} and l_{it} as the labor and capital inputs, and γ^X for $X \in \{K, L, M\}$ as the output elasticities. This equation shows how multiple unobservables simultaneously determine a firm's input choices. The markup μ_{it} enters negatively because higher markups reduce optimal input usage: when firms charge prices above marginal cost, they restrict output below the competitive level and thus demand fewer inputs. For a given markup, a higher price leads firms to want to sell more.

Transmission bias The canonical concern for estimating output elasticities is the transmission bias from unobserved productivity a_{it} to inputs: a_{it} enters both the production function (12) and the firm's optimal materials choice in (20). A firm that receives a positive productivity shock will, all else equal, use more materials. The resulting correlation between the input m_{it} and the unobserved error component a_{it} violates the OLS exclusion restriction $\mathbb{E}[m_{it}a_{it}] = 0$. All else equal, a naive OLS estimate of the materials elasticity, $\widehat{\gamma}^M$, is biased upward, confounding the true technological parameter with the firm's endogenous response to productivity.[27] Since the production approach calculates markups as $\mu_{it} = \widehat{\gamma}^M / s_{it}^M$, this

[27] All else is not always equal. With sticky prices and market power, the bias can be more complex. If all firms receive the same positive productivity shock but cannot adjust prices, markups rise (as marginal costs

confounding translates directly into biased markups. Intuitively, because higher a_{it} induces higher m_{it}, OLS attributes their correlation to technology and overstates $\widehat{\gamma}^M$; dividing by s_{it}^M then overstates μ_{it}, making OLS a (weak) upper bound.

Omitted price bias The problem deepens because most datasets provide firm revenue, not separate prices and quantities. Let $r_{it} = p_{it} + y_{it}$ denote the firm's log revenue. Adding price p_{it} to both sides of the production function (12), the estimating equation becomes:

$$r_{it} = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it} + a_{it} + p_{it}. \tag{21}$$

Prices p_{it} join productivity a_{it} as unobserved determinants in both the estimating equation (21) and the materials demand (20). Firm-specific demand shocks that affect prices also affect input choices, creating a second layer of endogeneity. Estimating this revenue production function via OLS therefore confounds the technological parameter γ^M with endogenous input responses to both productivity a_{it} and non-technological forces that influence output prices p_{it}—including competition.

The omitted price bias leads to a troubling circularity: output elasticities are contaminated by the very markups we seek to estimate. Under standard models of imperfect competition, firms set prices as a markup over marginal cost, so in logs $p_{it} = \log \mu_{it} + \log \lambda_{it}$. As a result, the error term embeds markup variation both directly (through pricing) and indirectly (through correlations between markups and productivity). Because input choices also respond to markups (see (20)), regressors correlate with the markup-contaminated error. We thus need to know markups to estimate output elasticities correctly (by accounting for the markup's effect on prices in the error), and we need those elasticities to estimate markups (via the production-approach FOC).

Given these challenges, researchers often turn to simple fixes, but these are generally inadequate. A common approach is to deflate firm revenues using an industry-level price index, p_t, leaving firm-level deviations, $\tilde{p}_{it} = p_{it} - p_t$, embedded in the error term. Under perfect competition, firms charge the same quality-adjusted price. Then $\tilde{p}_{it}$ is uncorrelated with input choices; and, indeed, if price gaps mostly reflect quality, revenue productivity may be the relevant object (De Loecker et al., 2016). But under imperfect competition—the case of interest—Klette and Griliches (1996) show that output prices and input choices are negatively correlated. As more productive firms charge lower prices and produce more, output elasticities are then biased downward alongside the inferred markups.

fall while prices remain fixed), dampening input demand through the $\log \mu_{it}$ term. This creates an opposing bias, leaving the overall effect ambiguous and showing how market frictions can alter the bias.

Indeed, Bond et al. (2021) argue that production-based markups estimated from revenue data are uninformative about the level of true markups. The key insight is that with revenue data, the regression recovers a revenue elasticity $\hat{\gamma}_{it}^{\text{revenue},X} = \gamma_{it}^X/\mu_{it}$ that conflates the true output elasticity γ_{it}^X and the true markup μ_{it}. Recall that the revenue share of a flexible input equals the ratio of an output elasticity and a markup, $s_{it}^X \equiv \frac{W_{it}^X X_{it}}{P_{it}Y_{it}} = \gamma_{it}^X/\mu_{it}$. Therefore, using a revenue elasticity to infer a production-based markup, $\hat{\mu}_{it}^{\text{revenue}} = \hat{\gamma}_{it}^{\text{revenue},X}/s_{it}^X = 1$, yields no information about market power. Extending Bond et al. (2021), Hashemi et al. (2022) show that industry-level deflation resolves the issue only in the special case of perfect competition or identical firm-level prices—precisely when markups equal one.

Collecting firm-specific price data to render $\tilde{p}_{it}$ observable would address this problem directly. Such efforts are increasingly feasible and valuable. Manufacturing and trade datasets now sometimes contain quantity information, particularly for homogeneous goods where units of output are well-defined. Services remain challenging (what's a unit of healthcare?) as do differentiated products. Still, when available, such data provide validation of revenue-based estimates and help researchers understand the magnitude of potential biases.

An emerging literature using data that distinguishes prices from quantities at a firm level shows that the difference between physical and revenue productivity matters for measuring markup levels and, at times, their trends. Using simulations and data from French manufacturing firms where unit prices can be constructed, De Ridder et al. (2024) provide the most systematic assessment. Their simulations focus on a repeated, static oligopoly model (both Cournot and Bertrand cases), yielding a 0.9 correlation between revenue-based and true log markups. In the French data, however, the correlation is only 0.3. This gap suggests that the true magnitude of the bias depends on the specifics of demand, competition, and production parameters. Given all the other moving parts to estimation, the biases might offset or compound in complex ways. Still, De Ridder et al. (2024) provide insights and, more importantly, a template for examining these biases in specific settings.

More generally, the concerns about endogeneity extend to other unobservables, creating additional confounding forces that bias the estimated output elasticities and hence the measured markups. A common and closely related one arises if we only observe expenditures on inputs ($e_{it}^M = w_{it}^M + m_{it}$). In this case, the revenue production function is:

$$r_{it} = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M e_{it}^M + a_{it} + p_{it} - \gamma^M w_{it}^M. \tag{22}$$

The unobserved, firm-specific input price, w_{it}^M, now also enters the error term. If firms with more materials market power face lower input prices and thus purchase more inputs, this creates another source of endogeneity, as discussed in section 3.1.

To see the full scope of the confounding problem, consider the error term we face. From equation (22), the error term in our revenue-based estimating equation is:

$$\varepsilon_{it} = a_{it} + p_{it} - \gamma^M w_{it}^M \tag{23}$$

Any method for estimating γ^M must confront this composite error, where each component represents a different confounding force. The traditional solutions in the next section—instruments, control functions, dynamic panels—were designed for a world where $\varepsilon_{it} = a_{it}$, addressing only the first layer of confounding. But in the real world of revenue data and unobserved prices, these methods must somehow address all three confounding components. Looking back at our materials demand equation (20), we see the fundamental tension: every component of this composite error drives input choice, yet all are unobserved. The very flexibility that makes materials suitable for the FOC approach ensures they respond to forces we cannot measure. As we'll see, this is a formidable challenge that existing methods only partially address.

5.2 Estimating output elasticities: Solutions

The literature has developed a range of tools to address the confounding problem. The stakes are high: in practice, the biases can be substantial. For pedagogical clarity, our tour of the main approaches evaluates how each handles the two primary layers of endogeneity: unobserved productivity (a_{it}) and unobserved output prices (p_{it}). While the third layer—unobserved input prices—is a real concern, we focus on the first two layers and note where input-price issues arise.

Fixed effects introduce firm-specific intercepts a_i to control for fixed firm characteristics (Mundlak, 1961). These intercepts absorb time-invariant productivity a_i and, with the composite error, also time-invariant prices and markups. Historically, econometricians developed these models for agricultural production where productivity largely reflects time-invariant characteristics like soil fertility. But in modern applications, unobserved productivity likely includes both a fixed and a time-varying component $a_{it} = a_i + e_{it}$. Similarly, prices may have fixed and time-varying components. Fixed effects remove the combined fixed component $(a_i + p_i)$ through demeaning or differencing, but time-varying components remain.

Fixed effects can also exacerbate measurement error, often yielding implausibly low or even negative capital coefficients. 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 estimates. One hypothesis is that capital is a quasi-fixed input—responding slowly to productivity shocks at annual horizons—so the within-firm variation is dominated by measurement error.[28] This also implies that changes in capital, which capture depreciation and investment, are potentially contaminated by measurement error. In an in-depth study of capital measurement issues, Becker et al. (2006) find that different ways of measuring capital that ought to be equivalent, such as using perpetual inventory methods or inferring capital investment from the capital producing sectors, lead to different results for a variety of outcomes, including parameter estimates of the production function and investment patterns. Measurement was the primary reason fixed effects were largely abandoned for production function estimation (De Loecker and Warzynski, 2012).

Instrumental variables offer the most conceptually straightforward solution. The method seeks to isolate variation in an input that is orthogonal to the composite error term. To solve both layers of the problem, a valid instrument z_{it} must be correlated with the input choice $\mathbb{E}[m_{it} z_{it}] \neq 0$ but strictly uncorrelated with both productivity and prices, such that $\mathbb{E}[z_{it}(a_{it} + p_{it})] = 0$. Industry time-series approaches (Section 1.2) typically used aggregate demand instruments.

Control functions offer a complementary approach to addressing endogeneity using proxy variables. But whereas instrumental variables are correlated with exogenous input variation (and thus uncorrelated with productivity), proxy variables are correlated with endogenous input variation (and thus explicitly correlated with productivity). Key to this approach is the requirement of monotonicity: productivity's effect on the proxy must be strictly increasing or decreasing, allowing us to use information from the proxy variable to control for otherwise unobserved productivity.

Olley and Pakes (1996) pioneered control functions in the productivity literature. They show a large class of models implies that, conditional on the capital stock, investment demand depends monotonically on productivity. This monotonic relationship makes it possible to invert the investment-demand function and express unobserved productivity as a function of the observed capital stock k and investment i:

$$i_{it} = g(k_{it}, a_{it}) \implies a_{it} = g^{-1}(k_{it}, i_{it}). \tag{24}$$

[28] The terminology here matters: IO economists typically call capital "quasi-fixed" or "predetermined," meaning it doesn't respond to productivity shocks within the period. This differs from "fixed costs," which are overhead expenses necessary to produce at all. The quasi-fixed nature of capital creates econometric challenges (little useful variation) which, as discussed in section 4.1, can bias all coefficient estimates.

Substituting out unobserved productivity a_{it} in the production function (12) with the control function $g^{-1}(k_{it}, i_{it})$ yields an estimable equation that is entirely a function of observables:

$$y_{it} = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it} + g^{-1}(k_{it}, i_{it}). \tag{25}$$

As investment is often lumpy and infrequent, Levinsohn and Petrin (2003) proposed using intermediate inputs as an alternative proxy.[29]

Intermediate inputs are typically adjusted more smoothly than investment, have fewer confounding determinants (e.g., any serially correlated unobservable), and can respond more directly to productivity shocks (think using more steel when you run extra shifts in the factory). To see how materials can proxy for productivity, start from our parametric materials demand equation (20). If—for the moment—we assume perfect competition ($\mu_{it} = 1$), common input prices, and hold labor and capital fixed at their observed values, materials demand simplifies to:

$$m_{it} = h(k_{it}, l_{it}; a_{it}) \tag{26}$$

where the semicolon indicates that k_{it} and l_{it} are held fixed. If this function is monotonic in a_{it}, we can invert it:

$$a_{it} = h^{-1}(k_{it}, l_{it}; m_{it}), \tag{27}$$

and the inverted function $h^{-1}(k_{it}, l_{it}; m_{it})$ expresses unobserved productivity in terms of observables while avoiding the pitfalls associated with using investment.

In addition to the earlier requirement of monotonicity, inversion of input demand also requires that productivity is the only unobserved factor affecting investment or materials decisions (i.e., the "scalar unobservable" assumption). In other words, forces like demand shocks or measurement error cannot independently drive the proxy. With revenue productivity $a_{it}+p_{it}$ in the error, this assumption becomes problematic: if firms' investment or materials choices respond to demand conditions that affect prices, the control function cannot be inverted. Note also that if prices and productivity move in opposite directions—the Klette-Griliches logic where more productive firms charge lower prices—the composite revenue productivity may fail to be monotonic in a_{it}, further undermining the invertibility of the control function.

Control functions rely on two more assumptions. First, productivity follows a Markov process: today's productivity depends only on yesterday's productivity. This process puts only limited structure on how productivity evolves but is a key facilitator of the implementation discussed below. Second, any difference between observed output and output predicted

[29] It's possible to control for productivity within the subset of observations with nonzero investment.

by the model is assumed to represent either measurement error or a productivity innovation that occurs *after* inputs are chosen. In this way, missing structural determinants (e.g., misspecification of the production process) are assumed *not* to play a role.

With these assumptions, control functions are implemented as follows:

1. Regress output flexibly on inputs and the control variable (e.g., with a third-order polynomial) to estimate the combined production and control function:

$$y_{it} = \phi_t(k_{it}, l_{it}, m_{it}, i_{it}) + \epsilon_{it} \tag{28}$$

   where $\phi_t(\cdot) = \gamma^K k_{it} + \gamma^L l_{it} + \gamma^M m_{it} + g^{-1}(k_{it}, i_{it})$. The residual ϵ_{it} is interpreted as measurement error. Why? If the control function perfectly captures productivity (the scalar unobservable assumption), then the equation should fit exactly except for measurement error. The fitted value $\hat{y}_{it} = \phi_t(\cdot)$ thus provides output "cleaned" of measurement error. Importantly, this step does not identify any individual production function parameters since we cannot separately identify $\gamma^M m_{it}$ from the part of $g^{-1}(\cdot)$ that depends on materials.

2. For any candidate set of production function parameters $(\gamma^K, \gamma^L, \gamma^M)$, back out productivity as the difference between cleaned output from step 1 and predicted output:

$$\hat{a}_{it}(\gamma) = \hat{y}_{it} - \gamma^K k_{it} - \gamma^L l_{it} - \gamma^M m_{it}. \tag{29}$$

3. Since productivity is Markovian, regress it on its lag to decompose it into its predictable component and its innovation: $\hat{a}_{it} = f(\hat{a}_{i,t-1}) + \xi_{it}$. As the innovation ξ_{it} is orthogonal to inputs chosen before the innovation is realized, lagged inputs $z_{i,t-1}$ form the basis of moment conditions for estimation: $\mathbb{E}[\xi_{it}(\gamma) \cdot z_{i,t-1}] = 0$. These moments identify production-function parameters through GMM or similar methods.

   But here's another place where the omitted price bias can pose a problem. With revenue data, we're not recovering physical productivity a_{it} but revenue productivity $a_{it} + p_{it}$. The control function approach works if this composite follows a Markov process, but that requires both productivity and prices to evolve in a coordinated way. When demand shocks affect prices independently of productivity, the Markov assumption likely fails and the moment conditions become invalid.

Though widely used, control function methods face identification challenges even apart from omitted prices. The output elasticity of a flexible input—precisely the sort needed for markup inference in equation (3)—turns out to be underidentified in many standard settings. The

identification challenge comes from the dual roles played by a proxy variable in estimation: (i) accounting for unobserved productivity and (ii) generating input variation that identifies the output elasticity. If a control variable can be flexibly adjusted, then it fully reflects variation in productivity. When materials optimally adjust to productivity, they become collinear with it, leaving no variation to identify the materials elasticity. Ackerberg et al. (2015) address this point by assuming labor is chosen before materials, though this shifts rather than solves the problem. In a GMM framework, this lack of variation means some of the moment conditions are redundant, causing identification to fail. Ackerberg et al. (2015), building on an earlier insight in Bond and Söderbom (2005), elaborate on these concerns for value-added production functions. Gandhi et al. (2020) show identification concerns of this sort are even more severe in gross-output settings.

While remedies exist to restore identification, many violate assumptions behind the inference of production-based markups. For instance, Bond and Söderbom (2005) show frictions and adjustment costs can help generate identifying variation for production-function estimation. But the production-based markups in equation (3) should be inferred from a flexibly adjusted input. Similarly, Gandhi et al. (2020) propose using a competitive firm's first-order condition equating the flexible input's output elasticity to its revenue share as an additional moment to recover identification. But this method assumes perfect competition, creating another circularity problem: markup estimation requires production function parameters, which themselves assume perfect competition ($\mu_i \equiv 1$).

Identification problems compound when firms do not flexibly adjust output in response to productivity shocks. An empirical and theoretical macro literature argues that, because of sticky prices or inelastic short-run demand (e.g., Galí, 1999; Basu et al., 2006), when productivity improves, firms might first reduce the use of a flexible input and then, with a lag, increase it. Customer-market frictions provide another reason: building relationships takes time, preventing immediate output expansion. This would violate the control-function monotonicity assumption. While such effects are well-documented in industry and aggregate data, they could, in principle, also arise in microdata.

Doraszelski and Jaumandreu (2019) argue that imperfect competition undermines the control-function approach: if firms' input choices depend on both productivity and unobserved markup determinants (such as demand conditions and firm conduct), the control function becomes uninvertible and cannot fully capture unobserved productivity. In such cases, Ackerberg et al. (2007) show that two independent controls are needed to account for the two latent state variables. One proposal is to use market shares as an additional control (e.g., De Loecker et al., 2020; De Ridder et al., 2024), though this requires assumptions about the

competitive environment that dilute the appeal of a production-based approach rooted solely in cost minimization. More generally, incorporating non-price variables into the control function can proxy for unobserved firm-specific prices. For instance, De Loecker et al. (2020) use revenue market shares, while Kirov et al. (2025) advocate including fixed effects and observable price controls directly in the first-order conditions (see also Kasahara and Sugita 2020; Ackerberg and De Loecker 2024). These strategies capture the price variation that drives input choices, but their validity rests on assumptions about firm behavior (e.g., Bertrand or Cournot competition)—precisely the structural detail the production approach aims to remain agnostic about. If such assumptions are indispensable, it may be more transparent to estimate production jointly with models of demand and competition.[30]

Dynamic panel methods offer an alternative that does not require the scalar unobservable assumption from the earlier discussion. These methods replace the control-function approach's informational assumptions with alternative assumptions about the time-series properties of unobservables. Consider a first-differenced production function:

$$\Delta y_{it} = \gamma^K \Delta k_{it} + \gamma^L \Delta l_{it} + \gamma^M \Delta m_{it} + \Delta a_{it}. \tag{30}$$

If a_{it} follows an AR(1) process, $a_{it} = \rho a_{i,t-1} + \varepsilon_{it}$, then $\Delta a_{it} = \rho \Delta a_{i,t-1} + \Delta \varepsilon_{it}$. Inputs from $t-2$ are now valid instruments for differenced inputs because they are uncorrelated with productivity innovation $\Delta \varepsilon_{it}$ but correlated with current inputs via persistence. This generates moment conditions like $\mathbb{E}[\Delta \varepsilon_{it} \cdot m_{i,t-2}] = 0$. System GMM combines these differenced equations with level equations to improve efficiency (Blundell and Bond, 1998, 2000; Arellano and Bond, 1991).

With revenue data, however, the error term is $\Delta a_{it} + \Delta p_{it}$; the validity of lagged inputs as instruments requires that both productivity and demand shocks follow similar time-series processes. This requirement seems implausible when demand responds to different forces than productivity. Moreover, if productivity follows a more complex process than AR(1)—say $a_{it} = \rho_2 a_{i,t-1}^2 + \rho_1 a_{i,t-1} + \varepsilon_{it}$—then the "innovation" includes predictable components that are correlated with lagged inputs, invalidating the moment conditions. As Bond and Söderbom (2005) argue, these lagged instruments typically only have identifying power if there are frictions like adjustment costs. However, such frictions violate the static first-order condition that underpins the markup calculation itself.

Recent work has refined these methods, but there remains ample scope for further research.

[30] A subtler risk is overfitting: an ideal control function uses proxies correlated with omitted prices but not directly determined by them, an exclusion restriction of sorts. See Kirov et al. (2025) for further discussion.

Brand (2019) proposes a method that treats observed output as a noisy signal of productivity, allowing unobservables to evolve nonlinearly. We can use lagged output as an instrument to distinguish current productivity from measurement error, provided productivity is persistent and measurement error is not (and with at least three periods of data). More generally, a valuable validation of dynamic-panel methods would use firm-level datasets that separate output prices from quantities to assess whether physical and revenue productivity measures have time-series properties consistent with the approach. After all, those assumptions underpin confidence in the estimated output elasticities—and hence in any markups inferred through the production approach.

Structural demand modeling addresses the problem of omitted prices by deriving estimating equations that can be written in terms of observed revenue rather than unobserved quantity of output. For instance, as originally suggested by Klette and Griliches (1996), we could specify an isoelastic CES demand curve, here with η as the price elasticity of residual demand and p_t and y_t as industry quantity and price indexes:

$$y_{it} = y_t - \eta(p_{it} - p_t). \tag{31}$$

By combining this demand curve with the production function (12), we can derive a revenue-based estimating equation:

$$r_{it} - p_t = \underbrace{\frac{\eta-1}{\eta}\gamma^K}_{\beta^K} k_{it} + \underbrace{\frac{\eta-1}{\eta}\gamma^L}_{\beta^L} l_{it} + \underbrace{\frac{\eta-1}{\eta}\gamma^M}_{\beta^M} m_{it} + \underbrace{\frac{1}{\eta}}_{\beta^Y} y_t + \frac{\eta-1}{\eta} a_{it}. \tag{32}$$

This framework lets us recover revenue elasticities β using methods like the control function approach. We can then separate output elasticities γ from the demand elasticity η: $\gamma = \beta \times \eta/(\eta - 1)$. While the above illustration uses the case of CES demand, similar ideas can be applied with richer models of demand. De Loecker (2011a) extend this approach to multiproduct firms; Ruzic and Ho (2021) and Choi et al. (2024) extend the method to models of heterogeneous markups and oligopolistic competition. Richer demand models (nested logit, random coefficients) provide more flexible substitution patterns.

Jointly estimating production and demand models not only addresses concerns regarding missing prices (through structural assumptions on demand), but also allows for counterfactual exercises. At the same time, structural modeling of demand undermines one of the production approach's key appeals: estimating markups without specifying market structure. If we must assume Bertrand competition or CES demand to estimate γ^M, we've essentially

returned to traditional IO methods where markups depend on our demand and conduct modeling choices. The cost of the approach, as with traditional IO demand estimation, is that markups are functions of the model structure and counterfactual conclusions regarding market power are conditional on this specified demand and conduct. The benefit is transparency: assumptions are explicit rather than buried in econometric fixes, and a fuller model is more suitable for counterfactual analysis.

Each proposed solution in this section grapples with the same core challenge: the error term contains revenue productivity $a_{it}+p_{it}$, not just productivity. Fixed effects and deflation work only under restrictive conditions. Instrumental variables need exogenous variation uncorrelated with both productivity and prices—a high bar. Control functions assume productivity is the sole unobservable, failing when demand affects input choices. Dynamic panels rely on assumptions about the evolution of multiple unobservables. Joint estimation addresses these issues by imposing structure but sacrifices the model-free appeal of the production approach.

Most solutions add frictions or structure that weaken the premise and the appeal of the production approach to markups. The ultimate choice of estimation strategy depends on the research question, the institutional setting and the available data. Each method makes different compromises between identification and the assumptions underlying markup inference. The apparent simplicity of needing just one output elasticity masks the full complexity of the identification problem—properly estimating that elasticity requires confronting all the challenges of production function estimation under imperfect competition.

5.3 Market power vs. technology, redux: Issues of specification

Despite the appeal of modeling output elasticities more flexibly—discussed in Section 3.2—practical specification choices affect how we separate market power from technology.

Returns to scale: Constant or not? A common technology restriction is to impose constant returns to scale (CRS). For a range of approaches, CRS simplifies the estimation and identification of output elasticities. However, any true heterogeneity of returns to scale—in the cross section or the time series—risks being interpreted as heterogeneity in markups.

A prime example of how CRS simplifies the estimation of output elasticities is with cost shares. As discussed in Sections 1.2 and 2, under CRS the output elasticity γ_{it}^X equals the factor's share in costs c_{it}^X. Cost shares elegantly address many challenges raised in this section: endogeneity (cost shares are equilibrium outcomes requiring no production function estimation), missing prices (we compare cost shares to revenue shares directly), and

functional-form flexibility (the FOC holds for any production function).[31] As a complement to other approaches, showing results with cost shares is a useful robustness exercise.

However, measuring total cost—the denominator of the cost share—poses practical challenges. The primary issue, and a key reason IO moved away from cost-share approaches, is the measurement of capital and its user cost. As highlighted in Sections 3 and 4, capital quantities in microdata are prone to significant measurement error. Economically, most capital is owned rather than rented at observable market rates, so user costs must be imputed and firm-specific risk premia incorporated. A large finance literature shows that these premia can be as variable as productivity itself. In addition, non-markup wedges $(1 + \tau_{it}^{X})$ in the FOC (14)—such as adjustment costs or other shadow costs—must also be accounted for. Omitting these unobserved wedges risks understating costs and overstating markups.

The CRS assumption is also frequently used to make the estimation of more flexible production functions more tractable. CES or translog, for instance, are attractive a priori because they allow more flexibility in output elasticities across producers. However, a two-input translog (as in footnote 9) has six parameters (one constant term, two linear terms, three quadratic terms); three inputs has 10 parameters (one constant, three linear, six quadratic); four inputs has 15 (one constant, four linear, 10 quadratic). Imposing constant returns to scale and perfect competition introduces cross-parameter restrictions that reduce the parameter space and make estimation more tractable (e.g., Jorgenson et al. 1987). For example, output elasticities, which are functions of the parameters, need to sum to one under CRS.

Flynn et al. (2019) also show that CRS can resolve identification problems with control-function identification. Namely, when lagged flexible inputs are used as instruments in the presence of markups the estimated markup distribution exhibits spurious skewness. Restrictions on returns to scale, such as CRS, resolve the identification problem and achieve more accurate markup estimates.

However, the CRS assumption is theoretically substantive and empirically underexplored. With CRS, marginal cost equals average cost and equation (9) shows that the markup maps one-to-one to the implied rate of economic profits. Some firm and industry data do suggest that returns to scale are, on average, close to constant (McAdam et al., 2024; Basu et al., 2006). However, any heterogeneity in returns to scale will be assigned to markups. Furthermore, research also suggests that variation in returns to scale, across industries and across time, can simultaneously rationalize long-run trends in factor shares and improve the measurement of misallocation (Ruzic and Ho, 2021).

[31] Because of firm-level measurement error, some practitioners average shares within an industry. But that returns us to the problem that true variations in output elasticity will be labeled as variations in markups.

Gross output or value added? Another common technology restriction that simplifies estimation is a value-added production function. If production is Leontief, so intermediates and value added are used in fixed proportions (perfect complements), we can write gross output Y as a function of intermediates M_{it}, a value-added function $F(\cdot)$, and a constant α:

$$Y_{it} = \min\{\alpha M_{it}, F(K_{it}, L_{it}, A_{it})\}. \tag{33}$$

By excluding intermediate inputs, value added reduces the dimensionality of the estimation problem. We do not need to separately identify elasticities for intermediates.

However, the strong Leontief assumption is probably not a realistic representation of technology. Otherwise, as Basu and Fernald (1995) and Basu and Fernald (1997) emphasize, value-added models implicitly assume that intermediate inputs are paid their marginal products, even when firms exercise market power. But markups drive a wedge between marginal products and factor payments, and some of the productive contribution of intermediates is incorrectly embedded in measured value added. Empirically, this misspecification can be substantial: Gandhi et al. (2017) find that imposing a value-added structure inflates measured productivity dispersion by a factor of five relative to a gross-output specification. Moreover, this value added specification assumes that intermediates are equally substitutable with capital and with labor. Ruzic (2024) provides evidence that intermediates disproportionately displace labor relative to capital—even for the aggregate economy—in a way that is inconsistent with value-added specifications of production.

On the whole, separating technology from market power requires careful judgment and reasoning. We want to be flexible enough with our technology specification to capture true shifts, especially over time, but not so flexible that it assumes away the possibility of market power. Recent work offers guidance: Foster et al. (2024) demonstrate how granular industry estimates can reverse aggregate trends, while Raval (2023) develops tests for production function stability. These advances help researchers navigate the flexibility-precision tradeoff. But these don't work well in micro-data cross-sections or short panels (De Loecker, 2011b). Finding such instruments in firm-level data is difficult.

The input demand equation (20) reveals exactly what instruments could theoretically work. Any shifter of input demand that doesn't appear in the production function is a potential instrument. Looking at the equation, valid instruments include input prices (w_{it}^M), the markup (μ_{it}), and the output price (p_{it})—but of course the latter two are unobserved and part of what we're trying to measure. Predetermined inputs (k_{it}, l_{it}) appear in both the demand equation and the production function, so they cannot serve as instruments. This

leaves input prices as the primary candidate. While firm-specific input prices can work when available (e.g., Doraszelski and Jaumandreu, 2013), they are often weak or potentially invalid if correlated with local demand shocks that also affect output prices (Gandhi et al., 2020).[32] Despite the limitations, instrument relevance is testable, and these approaches have seen renewed interest where input prices are strong instruments.

6 A Call to Arms

We began this review with a reference to a unifying FOC that highlights the close relationship between revenue shares, output elasticities and markups. Prominent papers have used this relationship to suggest that we live in a world of large and heterogeneous deviations from perfect competition. This research links these deviations to trends in the macroeconomy, ranging from slow growth to declining labor shares. Those findings serve as at least a "proof of concept" that markups could matter for major economic trends.

We end this review by using the same FOC to launch a three-pronged methodological call to arms. First, and most concretely, we propose that production-based markups be paired with a simple R^2 decomposition that quantifies how much of the variation in factor shares is explained by markups versus output elasticities. Such a decomposition helps assess the relative importance of demand and technology heterogeneity and provides a check on the credibility of key assumptions in implementing production-based markups. Second, we encourage systematic comparisons of production- and demand-based markups, the latter being more common in the IO literature. Understanding when the two approaches align would strengthen confidence in both. Third, we call for more work mapping firm-level markup heterogeneity into macroeconomic models. Progress requires sharper identification of the structural forces behind markups and richer analysis of when heterogeneity and production networks matter for aggregate outcomes.

6.1 Transparency: An R-squared for revenue shares

Our review has returned several times to the question raised by the central FOC: How much of the observed variation in revenue shares across firms and over time reflects demand-side forces like markups, and how much reflects differences in technology, such as output elasticities? Production-approach markups estimate output elasticities and infer markups as residuals.

[32] Even if a firm takes input prices as given, general equilibrium effects can invalidate them as instruments. Suppose all firms in a market get a positive productivity shock. If industry demand is not perfectly elastic, the market price for their output will fall, creating a correlation between the shock a_{it} and the price p_{it} that violates the exogeneity condition.

The more estimation restricts how much output elasticities can vary, the more variation in revenue shares gets shifted onto the markup. A priority for the literature is greater transparency on this front. Such transparency at the firm level complements our earlier suggestion (Section 4) to benchmark aggregate revenue shares against national accounts.

The R^2 of a simple linear regression can transparently quantify the extent to which revenue-share data is being explained by output elasticities versus by markups. Consider the following regression that can be estimated using ordinary least squares:

$$\ln s_{it}^X = \alpha + \beta \ln \mu_{it} + \varepsilon_{it},$$

with α the constant, β the slope coefficient and ε_{it} the error term. The R^2 of this regression quantifies the share of the total variation in $\ln s_{it}^X$ (the total sum of squares in the denominator) that can be linearly explained by the markup $\ln \mu_{it}$ (the residual sum of squares in the numerator). Moreover, since FOC (4) presents a log-linear relationship between revenue shares s_{it}^X, markups μ_{it} and the output elasticities γ_{it}^X:

$$\ln s_{it}^X = \ln \gamma_{it}^X - \ln \mu_{it}, \tag{34}$$

$1 - R^2$ captures the share of residual variation coming from $\ln \gamma_{it}^X$, after linearly projecting it onto $\ln \mu_i$. We can also add fixed effects (e.g., year, industry or year-industry) to the regression. The within R^2 and $1 - R^2$ then partition the variation of within-year, within-industry or within-industry-year revenue shares into those explained by the markups and those explained by the output elasticities.

Table 2 illustrates the R^2 decomposition using Compustat data and code from existing papers. The Markup columns report the R^2 from the above regression of markups on revenue shares (without and with industry-year fixed effects). The corresponding Output Elasticity columns report $1 - R^2$. Estimates correspond to the headline figure 1 from De Loecker et al. (2020), which uses Cost of Goods Sold (COGS) as the flexible input. We draw on their replication code and an online estimation toolbox from De Ridder et al. (2024).

The results in Table 2 suggest that the output elasticities explain remarkably little of the variation in COGS revenue shares. The lion's share loads on the residual claimant of FOC (4), the markup. Consider Figure 1 in De Loecker et al. (2020), which estimates (two-digit) sector-year Cobb-Douglas output elasticities. As an extreme benchmark, the two right-hand columns add industry-year fixed effects to regression (34). By construction, the output elasticities have no explanatory power (the R^2 is exactly zero), since a single output elasticity

Table 2: An R^2 Decomposition of Revenue-Share Variation: Output Elasticities vs Markups

	Full Sample		**Industry-Year**	
	Output Elasticity	Markup	Output Elasticity	Markup
De Loecker, Eeckhout & Unger (2020) *Replication File*				
Cobb-Douglas, Industry-Year	0.0102	0.9898	0	1
De Ridder, Grassi & Morzenti (2024) *Markup Toolbox: Translog, 2nd order*				
No interactions of inputs, Firm-Year	0.0400	0.9600	0.0058	0.9942
Interactions of inputs, Firm-Year	0.1929	0.8071	0.1002	0.8998

Note: Using Compustat data and different approaches to estimating output elasticities, we report the (within) R^2 from regressions of form $\ln s_{it}^X = \alpha + \beta \ln \mu_{it} + \varepsilon_{it}$ in the Markup columns. The Output Elasticity columns report 1 minus that R^2.

applies to all Compustat firms in a given industry-year. So 100% of the variation in revenue shares within an industry is attributed to markups.

What is striking is the extent to which the same pattern holds for the full sample (the first two columns). Time-varying output elasticities can soak up some variation in revenue shares. But these elasticities explain only 1 percent of the variation in the COGS revenue share, leaving markups to explain the remaining 99 percent. Similar results hold for our own calculations where we estimate industry-year elasticities using cost shares under the additional assumption of constant returns to scale. These results highlight the extent to which choices made in estimation can restrict the potential for output elasticities, and hence technology, to explain variation in revenue shares.

A potential way forward is to estimate output elasticities more flexibly. Translog production functions, which allow estimated coefficients to interact with firm-specific inputs, yield output elasticities that vary across firms and time.

The results in the bottom half of Table 2 suggest that even this translog approach continues to assign remarkably little revenue-share variation to output elasticities. In a version of the De Ridder et al. (2024) code that interacts production-function parameters with individual inputs and the inputs squared (but not with cross-input interactions), only about 1/2 percent of the within-industry-year variation in revenue shares is explained by output elasticities. Markups explain 99-1/2 percent, barely different than Cobb-Douglas. Even in the full sample, the explanatory power is only modestly different from the Cobb-Douglas

benchmark. In a second version of the translog production function—now allowing cross-input interactions—output elasticities account for between 0.1 and 0.2 of the revenue-share variation, and markups for *only* .80 to .90. Even with this additional flexibility in estimation, the output elasticities explain rather little of the variation in the data; meanwhile, the residual claimant that ensures FOC (4) holds—the markup—absorbs most of the variation.

Overall, these results raise an important intellectual question: how much of the variation in revenue shares across firms should we attribute to demand-side forces like markups and how much might plausibly reflect variation in technology, such as output elasticities? One important risk in answering that question is that the parameter that is inferred as a residual will always have greater explanatory power than the parameter that is estimated directly.

6.2 Validation: Stress-testing market power

Our second call to arms is also straightforward: researchers should look for opportunities to validate production-based markups.

First, production-based markups complement an extensive industrial-organization literature on demand-based markups. In principle, both measure the same object: the markup of price over marginal cost. But they do so through distinct implementations and assumptions. Comparing results in settings where researchers can apply both offers a valuable robustness check. When they diverge, the comparison can shed light on misspecification, data limitations, or the economic forces each method captures. When they align, we gain confidence in the empirical signal. At present, such cross-method validations remain uncommon—and that is a missed opportunity.

The demand approach derives markups from a model of consumer demand and firm pricing behavior. It is often tailored to a narrowly defined industry, such as the canonical study of the automobile market by Berry et al. (1995). Researchers specify a structural demand system—typically allowing for heterogeneous consumers and flexible substitution across differentiated products—and recover price elasticities from observed market shares and prices. Under an assumed model of firm conduct, often Bertrand competition, these elasticities yield markups through pricing first-order conditions. For instance, in the baseline case of a single-product firm facing constant-elasticity demand, the markup is a simple function of the absolute value of the demand elasticity: $\mu = |\eta|/(|\eta| - 1)$. Demand-side markups rely on rich market-level data and functional-form assumptions about preferences and competition. This approach enables detailed counterfactual and welfare analysis, but it can be sensitive to model misspecification and instrument choice.

Do demand and production approaches yield consistent markup estimates in practice? Recent studies offer mixed answers. Grieco et al. (2024) examine the U.S. automobile industry from 1980 to 2018 using both methods. Their demand-based estimates show markups declining moderately, while production-based estimates from De Loecker et al. (2020) show different levels and an upward trend. The authors attribute the divergence to demand-side improvements in product quality and variety that production methods may not fully capture—a key insight about what each approach measures. A more reassuring example comes from De Loecker and Scott (2016) study of the U.S. beer industry, where both methods deliver broadly similar mean markups with overlapping confidence intervals, at least for certain years and specifications.

Second, simulation studies help us understand when and why different methods succeed or fail. By specifying the data-generating process, researchers can evaluate how well production-based estimators recover known markups under various conditions. For example, section 5 mentioned the Monte Carlo simulations in De Ridder et al. (2024). Those simulations suggest that, at least in their model environment, using revenue data led to biased markup estimates, but trends and dispersion remain informative. These kinds of controlled assessments map out the conditions required for reliable estimation.

Third, quasi-experimental validation offers a path forward. When mergers, trade shocks, or regulatory changes generate plausibly exogenous shifts in competition, we can test whether markup estimates detect these changes. Miller et al. (2017) exploit the MillerCoors joint venture to show that demand-based markups correctly capture the resulting price increases. Carrillo et al. (2023) use a different approach, showing how exogenous demand shocks from public procurement contracts for construction services in Ecuador identify features of the marginal product distribution, testing for misallocation and quantifying welfare losses. Majerovitz and Hughes (2025) studies misallocation in Sri Lanka's construction sector using quasi-experimental variation from government procurement contracts. These settings can not only help validate production-based methods but, when combined with structural estimation, generate insights about firm behavior under imperfect competition impossible with reduced-form approaches alone.

Validation should become a standard component of markup estimation research. This does not mean every paper must implement every validation approach—that would be neither feasible nor productive. The production approach already demands substantial cross-area expertise and institutional knowledge, as we've highlighted in this review. Research benefits from specialization: Some researchers can push the frontiers of measurement; others can focus on validation through demand modeling, experimental design, or simulation methods.

Studies with access to rich data should compare multiple approaches. Papers introducing new methods should show performance in simulations. Empirical applications should seek quasi-experimental corroboration where available. Through this collective effort, we can build confidence in our measurements of market power and clarify what drives cross-method differences. The tools exist; we need to use them more systematically.

6.3 Aggregation: Macro implications of micro markups

Our final call to arms is to bridge the micro and the macro. Despite the conceptual and empirical challenges outlined earlier, the strength of the production approach lies in its ability to deliver granular markup estimates. To guide future work on how such estimates can inform an economy's productivity, welfare, and resilience to shocks, we highlight three research questions to advance this agenda. First, what economic primitives underlie measured markups? Second, how much heterogeneity is needed to account for macroeconomic outcomes? Third, how do markups interact with—and potentially shape—the economy's network structure?

First, for many questions in macroeconomics—as in industrial organization—it is not enough to establish that markups exist. To use them for counterfactual analysis, we need to know why firms are charging them. Markups can arise from very different primitives: consumer demand elasticities, firm conduct in oligopoly, the presence of fixed costs, or increasing returns to scale. The production approach, by design, abstracts from these structural underpinnings; it delivers equilibrium objects, not the reasons behind them. This might make the production approach powerful for documenting patterns in the data, but leaves macroeconomists without guidance for how to embed markups into structural models where the source of markup rents determines welfare and growth implications.

Quantifying the different economic forces behind markups matters: if we erroneously assign the entirety of a production-based markup to a single markup mechanism, we risk obtaining misleading predictions from the fuller model. For instance, markups can represent pure economic profits—market-power rents that reduce efficiency and justify antitrust concern. Or, they can be a necessary byproduct of cost recovery or scale economies: firms with high fixed costs or increasing returns may charge prices above marginal cost without generating excess profits. If we model production-based markups as though they all stem from one source—say, barriers to entry—when in fact they are partly compensating for fixed costs or returns to scale, our counterfactual predictions risk being misleading. Moving forward on these questions requires treating production parameters themselves, not just the markup

estimates, as central objects of interest.[33]

Future research can help quantify the relative prevalence of "good" and "bad" markups from the long-standing macro-IO debates. Some rents reflect entry barriers or inefficient firm conduct that depresses output and growth (e.g., Bresnahan 1989 and Berry et al. 2019). Others reflect Schumpeterian incentives for innovation: Aghion et al. (2005) find a U-shaped relationship between innovation and market power, balancing positive incentives with negative inefficiency effects (see Gilbert 2006 for a review). Klette and Kortum (2004) model markups as the return to innovation, as does Peters (2020). More recently, Autor et al. (2020) argue that the rise of superstar firms with high markups reflects efficient scale and innovation. Whereas Aghion et al. (2023) argue that comparable markups can instead hinder growth if they result from process inefficiencies and R&D misallocation, generating "bad" rents.[34] Quantifying the prevalence of different markup motivations can help researchers understand how to rationalize production-based markups in macroeconomic models.

Second, how much heterogeneity matters for macro outcomes is a central open question. On the household side, macroeconomists debate whether simplified two-agent New Keynesian (TANK) models are sufficient to capture meaningful heterogeneity in consumption responses, or whether richer heterogeneous-agent New Keynesian (HANK) models are required, despite being less tractable (Galí et al., 2007; Kaplan et al., 2018). A parallel question arises for firms: do we need the full distribution of markups and production parameters to analyze macro outcomes, or can simplified representative structures get us "close enough"? This issue is especially pressing because tractability pushes models toward parsimony, but ignoring key dimensions of heterogeneity risks distorting counterfactuals.

Thus far, theoretical work has not resolved how to connect micro estimates to representative macro models. Standard business-cycle and growth models typically rely on a representative firm producing value added (manufactured by capital and labor, but not intermediates). The mapping from production-based gross-output markups to such a representative firm is unclear. For example, Rotemberg and Woodford (1995, 1999) show that strong assumptions are needed to translate gross-output production with market power into value-added production. Basu and Fernald (1997) highlight that the appropriate aggregation of firm-level elasticities and markups depends on model specifics, including how reallocations across firms occur over the cycle. They provide a two-firm example in which the correct "aggregate" pa-

[33] Indeed, in the 1990s production-approach literature, markups and returns to scale were treated as equally relevant for modeling. For example, Basu and Fernald (1997) emphasized returns to scale because of the focus of some models at the time, but their estimates mapped to markups via equation (9).

[34] Aghion et al. (2023) and De Ridder (2024) argue that good rents can morph into bad rents as technological winners create barriers to entry, leading ultimately to reduced innovation and growth.

rameter could be either the simple average of micro estimates or the parameter that would be recovered from aggregate time series—depending on non-production details of household behavior and equilibrium allocation. To date, there is no general answer to how much micro heterogeneity needs to be preserved when calibrating representative-firm models.[35]

Ongoing empirical debates underscore why this heterogeneity question is so important. Revenue shares vary widely across firms, and these differences are not random. Large, expanding firms tend to have low labor shares, suggesting potentially important macro consequences (Autor et al., 2020; Kehrig and Vincent, 2021). In rationalizing these patterns, workhorse models often feature firm heterogeneity in productivity (Melitz, 2003) and demand elasticities (Atkeson and Burstein, 2008), but typically assume common production technologies within industries. If empirical evidence of heterogeneous output elasticities proves robust, existing modeling frameworks will need to evolve. Moreover, the choice of how to aggregate matters: Edmond et al. (2023) show that sales-weighted markups rise much more steeply than cost-weighted ones, with the gap reflecting allocative inefficiency from dispersion. These results highlight that the way heterogeneity is summarized—whether through cost-weighted, sales-weighted, or distributional statistics—shapes the macro conclusions we draw.

Third, more research should help evaluate how the economy's network structure shapes the welfare consequences of markups. A long tradition of input–output analysis, from Leontief (1941) to more recent work reviewed by Carvalho and Tahbaz-Salehi (2019) and Baqaee and Rubbo (2023), emphasizes that the economy's network of intermediate flows can amplify or attenuate distortions. What matters is not only the level and dispersion of markups but also where in the network they arise and how they propagate through production chains.

Notably, upstream and downstream markups differ fundamentally in how they propagate through the economy. At one extreme, markups may simply redistribute income between wages and profits without real effects, particularly if they fall entirely downstream on final consumption goods and labor is supplied inelastically. At the other extreme, when markups are applied upstream to intermediate inputs, they act like taxes on production, distorting relative input use and depressing aggregate productivity. This misallocation pushes the economy inside its production-possibilities frontier, violating the (Diamond and Mirrlees, 1971) production-efficiency theorem. Theoretical and quantitative work by Basu (1995), Basu and Fernald (2002), Bigio and La'O (2020), and Baqaee and Farhi (2020) formalizes

[35]Growth accounting has long emphasized microeconomic heterogeneity in explaining macro aggregates. Basu and Fernald (1997, 2002) added heterogeneity in markups and returns to scale to the accounting in Jorgenson et al. (1987). The misallocation literature also emphasizes micro heterogeneity. Baqaee and Rubbo (2023) review the more recent literature.

these effects, showing how markups on upstream goods differ in their consequences from markups applied at the final-goods stage.

At the same time, contracting arrangements shape how strongly these distortions bite. In models with spot-market transactions, "double marginalization" compounds markups along supply chains, magnifying inefficiencies. But if upstream and downstream firms bargain efficiently—as often happens in long-term supplier relationships—markups need not generate misallocation, even if measured markups appear high. This observation suggests that understanding the macroeconomic consequences of markup heterogeneity requires integrating both network structure and institutional detail. Future research can help clarify when markups primarily redistribute income, when they distort production, and how to incorporate these differences into tractable macro models.

7 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, they absorb whatever the model, the data, or the econometrics fail to capture. The implementation requires assumptions that are not innocuous. Researchers make a series of choices, each with its own risks.

The evidence so far requires humility. Some datasets and specifications show markups rising sharply. Others do not. As we discussed, the discrepancies may reflect non-markup frictions, specification, data, or econometric choices. The variation tells us something important: production-based markups depend on the validity of the auxiliary assumptions we impose. To move forward, we need transparency, validation, robustness, and better data.

What should macroeconomists do with the evidence? We should acknowledge how much remains uncertain, while pressing ahead on both theory and measurement. On the theoretical 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, in order to guide theorizing. Measurement leaves us with residuals. The challenge is to combine institutional knowledge and economic theory to show what those residuals mean for growth, welfare, and policy.

References

Acemoglu, D. and A. Tahbaz-Salehi (2025). The macroeconomics of supply chain disruptions. *Review of Economic Studies 92*(2), 656–695.

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007). Econometric tools for analyzing market outcomes. *Handbook of Econometrics 6*, 4171–4276.

Ackerberg, D., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–51.

Ackerberg, D. and J. De Loecker (2024). Production function identification under imperfect competition. Discussion Paper Dp19640, Cepr.

Aghion, P., A. Bergeaud, T. Boppart, P. Klenow, and H. Li (2023). A theory of falling growth and rising rents. *Review of Economic Studies 90*(6), 2675–2702.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005). Competition and innovation: An inverted-u relationship. *Quarterly Journal of Economics 120*(2), 701–728.

Ali, A., S. Klasa, and E. Yeung (2008). The limitations of industry concentration measures constructed with Compustat data: Implications for finance research. *Review of Financial Studies 22*(10), 3839–3871.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies 58*(2), 277–297.

Asker, J., A. Collard-Wexler, and J. De Loecker (2014). Dynamic inputs and resource (mis) allocation. *Journal of Political Economy 122*(5), 1013–1063.

Atkeson, A. and A. Burstein (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review 98*(5), 1998–2031.

Autor, D., D. Dorn, L. Katz, C. Patterson, and J. Van Reenen (2020). The fall of the labor share and the rise of superstar firms. *Quarterly Journal of Economics 135*(2), 645–709.

Baily, M., C. Hulten, and D. Campbell (1992). Productivity dynamics in manufacturing plants. *Brookings Papers on Economic Activity 1992*(1), 187–267.

Bain, J. (1951). Relation of profit rate to industry concentration: American manufacturing, 1936–1940. *Quarterly Journal of Economics 65*(3), 293–324.

Baqaee, D. and E. Farhi (2020). Productivity and misallocation in general equilibrium. *Quarterly Journal of Economics 135*(1), 105–163.

Baqaee, D. and E. Rubbo (2023). Micro propagation and macro aggregation. *Annual Review of Economics 15*(1), 91–123.

Baqaee, D. R., E. Farhi, and K. Sangani (2024). The darwinian returns to scale. *Review of Economic Studies 91*(3), 1373–1405.

Barkai, S. (2020). Declining labor and capital shares. *Journal of Finance 75*(5), 2421–63.

Basu, S. (1995). Intermediate goods and business cycles: Implications for productivity and welfare. *American Economic Review 85*(3), 512–531.

Basu, S. (2019). Are price-cost markups rising in the United States? A discussion of the evidence. *Journal of Economic Perspectives 33*(3), 3–22.

Basu, S. and J. Fernald (1995). Are apparent productive spillovers a figment of specification error? *Journal of Monetary Economics 36*(1), 165–188.

Basu, S. and J. Fernald (1997). Returns to scale in US production: Estimates and implications. *Journal of Political Economy 105*(2), 249–83.

Basu, S. and J. Fernald (2002). Aggregate productivity and aggregate technology. *European Economic Review 46*(6), 963–91.

Basu, S., J. Fernald, and M. Kimball (2006). Are technology improvements contractionary? *American Economic Review 96*(5), 1418–1448.

Basu, S., J. Fernald, and M. Shapiro (2001). Productivity growth in the 1990s: technology, utilization, or adjustment? *Carnegie-Rochester Conference Series on Public Policy 55*(1), 117–165.

Basu, S. and C. House (2016). Allocative and remitted wages: New facts and challenges for Keynesian models. In J. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics, Volume 2A*, Handbook of Macroeconomics, pp. 297–354. Elsevier.

Becker, R., J. Haltiwanger, R. Jarmin, S. Klimek, and D. Wilson (2006). Micro and macro data integration: The case of capital. In D. Jorgenson, J. S. Landefeld, and W. Nordhaus (Eds.), *A New Architecture for the US National Accounts*, pp. 541–610. University of Chicago Press.

Berger, D., K. Herkenhoff, and S. Mongey (2022). Labor market power. *American Economic Review 112*(4), 1147–1193.

Berman, E. and A. Jaffe (2024). Zvi Griliches (1930–1999). In R. Cord (Ed.), *The Palgrave Companion to Harvard Economics*, pp. 573–596. Palgrave Macmillan.

Berndt, E. and D. Wood (1975). Technology, prices, and the derived demand for energy. *Review of Economics and Statistics 57*(3), 259–268.

Berndt, E. R. and M. A. Fuss (1986). Productivity measurement with adjustments for variations in capacity utilization and other forms of temporary equilibrium. *Journal of Econometrics 33*(1), 7–29.

Berry, S., M. Gaynor, and F. S. Morton (2019). Do increasing markups matter? lessons from empirical industrial organization. *Journal of Economic Perspectives 33*(3), 44–68.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

Bigio, S. and J. La'O (2020). Distortions in production networks. *Quarterly Journal of Economics 135*(4), 2187–2253.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–143.

Blundell, R. and S. Bond (2000). GMM estimation with persistent panel data: An application to production functions. *Econometric Reviews 19*(3), 321–40.

Bond, S., A. Hashemi, G. Kaplan, and P. Zoch (2021). Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics 117*, 511–33.

Bond, S. and M. Söderbom (2005). Adjustment costs and the identification of cobb douglas production functions. Technical Report W05/04, Institute for Fiscal Studies, London.

Brand, J. (2019). Estimating productivity and markups under imperfect competition. Technical report, Working paper.

Bresnahan, T. F. (1989). Empirical studies of industries with market power. *Handbook of industrial organization 2*, 1011–1057.

Burnside, C., M. Eichenbaum, and S. Rebelo (1995). Capital utilization and returns to scale. In B. Bernanke and J. Rotemberg (Eds.), *NBER Macroeconomics Annual 1995, Volume 10*, NBER Chapters, Chapter 4, pp. 67–124. MIT Press.

Caballero, R. and R. Lyons (1992). External effects in U.S. procyclical productivity. *Journal of Monetary Economics 29*(2), 209–225.

Carrillo, P., D. Donaldson, D. Pomeranz, and M. Singhal (2023). Misallocation in firm production: A nonparametric analysis using procurement lotteries. Technical report, National Bureau of Economic Research.

Carvalho, V. M. and A. Tahbaz-Salehi (2019). Production networks: A primer. *Annual Review of Economics 11*(1), 635–663.

Choi, J., A. Levchenko, D. Ruzic, and Y. Shim (2024). Superstars or supervillains? Large firms in the South Korean growth miracle. Working Paper 32648, National Bureau of Economic Research.

Collard-Wexler, A. and J. De Loecker (2016). Production function estimation and capital measurement error. Working paper, National Bureau of Economic Research.

Cooper, R., J. Haltiwanger, and J. Willis (2024). Declining responsiveness at the establishment level: Sources and productivity implications. NBER Working Papers 32130, National Bureau of Economic Research.

Corrado, C., C. Hulten, and D. Sichel (2009). Intangible capital and US economic growth. *Review of income and wealth 55*(3), 661–685.

Crouzet, N. and J. Eberly (2019). Understanding weak capital investment: The role of market concentration and intangibles. Technical report, National Bureau of Economic Research.

Davis, S., J. Haltiwanger, R. Jarmin, and J. Miranda (2006). Volatility and dispersion in business growth rates: Publicly traded versus privately held firms. In *NBER Macroeconomics Annual*, Volume 21, pp. 107–180. MIT Press.

De Loecker, J. (2011a). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica 79*(5), 1407–1451.

De Loecker, J. (2011b). Recovering markups from production data. *International Journal of Industrial Organization 29*(3), 350–355.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *Quarterly Journal of Economics 135*(2), 561–644.

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

De Loecker, J. and P. Scott (2016). Estimating market power evidence from the US brewing industry. Working Paper 22957, National Bureau of Economic Research.

De Loecker, J. and F. Warzynski (2012). Markups and firm-level export status. *American Economic Review 102*(6), 2437–71.

De Ridder, M. (2024). Market power and innovation in the intangible economy. *American Economic Review 114*(1), 199–251.

De Ridder, M., B. Grassi, and G. Morzenti (2024). The hitchhiker's guide to markup estimation. POID Working Papers 063, Centre for Economic Performance, LSE.

Decker, R. and J. Williams (2023). A note on industry concentration measurement. FEDS Notes 2023-02-03, Board of Governors of the Federal Reserve System (U.S.).

Demirer, M. (2025). Production function estimation with factor-augmenting technology: An application to markups. Phd thesis, Mit.

Demsetz, H. (1973). Industry structure, market rivalry, and public policy. *Journal of Law and Economics 16*(1), 1–9.

Dhyne, E., A. K. Kikkawa, and G. Magerman (2022). Imperfect competition in firm-to-firm trade. *Journal of the European Economic Association 20*(5), 1933–1970.

Diamond, P. A. and J. A. Mirrlees (1971). Optimal taxation and public production i: Production efficiency. *American Economic Review 61*(1), 8–27.

Dobbelaere, S. and J. Mairesse (2013). Panel data estimates of the production function and product and labor market imperfections. *Journal of Applied Econometrics 28*(1), 1–46.

Doraszelski, U. and J. Jaumandreu (2013). R&D and productivity: Estimating endogenous productivity. *Review of Economic Studies 80*(4), 1338–1383.

Doraszelski, U. and J. Jaumandreu (2019). Using cost minimization to estimate markups. Working Paper Dp14114, Centre for Economic Policy Research.

Dunne, T., M. Roberts, and L. Samuelson (1988). Patterns of firm entry and exit in US manufacturing industries. *RAND Journal of Economics 19*(4), 495–515.

Edmond, C., V. Midrigan, and D. Y. Xu (2023). How costly are markups? *Journal of Political Economy 131*(1), 4–42.

Elsby, M., B. Hobijn, and A. Şahin (2013). The decline of the US labor share. *Brookings Papers on Economic Activity 2013*(2), 1–63.

Fernald, J. (2014). A quarterly, utilization-adjusted series on total factor productivity. Working Paper Series 2012-19, Federal Reserve Bank of San Francisco.

Fernald, J. (2024). Dale W. Jorgenson (1933–2022). In R. Cord (Ed.), *The Palgrave Companion to Harvard Economics*, pp. 597–631. Palgrave Macmillan.

Fernald, J., R. Inklaar, and D. Ruzic (2025). The productivity slowdown in advanced economies: Common shocks or common trends? *Review of Income and Wealth 71*(1), e12690.

Fernald, J. and E. Piga (2023). Comment on: Bottlenecks: Sectoral imbalances and the US productivity slowdown. In *NBER Macroeconomics Annual*, Volume 38 of *NBER Chapters*. University of Chicago Press.

Fisher, F. and J. McGowan (1983). On the misuse of accounting rates of return to infer monopoly profits. *American Economic Review 73*(1), 82–97.

Flynn, Z., J. Traina, and A. Gandhi (2019). Measuring markups with production data. Working Paper 3358472, Social Science Research Network.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review 98*(1), 394–425.

Foster, L. S., J. C. Haltiwanger, and C. Tuttle (2024). Rising markups or changing technology? Working Paper 30491, National Bureau of Economic Research.

Fox, J. and V. Smeets (2011). Does input quality drive measured differences in firm productivity? *International Economic Review 52*(4), 961–989.

Galí, J. (1999). Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations? *American Economic Review 89*(1), 249–271.

Galí, J., J. D. López-Salido, and J. Vallés (2007). Understanding the effects of government spending on consumption. *Journal of the European Economic Association 5*(1), 227–270.

Gandhi, A., S. Navarro, and D. Rivers (2017). How heterogeneous is productivity? A comparison of gross output and value added. Technical Report 201727, Center for Human Capital and Productivity, University of Western Ontario.

Gandhi, A., S. Navarro, and D. Rivers (2020). On the identification of gross output production functions. *Journal of Political Economy 128*(3), 810–53.

Gilbert, R. (2006). Looking for mr. schumpeter: Where are we in the competition–innovation debate? *Journal of Economic Literature 44*(3), 537–583.

Grieco, P., S. Li, and H. Zhang (2016). Production function estimation with unobserved input price dispersion. *International Economic Review 57*(2), 665–690.

Grieco, P., C. Murry, and A. Yurukoglu (2024). The evolution of market power in the US automobile industry. *Quarterly Journal of Economics 139*(2), 1201–1253.

Griliches, Z. and J. Mairesse (1998). Production functions: The search for identification. In S. Strøm (Ed.), *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pp. 169–203. Cambridge University Press.

Griliches, Z. and V. Ringstad (1971). *Economies of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data*. North-Holland Publishing Company.

Grunfeld, Y. and Z. Griliches (1960). Is aggregation necessarily bad? *Review of Economics and Statistics 42*(1), 1–13.

Gutiérrez, G. and T. Philippon (2017). Investmentless growth: An empirical investigation. *Brookings Papers on Economic Activity 2017*(2), 89–190.

Hall, R. (1980). Employment fluctuations and wage rigidity. *Brookings Papers on Economic Activity 11*(1, Tenth Anniversary Issue), 91–142.

Hall, R. (1986). Market structure and macroeconomic fluctuations. *Brookings Papers on Economic Activity 2*, 285–322.

Hall, R. (1988). The relation between price and marginal cost in US industry. *Journal of Political Economy 96*(5), 921–47.

Hall, R. (1990). Invariance properties of solow's productivity residual. In P. Diamond (Ed.), *Growth/ Productivity/Unemployment: Essays to Celebrate Robert Solow's Birthday*, pp. 71–112. MIT Press.

Hall, R. and D. Jorgenson (1967). Tax policy and investment behavior. *American Economic Review 57*(3), 391–414.

Hashemi, A., I. Kirov, and J. Traina (2022). The production approach to markup estimation often measures input distortions. *Economics Letters 217*, 110673.

Hsieh, C.-T. and P. Klenow (2009). Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics 124*(4), 1403–48.

Hulten, C. (1986). Productivity change, capacity utilization, and the sources of efficiency growth. *Journal of Econometrics 33*(1), 31–50.

Jarosch, G., J. S. Nimczik, and I. Sorkin (2024). Granular search, market structure, and wages. *The Review of Economic Studies 91*(6), 3569–3607.

Jorgenson, D., F. Gollop, and B. Fraumeni (1987). *Productivity and US Economic Growth*. Harvard University Press.

Jorgenson, D. and Z. Griliches (1967). The explanation of productivity change. *Review of Economic Studies 34*(99), 249–280.

Kaplan, G., B. Moll, and G. L. Violante (2018). Monetary policy according to hank. *American Economic Review 108*(3), 697–743.

Karabarbounis, L. and B. Neiman (2014). The global decline of the labor share. *Quarterly Journal of Economics 129*(1), 61–103.

Karabarbounis, L. and B. Neiman (2019). Accounting for factorless income. *NBER Macroeconomics Annual 33*(1), 167–228.

Kasahara, H. and Y. Sugita (2020). Nonparametric identification of production function, total factor productivity, and markup from revenue data. Working paper, arXiv.

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

Keil, J. (2017). The trouble with approximating industry concentration from Compustat. *Journal of Corporate Finance 45*(C), 467–479.

Kim, K. I., A. Petrin, and S. Song (2016). Estimating production functions with control functions when capital is measured with error. *Journal of Econometrics 190*(2), 267–279.

Kirov, I., P. Mengano, and J. Traina (2025). Measuring markups with revenue data. Working paper, Social Science Research Network.

Kirov, I. and J. Traina (2022). Labor market power and technological change in US manufacturing. Phd thesis, University of Chicago.

Klette, T. J. (1999). Market power, scale economies and productivity: Estimates from a panel of establishment data. *Journal of Industrial Economics 47*(4), 451–476.

Klette, T. J. and Z. Griliches (1996). The inconsistency of common scale estimators when output prices are unobserved and endogenous. *Journal of Applied Econometrics 11*(4), 343–61.

Klette, T. J. and S. Kortum (2004). Innovating firms and aggregate innovation. *Journal of Political Economy 112*(5), 986–1018.

Kudlyak, M. (2024). How cyclical is the user cost of labor? *Journal of Economic Perspectives 38*(2), 159–80.

Lenzu, S., D. Rivers, and J. Tielens (2023). Financial shocks, productivity, and prices. Working Paper Ssrn 3442156, Social Science Research Network.

Leontief, W. (1941). *The Structure of American Economy, 1919–1929: An Empirical Application of Equilibrium Analysis*. Harvard University Press.

Lev, B. (2001). *Intangibles: Management, Measurement, and Reporting*. Brookings Institution Press.

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies 70*(2), 317–41.

Liu, E. and A. Tsyvinski (2024). A dynamic model of input–output networks. *Review of Economic Studies 91*(6), 3608–3644.

Ližal, L. and K. Galuščák (2012). The impact of capital measurement error correction on firm-level production function estimation. Technical Report 1026, William Davidson Institute, University of Michigan.

Macchiavello, R. and A. Morjaria (2023). Relational contracts: recent empirical advancements and open questions. LSE Research Online Documents on Economics 123003, London School of Economics and Political Science, LSE Library.

Majerovitz, J. and D. Hughes (2025). Measuring misallocation with experiments. Technical report, Cepr.

Manning, A. (2003). *Monopsony in Motion: Imperfect Competition in Labor Markets*. Princeton University Press.

Marschak, J. and W. Andrews (1944). Random simultaneous equations and the theory of production. *Econometrica 12*(3/4), 143–205.

McAdam, P., P. Meinen, C. Papageorgiou, and P. Schulte (2024). Returns to scale: New evidence from administrative firm-level data. Working paper, Deutsche Bundesbank.

Melitz, M. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica 71*(6), 1695–725.

Miller, N. H., M. Remer, C. Ryan, and G. Sheu (2017). Understanding the price effects of the MillerCoors joint venture. *Econometrica 85*(6), 1763–1791.

Mundlak, Y. (1961). Empirical production function free of management bias. *Journal of Farm Economics 43*(1), 44–56.

Olley, S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–97.

Ornaghi, C. (2006). Assessing the effects of measurement errors on the estimation of production functions. *Journal of Applied Econometrics 21*(6), 879–891.

Peters, M. (2020). Heterogeneous markups, growth, and endogenous misallocation. *Econometrica 88*(1), 305–43.

Raval, D. (2023). Testing the production approach to markup estimation. *Review of Economic Studies 90*(5), 2592–2611.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics 11*(4), 707–20.

Restuccia, D. and R. Rogerson (2017). The causes and costs of misallocation. *Journal of Economic Perspectives 31*(3), 151–74.

Robinson, J. (1933). *The Economics of Imperfect Competition* (2 ed.). Macmillan.

Rogerson, R., R. Shimer, and R. Wright (2005). Search-theoretic models of the labor market: A survey. *Journal of Economic Literature 43*(4), 959–988.

Romer, P. (1990). Endogenous technological change. *Journal of Political Economy 98*(5, Part 2), S71–s102.

Rosen, S. (1985). Implicit contracts: A survey. *Journal of Economic Literature 23*(3), 1144–1175.

Rotemberg, J. and M. Woodford (1995). Dynamic general equilibrium models with imperfectly competitive product markets. In T. Cooley (Ed.), *Frontiers of Business Cycle Research*. Princeton University Press.

Rotemberg, J. and M. Woodford (1999). The cyclical behavior of prices and costs. In J. Taylor and M. Woodford (Eds.), *Handbook of Macroeconomics*, Volume 1, Chapter 16, pp. 1051–1135. North-Holland.

Rubens, M. (2023). Market structure, oligopsony power, and productivity. *American Economic Review 113*(9), 2382–2410.

Ruzic, D. (2024). The factor bias of external inputs: Implications for substitution between capital and labor. Working paper, Insead.

Ruzic, D. and S.-J. Ho (2021). Returns to scale, productivity measurement, and trends in U.S. manufacturing misallocation. *Review of Economics and Statistics 103*(1), 23–37.

Sato, K. (1967). A two-level constant-elasticity-of-substitution production function. *Review of Economic Studies 34*(2), 201–218.

Schmalensee, R. (1989). Inter-industry studies of structure and performance. *Handbook of Industrial Organization 2*, 951–1009.

Solow, R. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics 39*(3), 312–20.

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature 49*(2), 326–65.

Traina, J. (2018). Is aggregate market power increasing? Production trends using financial statements. Working paper, Stigler Center.

van Heuvelen, G. H., L. Bettendorf, and G. Meijerink (2021). Markups in a dual labour market: The case of the netherlands. *International Journal of Industrial Organization 77*(C), None.

Yeh, C., C. Macaluso, and B. Hershbein (2022). Monopsony in the US labor market. *American Economic Review 112*(7), 2099–2138.