Conflicting Markup Estimates and Cost Minimization: The Semi-Nonparametric Translog Model

Jibin Jose and GianCarlo Moschini*

November 11, 2026

Abstract

The production approach estimates markups from output elasticities of flexible inputs and their cost shares in revenue. Recent studies document that markups recovered from different inputs frequently disagree. We show that this conflict arises because standard estimation does not fully exploit the maintained cost-minimization assumption. We develop a semi-nonparametric framework that embeds cost-minimization conditions into estimation, while retaining Hicks-neutral productivity. Markups are invariant across inputs whenever estimation enforces consistency of production function parameters with cost-minimization conditions, whether directly, as here, or via factor-augmenting productivity. We illustrate these findings with manufacturing data from Colombia, Chile, and the United States.

Keywords: Markups, Production function estimation, Cost minimization, Semi-nonparametric estimation

JEL Codes: D24, L11, C14, D22

*Jibin Jose (jjose@iastate.edu) is a Ph.D. candidate, and GianCarlo Moschini (moschini@iastate.edu) is Charles F. Curtiss Distinguished Professor and the Pioneer Endowed Chair in Science and Technology Policy, both in the Department of Economics at Iowa State University. We thank James Traina and participants in the Labor-Public Workshop at Iowa State University for helpful feedback. We acknowledge the use of Claude (Anthropic) for coding assistance and feedback on mathematical derivations. The authors have no conflicts of interest to disclose.

1 Introduction

Markups—the ratio of price to marginal cost—are central to evaluating market power and its policy implications. The production approach to markup estimation, originated in Hall (1988) and revived by De Loecker and Warzynski (2012), has gained considerable prominence in recent years because of its conceptual simplicity. It eschews the need to model the demand side—as in standard structural Industrial Organization (IO) models—and recovers markups simply as the ratio of a flexible input's output elasticity to its cost share in revenue. The only required condition is that firms minimize costs, an uncontroversial assumption that arguably applies to a broad set of firms' objective functions and forms of market competition. Estimating markups thus reduces to estimating production function parameters, a challenge that an extensive literature has sought to address (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Blundell and Bond, 1998, 2000). Combined with the increasing availability of production microdata, this approach has been applied broadly across firms, industries, and countries, generating a burgeoning empirical literature on firm-level markups and market power.

Recent evidence, however, has highlighted several unresolved issues. Studies applying it to broadly similar data have reached different conclusions about the levels and trends of markups. For instance, De Loecker et al. (2020) documents a sharp rise in markups in the United States over recent decades, while Foster et al. (2022) and Traina (2018) report markedly muted patterns. The use of financial data in lieu of actual production data to estimate markups has also come under scrutiny (Bond et al., 2021; De Ridder et al., 2026). A particularly puzzling feature is documented by Raval (2023b), who shows that markups estimated from different flexible inputs—such as labor and materials—differ systematically within the same dataset, often implying opposite trends. Doraszelski and Jaumandreu (2019) similarly document that, in the Spanish data they study, exporters appear to charge higher markups than non-exporters when markups are computed from labor, but lower markups

when computed from materials. Such divergence is inconsistent with the framework's main stated assumption: cost minimization implies that any flexible input should yield exactly the same markup.

A possible explanation for these perplexing results, suggested by ongoing work, relates to the treatment of firm-level productivity. Many models, including Raval (2023b), postulate Hicks-neutral productivity, a setting that provides considerable benefits at the estimation stage. Indeed, recent work shows that models that allow for biased technical change, specifically labor-augmenting productivity, appear to yield estimated markups that are consistent across alternative flexible inputs (Doraszelski and Jaumandreu, 2019; Raval, 2023a; Demirer, 2026).

In this paper, we re-examine this puzzling empirical finding of conflicting markup estimates and propose an alternative and complementary explanation. We start by noting that the standard production approach with Hicks-neutral productivity (De Loecker and Warzynski, 2012), as implemented by Raval (2023b), contains nothing that requires the observed data and the parameterized production function to be jointly consistent with the maintained assumption of cost minimization. Specifically, the optimality conditions of cost minimization explicitly link input cost shares to the structure of the production function, but this relationship is not reflected in the estimation process. Nor can this property be readily incorporated. With two or more flexible inputs, cost minimization ties the ratio of output elasticities to the ratio of observed input costs. A production function with a fixed set of parameters cannot generally match a ratio that takes a different value at every firm-year observation in the data.

We propose a novel procedure to embed the cost-minimization conditions directly into a production function estimand while maintaining Hicks-neutral productivity. We start with a general flexible representation of the production function expressed in terms of production elasticities—the building blocks of the markup estimation procedure. These production elasticities are then expressed in terms of a parameterized scale elasticity and observed cost

shares of variable inputs, so as to maintain the desired cost-minimization conditions. Put differently, the ratio of output elasticities remains unparameterized such that cost minimization can be met by observed cost shares at each data point. This approach yields a new parameterization—the semi-nonparametric translog (SNTL)—which, by construction, produces identical markup estimates across flexible inputs. We illustrate these results using the Colombian Annual Manufacturing Survey data, with additional evidence from Chile and the United States—all among the datasets used by Raval (2023b) to document the problem of conflicting markups.

This invariance result provides a deeper perspective on the puzzle resolution that postulates a factor-augmenting productivity process—the estimation strategy required for such a setting recovers factor-augmenting productivity by inverting the ratio of cost-minimizing first-order conditions, precisely those the standard approach does not impose in estimation. It follows that, insofar as the invariance of markup estimates is concerned, embedding cost minimization into estimation suffices on its own.

While this paper identifies a source of conflicting markup estimates, some limitations remain. First, as in Raval (2023b) and most of this literature, our empirical applications use deflated revenue rather than physical output, which Bond et al. (2021) and De Ridder et al. (2026) show affects the informativeness of production-based markups. The SNTL inherits this concern in the same form as the standard approach; the comparison between the two estimators is unaffected, as both use the same data. Second, we maintain throughout that firms take the prices of flexible inputs as given. Wedges in the first-order conditions, such as monopsony power or adjustment costs, would distort the cost shares the SNTL takes as data, as they would under any production-based approach.[2]

This paper contributes to a growing literature that critically examines the production approach to markup estimation (Bond et al., 2021; De Ridder et al., 2026; Fernald et al., 2026; Kirov and Traina, 2023; Kirov et al., 2026; Hashemi et al., 2022; Flynn et al., 2019;

[2]We return to input-market wedges in the context of discussing labor-augmenting productivity in section 7.

Brand, 2019; Foster et al., 2022; Ackerberg and De Loecker, 2021; Jaumandreu, 2024). In particular, we contribute to the strand concerned with conflicting markups and their explanations (Raval, 2023b; Doraszelski and Jaumandreu, 2019, 2021; Raval, 2023a; Demirer, 2026). Where the current dominant account appeals to non-neutral technical change, we trace the divergence instead to an estimation procedure that leaves the implications of cost minimization unused. The semi-nonparametric translog (SNTL) developed here provides an internally consistent framework for embedding those conditions directly into production function estimation, while maintaining only Hicks-neutral productivity. In doing so, we delineate the distinct roles of cost minimization and non-neutral productivity: the invariance of markups across flexible inputs is delivered by the former, and does not require the latter.

The second strand to which we contribute concerns flexible representations of the production function and the incorporation of first-order conditions into estimation. That the first-order condition for a flexible input links its output elasticity to its cost share in revenue or variable cost has long been recognized (Marschak and Andrews, 1944; Solow, 1957). Under perfect competition and constant returns, this renders elasticities directly observable, as exploited by index number methods (Caves et al., 1982). Subsequent work relaxes these assumptions while imposing the first-order conditions directly in estimation. These conditions are used to recover unobserved productivity or input quantities within parametric specifications (Doraszelski and Jaumandreu, 2013; Grieco et al., 2016; Doraszelski and Jaumandreu, 2018), or to identify the output elasticity of a single flexible input nonparametrically (Gandhi et al., 2020).

Closest to our approach are two recent papers that substitute cost shares for the elasticities of inputs, under the assumption of cost minimization. Jaumandreu (2026) assumes a homogeneous production function of unknown degree, replaces the elasticities of the variable inputs by the product of a common scale parameter and their observed cost shares, and estimates that parameter. Gottardo (2026) develops a method to recover the cost of capital from a noisy proxy, includes it in the cost shares, and fixes the returns to scale at a known

value so that output elasticities are read directly off those shares. In both, a constant scale elasticity—estimated in one case, fixed in the other——governs the mapping from shares to elasticities, which requires the technology to be homogeneous. Our contribution is to represent the production function in terms of its output elasticities without committing to a parametric structure for the latter. Instead, output elasticities are expressed in terms of cost shares and a function representing scale elasticity that varies across observations, so neither homogeneity nor constant returns is imposed, while the cost-minimization conditions hold at every observation when more than one input is flexible. The representation nests commonly used specifications such as the Cobb–Douglas and the translog.

Finally, our framework bridges two traditions within the production approach itself. Hall (1988) and Klette (1999) impose the cost-minimization conditions directly in estimation, but recover only an industry-level markup; De Loecker and Warzynski (2012) recover firm-year-specific markups, but invoke those conditions only after the production function has been estimated. We combine the two, imposing the cost-minimizing conditions through cost shares rather than revenue shares while retaining firm-year-specific markups. The restriction is embedded in the estimating equation itself and is therefore compatible with any of the standard procedures for handling unobserved productivity.

The rest of the paper is organized as follows. Section 2 reviews how cost minimization can be exploited to derive an expression for markups and the standard estimation procedures used in this setting. Section 3 replicates Raval (2023b)'s finding of conflicting markups across flexible inputs using the Colombian Annual Manufacturing Survey, which serves as our running empirical example. Section 4 traces the conflict to its source, showing that the standard approach imposes no requirement that the model-implied cost shares match the observed ones, and recasts the markup expression in terms of the variable-scale elasticity to make its invariance under cost minimization transparent. Section 5 develops our alternative, the semi-nonparametric translog (SNTL), which embeds the cost-minimization conditions directly into estimation and delivers invariant, firm-year-specific markups with Hicks-neutral

productivity alone. Section 6 shows that introducing labor-augmenting productivity also resolves the conflict, but only by exploiting the cost-minimization optimality conditions to identify it. Section 7 places the SNTL alongside the model with labor-augmenting productivity, examines the assumptions each requires, and discusses other sources of cost-share heterogeneity. Section 8 corroborates these findings with manufacturing data from Chile and the United States. Section 9 applies the transparency diagnostic proposed by Fernald et al. (2026), measuring how much of the variation in input cost shares in revenue is explained by the estimated output elasticities, with the remainder attributed to markups. Section 10 concludes.

2 Recovering Markups Using the Production Approach

This section reviews the analytical framework that underlies the production approach to markup estimation and discusses the estimation procedures commonly employed in the literature. In line with De Loecker and Warzynski (2012), we consider a firm i using the following production technology at time t:

$$Q_{it} = Q_{it}(\mathbf{X}_{it}, A_{it}; \boldsymbol{\beta}), \tag{1}$$

where Q_{it} is the planned output, $\mathbf{X}_{it} = (X_{it}^1, ..., X_{it}^n)$ is the vector of all n inputs (indexed by $j \in J \equiv J^V \cup J^K = \{1, ..., n\}$), where J^V denotes the set of variable inputs and J^K the set of predetermined inputs (e.g., capital), $\boldsymbol{\beta}$ is the vector of production function parameters, and A_{it} denotes productivity.[3] Producers are assumed to be cost minimizers. The Lagrangian

[3] A variable (flexible) input is chosen at t after the productivity shock is realized; a predetermined (dynamic) input is chosen at $t-1$ for use at t. In empirical applications, materials are typically treated as flexible and capital as predetermined; the classification of labor depends on the assumed adjustment frictions.

associated with the static cost minimization problem is as follows:

$$\mathcal{L}(\mathbf{X}_{it}, \lambda_{it}) = \sum_{j \in J^V} W_{it}^j X_{it}^j + \lambda_{it}[Q_{it} - Q_{it}(\cdot)], \tag{2}$$

where W_{it}^j denotes the price of variable input j and λ_{it} is the Lagrange multiplier. Assuming there is no input market power nor other adjustment frictions, the optimality conditions require marginal products of variable inputs to be proportional to their factor prices. Specifically, multiplying both sides of the cost-minimizing first-order condition (FOC) for any variable input j by $\frac{X_{it}^j}{Q_{it}}$ and rearranging yields:

$$\frac{\partial Q_{it}(\cdot)}{\partial X_{it}^j} \frac{X_{it}^j}{Q_{it}} = \frac{1}{\lambda_{it}} \frac{W_{it}^j X_{it}^j}{Q_{it}}, \quad \forall j \in J^V. \tag{3}$$

The Lagrange multiplier, λ_{it}, is the marginal cost of production at a given level of output. Define markups to be the ratio of output price to marginal cost, $\mu_{it} \equiv \frac{P_{it}}{\lambda_{it}}$, where P_{it} is the output price. Equation (3) may then be rewritten as:

$$\mu_{it} = \frac{\theta_{it}^j}{R_{it}^j}, \tag{4}$$

where $\theta_{it}^j \equiv \frac{\partial Q_{it}(\cdot)}{\partial X_{it}^j} \frac{X_{it}^j}{Q_{it}}$ is the output elasticity of variable input j and $R_{it}^j \equiv \frac{W_{it}^j X_{it}^j}{P_{it} Q_{it}}$ is its share of cost in planned revenue—hereafter, "revenue share." Equation (4) characterizes how production functions and output market power are connected through factor demand, and is the key relationship underlying the production approach to markup estimation. Were the output elasticity of a variable input to exceed its revenue share, the inference would be that the firm is restricting input use to reduce output and thereby charge a price above marginal cost. Notably, the only behavioral condition invoked to derive equation (4) is cost minimization—no assumptions on the output market structure or the demand curves faced by firms are necessary.

Estimating output elasticities, therefore, is the main empirical challenge for recovering

firm-level markups. Researchers typically estimate the production function and output elasticities using the control function approach (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015) or the dynamic panel method (Arellano and Bond, 1991; Blundell and Bond, 1998, 2000), and then recover markups—essentially as a residual (Fernald et al., 2026)—via equation (4).

Estimation Procedures. Production function estimation in this context typically relies on the following assumptions. First, observed output differs from firms' planned output by an unobservable term (whereas this is not the case for inputs). Second, productivity is Hicks-neutral; the productivity term is known to firms when they make their production decision but it is unobserved by the econometrician. The production function is thus represented as: $Y_{it} = F(\mathbf{X}_{it}; \boldsymbol{\beta}) \exp(A_{it}^H) \exp(\epsilon_{it}^Y)$, where Y_{it} is the observed output, A_{it}^H denotes Hicks-neutral productivity such that planned output is $Q_{it} \equiv F(\cdot) \exp(A_{it}^H)$, and ϵ_{it}^Y collects measurement error and shocks to output that are unanticipated by the firm at time t. Log-transforming yields:

$$y_{it} = f(\mathbf{x}_{it}; \boldsymbol{\beta}) + A_{it}^H + \epsilon_{it}^Y, \tag{5}$$

where lower-case letters represent log-transformed values. In the standard scenario, inputs include capital (K_{it}), labor (L_{it}), and materials (M_{it}), with capital treated as a predetermined input.

The control function approach and the dynamic panel method address the primary challenge of estimating production functions—that the unobserved productivity is correlated with input choices—under different assumptions. The control function approach proxies for productivity by inverting an observed factor demand, whereas the dynamic panel method differences out its persistent component under a linear Markov process. In the control function approach, observed output is first purged of ϵ_{it}^Y to recover planned output as $\hat{Q}_{it} = Y_{it} \exp(-\hat{\epsilon}_{it}^Y)$, which is required to construct revenue shares; the production function

parameters $\boldsymbol{\beta}$ are then estimated.[4] Given $\hat{\boldsymbol{\beta}}$, markups are recovered as:

$$\hat{\mu}_{it} = \frac{\hat{\theta}^j_{it}(\mathbf{x}_{it}; \hat{\boldsymbol{\beta}})}{W^j_{it} X^j_{it} \,/\, [P_{it} Y_{it} \exp(-\hat{\epsilon}^Y_{it})]}, \tag{6}$$

where the denominator is the estimated revenue share of input j evaluated at planned output. The functional form of $\hat{\theta}^j_{it}(\mathbf{x}_{it}; \hat{\boldsymbol{\beta}})$ depends on the structure of the production function—for instance, $\hat{\theta}^j_{it} = \hat{\beta}_j$ when the production function is Cobb-Douglas. We describe both estimation procedures in detail in Appendix A.

3 Conflicting Markup Estimates

Under the production approach, the output elasticity of *any* flexible input is sufficient to recover markups; therefore, markups are "over-identified" whenever there are multiple flexible inputs. Exploiting this feature of the underlying framework, Raval (2023b) compares markups estimated using different flexible inputs—such as labor and materials, or different subsets of materials—across seven different data sets and strongly rejects the hypothesis that different flexible inputs yield the same markup in each of the seven data sets.

To illustrate the problem clearly, here we replicate the main results of Raval (2023b) using the Colombian Annual Manufacturing Survey data for the time period 1978-1991.[5] Specifically, the following translog production function in capital, labor, and materials with

[4] For the dynamic panel method, planned output is constructed analogously; see Appendix C.

[5] We describe the data in detail in Appendix B.

Hicks-neutral productivity is considered:[6]

$$y_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + \tfrac{1}{2}\beta_{KK} k_{it}^2 + \tfrac{1}{2}\beta_{LL} l_{it}^2 + \tfrac{1}{2}\beta_{MM} m_{it}^2 + \beta_{KL} k_{it} l_{it} + \beta_{KM} k_{it} m_{it} + \beta_{LM} l_{it} m_{it} + A_{it}^H + \epsilon_{it}^Y, \tag{7}$$

where $\boldsymbol{\beta} = \{\beta_K, \beta_L, \beta_M, \beta_{KK}, \beta_{LL}, \beta_{MM}, \beta_{KL}, \beta_{KM}, \beta_{LM}\}$ is the set of production function parameters, A_{it}^H is the Hicks-neutral productivity, and ϵ_{it}^Y represents unanticipated shocks to production and measurement error in output. Raval (2023b) implements the two-stage control function procedure (De Loecker and Warzynski, 2012; Ackerberg et al., 2015), treating capital as a predetermined input and labor and materials as flexible inputs.[7] After netting out the measurement error in output (ϵ_{it}^Y) in stage 1, the production function parameters are estimated in stage 2 using moments with the following instruments:

$$Z_{it} = \big(\, k_{it},\ l_{it-1}, m_{it-1},\ k_{it}^2, l_{it-1}^2, m_{it-1}^2, k_{it} l_{it-1}, k_{it} m_{it-1}, l_{it-1} m_{it-1} \big)'. \tag{8}$$

After obtaining parameter estimates, output elasticities are recovered as:

$$\hat{\theta}_{it}^j = \hat{\beta}_j + \sum_{\ell} \hat{\beta}_{j\ell}\, x_{it}^{\ell}, \quad j \in \{L, M\};\ \ell \in \{K, L, M\}. \tag{9}$$

Further, the first-stage residual is used to estimate planned output from observed sales: $\hat{Q}_{it} = Y_{it} \exp(-\hat{\epsilon}_{it}^Y)$. The markups are then estimated using output elasticities and revenue shares for flexible inputs, labor, and materials, according to equation (6). The results, reported in Table 1, show that markup estimates based on labor and materials differ significantly in levels. Furthermore, to examine the time trends in estimated markups, Raval (2023b)

[6] Raval (2023b) considers two functional forms (Cobb-Douglas and translog), different sets of inputs, and various estimation procedures. We choose the most flexible functional form and the most prevalent estimation strategy—the control function approach—to illustrate what drives this empirical finding.

[7] The production function is estimated at the 3-digit industry level. Estimation details are provided in Appendix A.

estimates the following specification:

$$\log(\hat{\mu}_{it}^{j}) = \alpha + \tau_t + \delta_h + \epsilon_{it}^{\mu^j}, \tag{10}$$

where τ_t and δ_h are year and industry fixed effects. The estimated year effects are plotted in Figure 1.

Figure 1, which closely replicates one of the results reported by Raval (2023b), indicates that markups based on labor and materials not only differ significantly in levels, but they also imply divergent trends. This pattern persists across all the data sets Raval (2023b) considers—plant-level data for Chile, India, and Colombia, firm-level data for Indonesia, Compustat data for public firms in the US, a firm panel for Southern Europe, and unique store-level data from a US retailer—and across different specifications and estimation procedures. As noted earlier, Doraszelski and Jaumandreu (2019), using data from Spain, also documents significant discrepancies between markups obtained from labor versus material.

4 Invariance of Markup Estimates and Cost Minimization

What leads to the conflicting markup estimates? We maintain that the key reason is as follows: whereas the underlying economic framework appeals to cost minimization, the econometric model of the production function approach contains nothing that requires the data and the parametrized production function to be jointly consistent with that hypothesis. To make our argument concrete, recall the cost-minimizing first-order condition from the firm's problem can be rewritten as:

$$W_{it}^{j} X_{it}^{j} = \lambda_{it} \frac{\partial Q_{it}(\cdot)}{\partial X_{it}^{j}} X_{it}^{j}, \quad \forall j \in J^{V}. \tag{11}$$

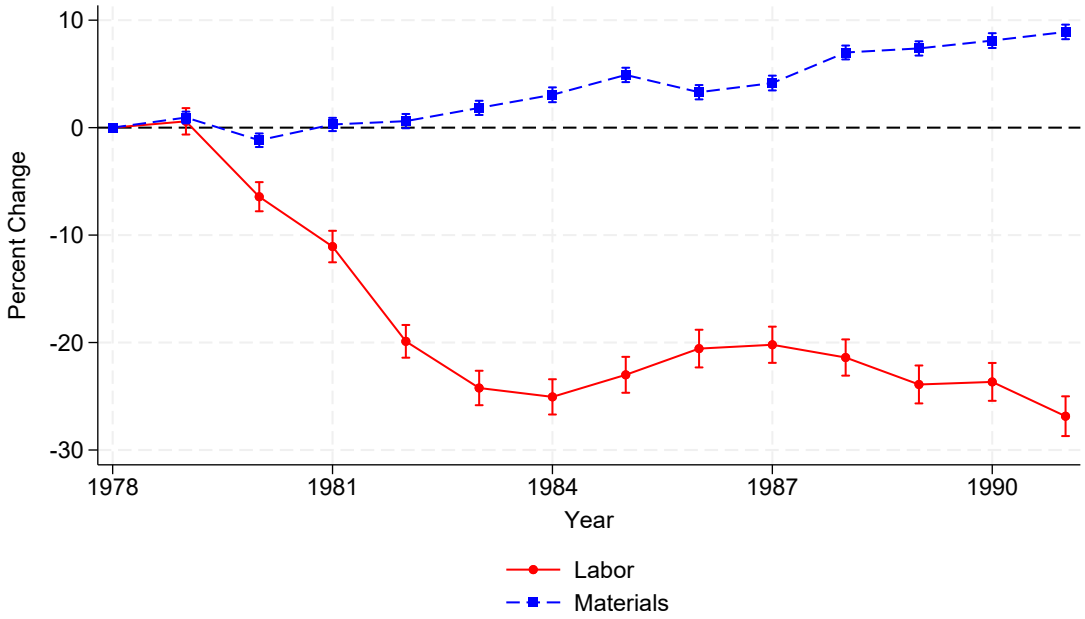
Summing over the variable inputs yields the minimum variable cost as $\sum_{j \in J^V} W_{it}^{j} X_{it}^{j} = \lambda_{it} \sum_{j \in J^V} \frac{\partial Q_{it}(\cdot)}{\partial X_{it}^{j}} X_{it}^{j}$. Cost-minimizing input cost shares, defined as $S_{it}^{j} \equiv W_{it}^{j} X_{it}^{j} / \sum_{\ell \in J^V} W_{it}^{\ell} X_{it}^{\ell}$,

Table 1: Markup Estimates Based on Raval (2023b)

Input	Mean	Median	SD	P10	P90
Labor	1.7190	1.5832	1.5045	0.7555	3.0395
Materials	1.2475	1.2099	0.8232	0.9756	1.6618

Notes: Markups estimated using the control function procedure following Raval (2023b). Markup is computed as the output elasticity divided by the revenue share for each input.

Figure 1: Markup Trends (Raval, 2023b)



Notes: Figure 1 is similar to figure 2 (b) of Raval (2023b). Estimates are based on equation (10). Vertical bars represent 95% confidence intervals based on clustering at the plant level. All estimates are relative to the first year, which is set to zero.

therefore must satisfy:

$$S_{it}^{j} = \frac{\theta_{it}^{j}(\mathbf{X}_{it};\boldsymbol{\beta})}{\gamma_{it}^{V}(\mathbf{X}_{it};\boldsymbol{\beta})}, \tag{12}$$

where $\gamma_{it}^{V} \equiv \sum_{j} \theta_{it}^{j}$ denotes the short-run scale elasticity—the sum of output elasticities of all variable inputs. The structure of the production function, which determines θ_{it}^{j} and γ_{it}^{V}, therefore constrains the cost shares of the variable inputs. The same logic, of course, constrains the revenue shares used in the standard markup expression of equation (4):

$$R_{it}^{j} = \frac{W_{it}^{j} X_{it}^{j}}{P_{it} Q_{it}} = S_{it}^{j} \times \frac{\text{Variable Cost}_{it}}{\text{Planned Revenue}_{it}}. \tag{13}$$

Combining (13) and (12), the markup formula (4) can be rewritten in terms of the variable-scale elasticity:

$$\mu_{it} = \gamma_{it}^{V} \frac{P_{it} Q_{it}}{\sum_{j \in J^{V}} W_{it}^{j} X_{it}^{j}}. \tag{14}$$

Equation (14) makes the source of markup invariance under cost minimization transparent: markups recovered from γ_{it}^{V} and observables are invariant across variable inputs by construction. The standard econometric approach, however, imposes no explicit connection between the *model-implied* cost shares—the right-hand-side of equation (12) —and the *observed* cost shares of variable inputs. The conflicting markup estimates documented in Section 3 primarily reflect this mismatch.

Inflexibility of Parametric Production Functions. Equation (12) is a restriction that must hold at each observation. With J^{V} variable inputs, it imposes $J^{V} - 1$ independent restrictions per firm-year as cost shares sum to one, or $N(J^{V} - 1)$ restrictions in a sample of size N. To see this, take the ratio of (12) for two variable inputs j and ℓ:

$$\frac{\theta_{it}^{j}}{\theta_{it}^{\ell}} = \frac{S_{it}^{j}}{S_{it}^{\ell}}, \qquad j, \ell \in J^{V}. \tag{15}$$

i.e., cost minimization implies that the ratio of output elasticities of any two variable inputs must be equal to the ratio of their cost shares; the restrictions in (12) are precisely the $J^V - 1$ independent ratios of this form. When there is a single variable input, then $\gamma_{it}^V = \theta_{it}^j$ by definition and (12) reduces to $S_{it}^j = 1$, an accounting identity, satisfied by any parameter vector. When $J^V \geq 2$, however, the restrictions bear on a parameter vector $\boldsymbol{\beta}$ of fixed, finite dimension, and generically no such vector satisfies them all. The difficulty is transparent under the Cobb–Douglas specification in three inputs (k_{it}, l_{it}, m_{it}) with capital treated as predetermined, where $\theta_{it}^j = \beta_j$ and (12) requires:

$$S_{it}^L = \frac{\beta_L}{\beta_L + \beta_M}, \tag{16}$$

for all i and t, so that every firm must exhibit the same variable cost share in every period. The translog in (7) relaxes this structure—the model-implied share becomes a function of (k_{it}, l_{it}, m_{it}) and parameters (9)—but does not solve the problem: a fixed parameter count is asked to reproduce a number of observed cost shares that grows with the sample. The obstacle is finite parameterization itself rather than any particular functional form; employing a flexible parametric functional form narrows the gap without closing it.[8]

The current dominant approach to address the foregoing problem is to allow for non-neutral technical change, such as labor-augmenting productivity, so that the elasticity ratio is no longer a deterministic function of inputs and parameters (Doraszelski and Jaumandreu, 2019; Raval, 2023a; Demirer, 2026). As we discuss in detail in Section 6, this resolution works because cost-minimizing conditions—missing from the standard estimation approach—are explicitly invoked to identify non-neutral technical change.

We pursue an alternative and novel route. Observe that (12) restricts only how the scale elasticity is allocated across variable inputs, placing no restriction on the scale elasticity

[8] Raval (2023b) makes the same observation, noting that under Hicks-neutral productivity the ratio of output elasticities is a deterministic function of parameters and inputs, carrying no unobservable term, and therefore cannot reproduce the observed dispersion in input cost ratios across plants. We return to this point in Section 6.

itself—and by (14), the scale elasticity is the only object markups require. Accordingly, the semi-nonparametric methodology developed below starts by representing the production function in terms of its (unrestricted) output elasticities, and makes the approach operational by parameterizing the scale elasticity γ_{it}^{V} directly.

5 Embedding Cost Minimization in Production Function Estimation

Two observations emerge from the preceding discussion. First, with multiple flexible inputs, the structure of the production function, together with the assumption of cost minimization, constrains input cost shares. The standard econometric model of the production approach, however, does not restrict the *model-implied* cost shares to equal the *observed* cost shares, leading to conflicting markup estimates. Second, labor-augmenting productivity resolves this conflict. As noted in the foregoing, it does so because the underlying framework uses cost minimization to identify the additional unobserved productivity term; Section 6 shows this in detail.

In this section, we develop an alternative approach—the semi-nonparametric translog (SNTL) production function—where we do not invoke factor-biased productivity, but we do embed the maintained cost minimization conditions in the production function estimand. We show that this approach delivers identical markups across all flexible inputs, notwithstanding its reliance on a Hicks-neutral productivity specification. We develop our approach in two steps. First, we express the production function in terms of output elasticities, and second, we embed cost minimization conditions in the production function, exploiting the relationship between output elasticities and cost shares, as stated in (12).

5.1 Production Function and Output Elasticities

We seek a representation of the production function in terms of its output elasticities, which provides the ideal bridge to connect the production function with the first-order conditions of cost minimization. For notational clarity, in this section we drop the subscripts it. The primitive function is $F(\mathbf{X}) : \mathbb{R}^n_{++} \to \mathbb{R}_{++}$, defined in terms of a vector of inputs $\mathbf{X} = (X^1, ..., X^n) \in \mathbb{R}^n_{++}$. Representing this production function in terms of log-transformed variables $x^j \equiv \ln X^j$ yields the mapping $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$. Assume that $f(\mathbf{x})$ is continuously differentiable, and define its output elasticities as:

$$\theta^j(\mathbf{x}) \equiv \frac{\partial f(\mathbf{x})}{\partial x^j}. \tag{17}$$

Then, from the fundamental theorem of calculus, the function $f(\mathbf{x})$ can be expressed as:

$$f(\mathbf{x}) = f(\mathbf{0}) + \int_0^1 \sum_{j \in J} \theta^j(r\mathbf{x})\, x^j \, dr. \tag{18}$$

For any $\mathbf{x} \in \mathbb{R}^n$, define the scalar function:

$$G(r) \equiv \sum_{j \in J} x^j \, \theta^j(r\mathbf{x}), \qquad r \in [0, 1]. \tag{19}$$

Then, from (18), the production function $f(\mathbf{x})$ admits the exact line-integral representation:

$$f(\mathbf{x}) = f(\mathbf{0}) + \int_0^1 G(r)\, dr. \tag{20}$$

The trapezoidal approximation of a line integral then implies:

$$f(\mathbf{x}) \cong f(\mathbf{0}) + \tfrac{1}{2}\, G(0) + \tfrac{1}{2}\, G(1).$$

Recalling the definition of the scalar function $G(r)$ in (19), the function $f(\mathbf{x})$ can be expressed as:

$$f(\mathbf{x}) \cong f(\mathbf{0}) + \frac{1}{2}\sum_j \theta^j(\mathbf{0})\, x^j + \frac{1}{2}\sum_j \theta^j(\mathbf{x})\, x^j. \tag{21}$$

By defining $\beta_0 \equiv f(\mathbf{0})$ and $\beta_j \equiv \theta^j(\mathbf{0})$, we obtain:

$$f(\mathbf{x}) \cong \beta_0 + \frac{1}{2}\sum_j \beta_j\, x^j + \frac{1}{2}\sum_j \theta^j(\mathbf{x})\, x^j. \tag{22}$$

Equation (22) differs from the common Taylor expansion in that it averages marginal effects evaluated both at the base point and at the evaluation point, as is the nature of a trapezoidal approximation to the exact line integral. In fact, it can be equivalently stated as:

$$f(\mathbf{x}) \cong \beta_0 + \frac{1}{2}\sum_j \big[\beta_j + \theta^j(\mathbf{x})\big] x^j, \tag{23}$$

which makes it explicit that the approximation weights each input by the average of the output elasticity at the base point and at the current point. The only maintained assumption is that $f(\mathbf{x})$ is continuously differentiable, and thus also continuous, everywhere.

Equation (22) is the starting point for the semi-nonparametric translog (SNTL) that we develop in the next section. The base point $x^j = 0$ corresponds to input quantities $X^j = 1$ for all inputs. To make this approach operational, it helps to think of the point $X^j = 1$, $\forall j$, as particularly relevant for the purpose of the desired approximation, e.g., the mean point of the data (a normalization that we will enforce in the empirical applications that follow).

5.2 The Semi-Nonparametric Translog (SNTL) Production Function

In line with the production approach to markup estimation, we consider a cost-minimizing firm with a stochastic, Hicks-neutral production technology, as in equation (5). Applying

the representation (22) to $f(\mathbf{x})$, the production function can be written as:

$$y_{it} = \beta_0 + \frac{1}{2}\sum_{j\in J} \beta_j \, x_{it}^j + \frac{1}{2}\sum_{j\in J} \theta_{it}^j \, x_{it}^j + A_{it}^H + \epsilon_{it}^Y, \tag{24}$$

where $\theta_{it}^j \equiv \theta^j(\mathbf{x}_{it})$ denote the output elasticities. In this representation, the functional forms of the elasticities $\theta^j(\mathbf{x})$ need not be specified. The definition $\partial f(\mathbf{x})/\partial x^j \equiv \theta^j(\mathbf{x})$ nonetheless continues to hold, so that, from (24):

$$\theta^j(\mathbf{x}) = \beta_j + \sum_{\ell} \frac{\partial \theta^\ell(\mathbf{x})}{\partial x^j} \, x^\ell. \tag{25}$$

Equation (25) comprises a system of partial differential equations in the n unknown functions $\theta^j(\mathbf{x})$ that implicitly defines the structure of the output elasticities, provided the standard integrability (symmetry) condition $\frac{\partial \theta^j(\mathbf{x})}{\partial x^\ell} = \frac{\partial \theta^\ell(\mathbf{x})}{\partial x^j}$, $\forall j, \ell \in J$, holds.

We note at this juncture that the foregoing approach nests commonly-used parameterizations as special cases. Specifically, when $\partial \theta^\ell(\mathbf{x})/\partial x^j = 0$, then $\theta^j(\mathbf{x}) = \beta_j$, and the production function in (24) takes the Cobb–Douglas form:

$$y_{it} = \beta_0 + \sum_{j} \beta_j \, x_{it}^j + A_{it}^H + \epsilon_{it}^Y.$$

Alternatively, when $\partial \theta^\ell(\mathbf{x})/\partial x^j = \beta_{j\ell}$, then $\theta^j(\mathbf{x}) = \beta_j + \sum_\ell \beta_{j\ell} \, x^\ell$, and the production function in (24) takes the standard translog form:

$$y_{it} = \beta_0 + \sum_{j} \beta_j \, x_{it}^j + \frac{1}{2}\sum_{j}\sum_{\ell} \beta_{j\ell} \, x_{it}^j \, x_{it}^\ell + A_{it}^H + \epsilon_{it}^Y.$$

In what follows, however, we eschew both of these avenues in favor of a parameterization that permits the cost minimization conditions to be explicitly maintained in the specification of the production function.

Embedding Cost Minimization When All Inputs Are Flexible. We return to the production function in (24). To fix ideas, let's assume for now that all inputs are variable (we revisit the distinction between variable and predetermined inputs in the next section). The first-order conditions of cost minimization imply that:

$$\theta_{it}^{j} = S_{it}^{j}\,\gamma(\mathbf{x}_{it}), \qquad j \in J. \tag{26}$$

where $S_{it}^{j} \equiv W_{it}^{j}X_{it}^{j}/\sum_{\ell} W_{it}^{\ell}X_{it}^{\ell}$ denote the cost-minimizing input cost shares and $\gamma(\mathbf{x}_{it}) \equiv \sum_{j}\theta_{it}^{j}$ is the scale elasticity. Using (26) in (24), the production function can be re-stated as:

$$y_{it} = \beta_0 + \frac{1}{2}\sum_{j}\beta_j\,x_{it}^{j} + \frac{1}{2}\,\gamma(\mathbf{x}_{it})\sum_{j} S_{it}^{j}\,x_{it}^{j} + A_{it}^{H} + \epsilon_{it}^{Y}. \tag{27}$$

Next, approximate the scale elasticity by a linear-in-logs function, $\gamma(\mathbf{x}) = a_0 + \sum_j a_j\,x^j$. To ensure consistency with the output elasticities in (25), we set $a_0 \equiv \sum_j \beta_j$.[9] The proposed parameterization of the scale elasticity is therefore:

$$\gamma(\mathbf{x}) = \sum_{j}\beta_j + \sum_{j} a_j\,x^j. \tag{28}$$

Combining (28) with (27) yields the final form of the SNTL model. Let $z_{it} \equiv \sum_j S_{it}^{j}\,x_{it}^{j}$ denote the cost-share-weighted log index of inputs (essentially, a level version of a Divisia input index). Then the production function can be written as a restricted "translog" defined on the individual inputs and the Divisia input index z_{it}. Specifically:

$$y_{it} = \beta_0 + \frac{1}{2}\sum_{j}\beta_j\big(x_{it}^{j} + z_{it}\big) + \frac{1}{2}\sum_{j} a_j\,x_{it}^{j}\,z_{it} + A_{it}^{H} + \epsilon_{it}^{Y}. \tag{29}$$

[9] To see the logic of this normalization, evaluate (25) at the base point $\mathbf{x} = \mathbf{0}$, which gives $\theta^j(\mathbf{0}) = \beta_j$ and hence $\gamma(\mathbf{0}) = \sum_j \theta^j(\mathbf{0}) = \sum_j \beta_j$. Recall the approximation strategy of Section 5.1 weights each input by the average of its output elasticity at the base point and at the current point; the restriction therefore ties the intercept of $\gamma(\mathbf{x})$ to the same base point elasticities.

Note that the production function in (29) is considerably more parsimonious than standard flexible functional forms. Abstracting from the Hicks-neutral process in A_{it}^H, for the case of J inputs, the SNTL in (29) has $(2J+1)$ parameters; by contrast, the standard (symmetric) translog parameterization would require $\frac{1}{2}(J+1)(J+2)$ parameters. What equation (29) offers is the minimal structure needed to identify a flexible parametrization of the scale elasticity $\gamma(\mathbf{x})$. Given cost minimization, the scale elasticity is all that is required to recover output elasticities, as $\theta_{it}^j = S_{it}^j \gamma(\mathbf{x}_{it})$, and, consequently, markups. Needless to say, other properties of the production structure (such as elasticities of substitution) cannot be uncovered by the SNTL in (29).

The SNTL When There Are Predetermined Inputs. It is straightforward to adapt the foregoing semi-nonparametric setup to the empirically-relevant case where the distinction between flexible and pre-determined inputs matters. For concreteness, here we do so for the setting that is common in the literature and that pertains to all applications presented in this paper—three inputs, where capital is predetermined and labor and materials as flexible. The SNTL derived above, for this standard three-input case, starts from:

$$y_{it} = \frac{1}{2}\big(\beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it}\big) + \frac{1}{2}\big(\theta_{it}^K k_{it} + \theta_{it}^L l_{it} + \theta_{it}^M m_{it}\big) + A_{it}^H + \epsilon_{it}^Y, \tag{30}$$

where input K (capital) is predetermined and inputs L (labor) and M (materials) are flexible. The first-order conditions of cost minimization in (26) therefore apply only to inputs L and M. Hicks-neutral productivity, A_{it}^H, is assumed to follow a first-order linear Markov process.[10]

We use the standard translog parameterization for θ^K.[11] Let S^L and S^M be the cost shares relative to the variable-input cost only. Further, define $\gamma^V(\mathbf{x})$ as the short-run scale

[10] The constant β_0 is not separately identified from the constant in the productivity process. See Appendix A for details.

[11] One could, of course, envision additional parametric restrictions on the specification of θ^K. For example, one could impose $\beta_{KL} = \beta_{KM} = 0$, such that $\theta_{it}^K = \beta_K + \beta_{KK} k_{it}$. The argument would be that, as noted earlier, the SNTL is not meant to capture substitution possibilities, but only to estimate the scale elasticity.

elasticity (the scale elasticity of the variable inputs only). Then:

$$\begin{aligned}\theta_{it}^{K} &= \beta_K + \beta_{KK}k_{it} + \beta_{KL}l_{it} + \beta_{KM}m_{it},\\ \theta_{it}^{L} &= S_{it}^{L}\,\gamma_{it}^{V},\\ \theta_{it}^{M} &= S_{it}^{M}\,\gamma_{it}^{V},\end{aligned} \tag{31}$$

where the last two equations are the short-run cost-minimization conditions, with variable-cost shares satisfying $S_{it}^{L} + S_{it}^{M} = 1$ (and thus, as desired, $\gamma_{it}^{V} = \theta_{it}^{L} + \theta_{it}^{M}$). The input index here is defined over the variable inputs: $z_{it} \equiv S_{it}^{L}l_{it} + S_{it}^{M}m_{it}$. Substituting into (30) yields:

$$y_{it} = \frac{1}{2}\big(\beta_K k_{it}+\beta_L l_{it}+\beta_M m_{it}\big)+\frac{1}{2}\big(\beta_K k_{it}+\beta_{KK}k_{it}^2+\beta_{KL}k_{it}l_{it}+\beta_{KM}k_{it}m_{it}\big)+\frac{1}{2}\,\gamma_{it}^{V}\,z_{it}+A_{it}^{H}+\epsilon_{it}^{Y}.$$

Next, parameterize the short-run scale elasticity as:

$$\gamma_{it}^{V} = (\beta_L + \beta_M) + a_K k_{it} + a_L l_{it} + a_M m_{it}. \tag{32}$$

Given (31) and (32), the integrability (symmetry) condition implies that $a_K = \beta_{KL} + \beta_{KM}$. With this parameterization, we obtain:

$$\begin{aligned} y_{it} = \beta_K k_{it} &+ \frac{1}{2}\beta_{KK}k_{it}^2\\ &+ \frac{1}{2}\big[\beta_L(l_{it}+z_{it}) + \beta_M(m_{it}+z_{it})\big]\\ &+ \frac{1}{2}\big[\beta_{KL}k_{it}(l_{it}+z_{it}) + \beta_{KM}k_{it}(m_{it}+z_{it})\big]\\ &+ \frac{1}{2}\big(a_L l_{it} + a_M m_{it}\big)z_{it} + A_{it}^{H} + \epsilon_{it}^{Y}. \end{aligned} \tag{33}$$

We estimate equation (33) using the Colombian Annual Manufacturing Survey data and the estimation strategy of Section 3, so that the results are directly comparable to those of Raval

(2023b).[12]

5.3 Estimation Procedure and Results

We implement the two-stage control function approach to estimate the set of parameters $\boldsymbol{\beta} = (\beta_K, \beta_L, \beta_M, \beta_{KK}, \beta_{KL}, \beta_{KM}, a_L, a_M)$ using the moment conditions with the following instruments:

$$\begin{aligned} Z_{it} = \big(&k_{it},\ l_{it-1} + z_{it-1},\ m_{it-1} + z_{it-1},\ k_{it}^2, \\ &k_{it}\left(l_{it-1} + z_{it-1}\right),\ k_{it}\left(m_{it-1} + z_{it-1}\right),\ l_{it-1}z_{it-1},\ m_{it-1}z_{it-1}\big)'. \end{aligned} \tag{34}$$

After estimating $\boldsymbol{\beta}$, we recover the variable-scale elasticity as $\hat{\gamma}_{it}^V = \hat{\beta}_L + \hat{\beta}_M + (\hat{\beta}_{KL} + \hat{\beta}_{KM})k_{it} + \hat{a}_L l_{it} + \hat{a}_M m_{it}$. Output elasticities for the flexible inputs are obtained from equation (31) as $\hat{\theta}_{it}^j = \hat{\gamma}_{it}^V S_{it}^j$, and markups are computed from equation (6). Table 2 provides the summary statistics of estimated markups using our alternative approach, and Figure 2 plots the time trends in estimated markups. As expected, the markup estimates are identical across both labor and materials.

The markup estimates from the semi-nonparametric translog in equation (33) indicate that embedding cost-minimization conditions into the production function estimation *alone* is sufficient to obtain estimates that are invariant across multiple flexible inputs. Our framework retains the Hicks-neutral specification—making no use of factor-augmenting productivity—and delivers invariant markups by imposing cost-minimization conditions. Therefore, insofar as invariance of markup estimates is concerned, cost minimization is sufficient on its own; factor-augmenting productivity is not required. Section 6 presents a model

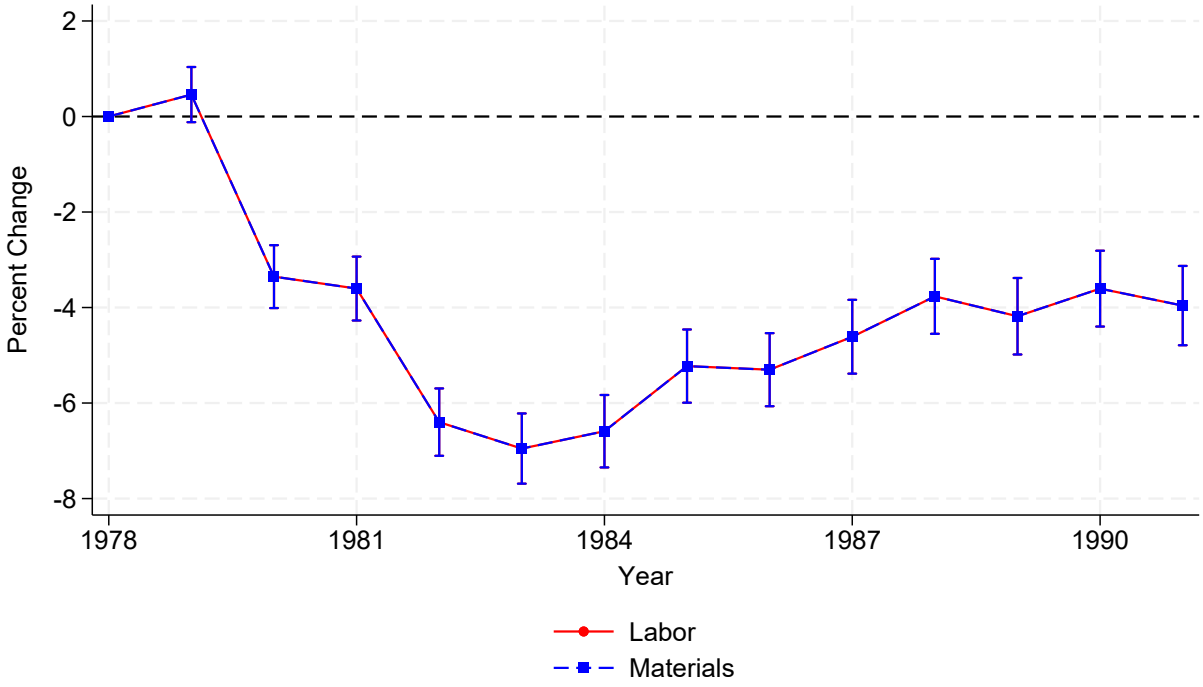
[12]Estimation details are in Appendix A. For estimation in Sections 5 and 6, we normalize the data in the following manner. For each industry h, define $\bar{x}_h = N_h^{-1} \sum_{it \in h} x_{it}$ for $x \in \{k, l, m, y\}$ and replace each log variable in the estimating equation with its within-industry demeaned value, $\tilde{x}_{it} = x_{it} - \bar{x}_h$. Cost shares and revenue shares are functions of observed expenditures and are invariant to this rescaling. Since the estimating equation in Section 3 [equation (7)] does not have cost shares, those estimation results will be invariant to using normalized data.

Table 2: Markup Estimates—Semi-Nonparametric Translog

Input	Mean	Median	SD	P10	P90
Labor	1.3638	1.3225	0.3474	1.0142	1.7496
Materials	1.3638	1.3225	0.3474	1.0142	1.7496

Notes: Markups estimated using the control function procedure. Markup is computed as the output elasticity divided by the revenue share for each input.

Figure 2: Markup Trends—Semi-Nonparametric Translog



Notes: This figure plots the estimated year effects based on equation (10) and is comparable to figure 1. Vertical bars represent 95% confidence intervals based on clustering at the plant level. All estimates are relative to the first year, which is set to zero.

with labor-augmenting productivity and shows that it achieves invariance through the same channel.

5.4 Two Traditions in the Production Approach

Incorporating cost-minimizing first-order conditions into production function estimation aligns the SNTL with the tradition of Hall (1988). He considers a production function $Q_{it} = F(K_{it}, L_{it}, M_{it}) \exp(A^H_{it})$, where capital is predetermined, and labor and materials are flexible. Taking logs and total differentials yields the familiar growth identity in terms of output elasticities, $\Delta q_{it} = \theta^K_{it} \Delta k_{it} + \theta^L_{it} \Delta l_{it} + \theta^M_{it} \Delta m_{it} + \Delta A^H_{it}$. Imposing constant returns to scale and substituting the cost-minimization conditions, $\theta^L_{it} = \mu_{it} R^L_{it}$ and $\theta^M_{it} = \mu_{it} R^M_{it}$, Hall (1988) obtains:[13]

$$\Delta q_{it} - \Delta k_{it} = \mu_{it} \big[R^L_{it} (\Delta l_{it} - \Delta k_{it}) + R^M_{it} (\Delta m_{it} - \Delta k_{it}) \big] + \Delta A^H_{it}, \tag{35}$$

where the right-hand side is a revenue-share-weighted aggregate of flexible inputs relative to capital, and ΔA^H_{it} is the rate of Hicks-neutral technical progress. Essentially, if one were to ignore the endogeneity issues caused by the unobserved productivity, a least-squares regression of $\Delta q_{it} - \Delta k_{it}$ on the revenue-share-weighted aggregate would yield an estimate of the industry-level average markup as the slope coefficient. While Hall (1988) does not estimate equation (35), Klette (1999) extends this approach to firm-level data, relaxes constant returns to scale, and employs a specification similar to equation (35) to estimate the markup directly as an industry-specific, time-invariant parameter.

Estimating equations such as (35) have two limitations, beyond those encountered in standard production function estimation. First, the revenue shares R^L_{it} and R^M_{it} are not strictly observed: the maintained assumption is that planned output Q_{it} differs from observed output Y_{it} by measurement error and/or unanticipated shocks to production. The empirically

[13] Equation (35) is a restated version of equation (19) in Section III (page 931) of Hall (1988).

constructed shares are thus noisy proxies for their true counterparts and introduce endogeneity into the constructed regressor on the right-hand side. Second, the specification recovers only an industry-level average markup, not firm-year-specific estimates.

De Loecker and Warzynski (2012) introduced what is now termed the *new* production approach, which addresses several limitations of the Hall/Klette framework. They estimate the production function using control function methods that handle unobserved productivity flexibly and recover firm-year-specific markups indirectly via $\hat{\mu}_{it}^{j} = \hat{\theta}_{it}^{j}/\hat{R}_{it}^{j}$. Their framework—now the standard model of the production approach—delivers heterogeneous markups, accommodates flexible production technologies, and dispenses with the constant-returns-to-scale assumption required by Hall (1988). The cost-minimization conditions, however, enter only through the markup expression; they are not employed during the estimation of the production function itself. As shown in Section 4, this is the source of the conflicting markup estimates across flexible inputs.

Our approach bridges the two strands. The cost-minimization first-order conditions are imposed, in the spirit of Hall (1988) and Klette (1999)—now through cost shares rather than revenue shares—directly into the estimation step, combined with the modern proxy-variable techniques, following De Loecker and Warzynski (2012). The use of cost shares circumvents the additional endogeneity problem of the Hall/Klette specifications, since cost shares are constructed from observed input expenditures alone and do not depend on output. The result is a framework that retains the firm-year-specific, structurally estimated markups of the *new* production approach while restoring the cost-minimization restriction central to the original Hall/Klette tradition.

6 Non-Neutral Technical Change and Identical Markups

Section 4 noted that the prevailing explanation for conflicting markups attributes them to non-neutral technical change. As equation (15) shows, cost minimization equates the ratio of output elasticities of two flexible inputs to the ratio of their costs. Because Hicks-neutral

productivity does not enter output elasticities, this ratio is a deterministic function of production function parameters and inputs, and cannot account for the variation in observed input cost ratios. Introducing an unobserved non-neutral productivity term makes the output elasticities stochastic and allows the model "to saturate" (Raval, 2023b).[14] Following this argument, an emerging literature (Doraszelski and Jaumandreu, 2019; Raval, 2023a; Demirer, 2026) employs production functions with both Hicks-neutral and labor-augmenting productivity and recovers identical markups across labor and materials. We now explain why this resolution arguably works: the cost-minimization conditions—absent from estimation in the standard framework—are explicitly invoked to identify labor-augmenting productivity.

6.1 A Model with Labor-Augmenting Productivity

Following Doraszelski and Jaumandreu (2019), we extend the translog production function in (7) by introducing labor-augmenting productivity, denoted by A^L_{it}. To make the model tractable and identify A^L_{it}, Doraszelski and Jaumandreu (2019) imposes two restrictions on the technology. First, the production function is separable in capital: $\beta_{KL} = \beta_{KM} = 0$. Second, it is homogeneous of degree $\beta_L + \beta_M$ in L and M, such that $-\beta_{LL} = -\beta_{MM} = \beta_{LM} \equiv \beta$. Under these restrictions, the production function simplifies to:

$$y_{it} = \beta_0 + \beta_K k_{it} + \tfrac{1}{2}\beta_{KK} k_{it}^2 + \beta_L (A^L_{it} + l_{it}) + \beta_M m_{it} - \tfrac{\beta}{2}(m_{it} - A^L_{it} - l_{it})^2 + A^H_{it} + \epsilon^Y_{it}. \quad (36)$$

The output elasticities of labor and materials are now functions of A^L_{it}:

$$\theta^L_{it} = \beta_L + \beta\,(m_{it} - A^L_{it} - l_{it}), \qquad \theta^M_{it} = \beta_M - \beta\,(m_{it} - A^L_{it} - l_{it}), \quad (37)$$

implying a constant short-run scale elasticity $\gamma^V_{it} \equiv \theta^L_{it} + \theta^M_{it} = \beta_L + \beta_M$.

[14] On page 2593.

Identification of A_{it}^L. The key to identifying labor-augmenting productivity is that it alters the firm's optimal input mix, whereas Hicks-neutral productivity only shifts the scale of input usage (Doraszelski and Jaumandreu, 2018; Zhang, 2019; Demirer, 2026). Taking the ratio of cost-minimizing first-order conditions for labor and materials yields:

$$\frac{W_{it}^L L_{it}}{W_{it}^M M_{it}} = \frac{\beta_L + \beta(m_{it} - A_{it}^L - l_{it})}{\beta_M - \beta(m_{it} - A_{it}^L - l_{it})}. \tag{38}$$

Solving for A_{it}^L:

$$A_{it}^L = (m_{it} - l_{it}) + \frac{\beta_L}{\beta} - \frac{\beta_L + \beta_M}{\beta} S_{it}^L, \tag{39}$$

where $S_{it}^L = W_{it}^L L_{it}/(W_{it}^L L_{it} + W_{it}^M M_{it})$ is the observed labor cost share in variable cost. The ratio of cost-minimizing first-order conditions is thus inverted to express A_{it}^L as a function of observables and parameters. Note that, unlike the standard approach, this procedure directly uses cost minimization for *identification* via the tangency conditions in (38).

Substituting the identified A_{it}^L from (39) into the output elasticities in (37) yields:

$$\theta_{it}^L = (\beta_L + \beta_M)\, S_{it}^L, \qquad \theta_{it}^M = (\beta_L + \beta_M)\, S_{it}^M, \tag{40}$$

where $S_{it}^M \equiv 1 - S_{it}^L$. The output elasticities of the two flexible inputs are proportional to their observed cost shares, with the common factor of proportionality equal to the short-run scale elasticity $\gamma^V \equiv \beta_L + \beta_M$. Further, substituting (39) into (36) eliminates the unobserved A_{it}^L from the production function, yielding (Doraszelski and Jaumandreu, 2019):

$$y_{it} = \beta_0 + \frac{\beta_L^2}{2\beta} + \beta_K k_{it} + \tfrac{1}{2}\beta_{KK} k_{it}^2 + (\beta_L + \beta_M)\, m_{it} - \frac{(\beta_L + \beta_M)^2}{2\beta} (S_{it}^L)^2 + A_{it}^H + \epsilon_{it}^Y. \tag{41}$$

To estimate equation (41), we must remove the unobserved Hicks-neutral productivity term A_{it}^H, which is accomplished via a dynamic panel transformation. We use the same Colombian

Annual Manufacturing Survey data as in Section 3.[15] Table 3 and Figure 3 report the markup estimates and trends.

Table 3 and Figure 3 confirm what is already implied in the foregoing analysis: markups recovered from labor and materials coincide at every firm-year data point. The equality is, essentially, by construction:

$$\hat{\mu}_{it}^{L} \equiv \frac{\hat{\theta}_{it}^{L}}{R_{it}^{L}} = \frac{(\hat{\beta}_L + \hat{\beta}_M)\, S_{it}^{L}}{R_{it}^{L}} = \hat{\gamma}^{V}\, \frac{P_{it}Q_{it}}{W_{it}^{L}L_{it} + W_{it}^{M}M_{it}} = \frac{\hat{\theta}_{it}^{M}}{R_{it}^{M}} \equiv \hat{\mu}_{it}^{M}. \tag{42}$$

Recall that (39) recovers A_{it}^{L} by inverting the ratio of cost-minimizing first-order conditions. The output elasticities in (40) therefore satisfy $\theta_{it}^{L}/\theta_{it}^{M} = S_{it}^{L}/S_{it}^{M}$, yielding identical markups. This framework and the SNTL thus reach invariance by the same route: both tie the output elasticities of the flexible inputs to their observed cost shares, so that (12) holds at every observation. What differs is how that restriction is arrived at—here by identifying A_{it}^{L}, in Section 5 by imposing the conditions directly.

7 Discussion

At this juncture, we wish to note that our analysis is not meant to detract from the relevance of factor-augmenting productivity. A large theoretical and empirical literature has explored factor-biased technical change, motivated in part by persistent firm-level heterogeneity in input ratios that is hard to reconcile with Hicks-neutral productivity (Doraszelski and Jaumandreu, 2018; Zhang, 2019; Raval, 2019; Demirer, 2026; Wang, 2025). Our point is narrower and directed at a specific problem: the lack of invariance in markup estimates can be resolved by cost-minimization conditions alone, without invoking factor-augmenting productivity.

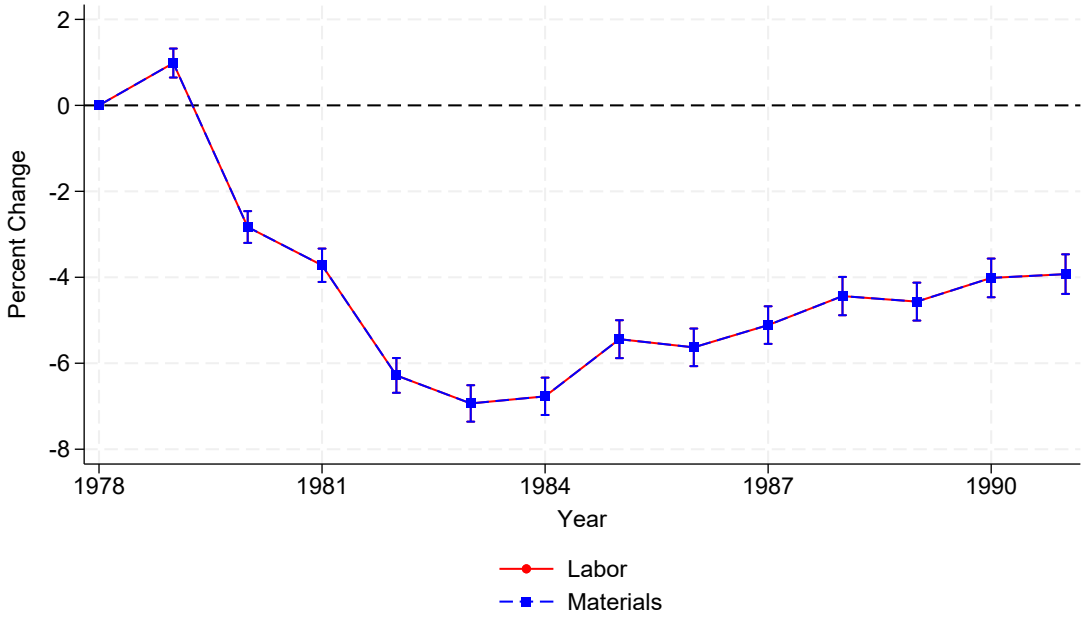
[15] As noted earlier, prior to estimation, input data are normalized to equal one at the (industry) geometric mean, similar to the procedure used in Section 5. Cost shares and revenue shares are functions of observed expenditures and are invariant to this rescaling. Doraszelski and Jaumandreu (2019) further restrict $\beta_{KK} = 0$ in estimating (41); we, however, do not impose that restriction. Estimation details are provided in Appendix C.

Table 3: Markup Estimates Based on Doraszelski and Jaumandreu (2019)

Input	Mean	Median	SD	P10	P90
Labor	1.4176	1.3586	0.3480	1.0988	1.9443
Materials	1.4176	1.3586	0.3480	1.0988	1.9443

Note: Markup is computed as the output elasticity divided by the revenue share for each input.

Figure 3: Markup Trends (Doraszelski and Jaumandreu, 2019)



Notes: This figure plots the estimated year effects based on equation (10) and is comparable to figures 1 and 2. Vertical bars represent 95% confidence intervals based on clustering at the plant level. All estimates are relative to the first year, which is set to zero.

It is useful to place the SNTL and the model with labor-augmenting productivity (LAP) side by side to appreciate some of their implications. Both approaches consider cost-minimizing firms that take prices of flexible inputs as given, and ask why the *model-implied* cost shares of the standard approach differ from the *observed* cost shares. Both point towards a misspecification problem in the standard approach, but locate it differently: the SNTL attributes it to the inflexibility of parametric production functions, whose fixed set of parameters cannot reproduce the observed cost ratio at every data point; the LAP model attributes it to an omitted dimension of unobserved heterogeneity. Under both models, the equality of markup estimates across flexible inputs holds by construction and can therefore no longer be *tested*, as in Raval (2023b).

Identifying factor-augmenting productivity, however, requires assumptions that go beyond cost minimization. Both Doraszelski and Jaumandreu (2019) and Demirer (2026) recover A_{it}^L only after restricting the technology to the form $F_t\big(K_{it}, h_t(A_{it}^L L_{it}, M_{it})\big)$, where labor and materials enter only through the composite h_t, which is homogeneous, and F_t is increasing in h_t. For given capital, output is then an increasing transformation of a homogeneous aggregator of $(A_{it}^L L_{it}, M_{it})$. The technology is thus separable in capital and homothetic in labor and materials. As illustrated in Section 6, Doraszelski and Jaumandreu (2019) consider a restricted translog, making the technology homogeneous of a fixed degree $\gamma^V = \beta_L + \beta_M$ in $(A_{it}^L L_{it}, M_{it})$, the same for every firm-year. Demirer (2026) leaves F_t nonparametric, so homogeneity is not imposed: the short-run elasticity of scale, γ_{it}^V, varies with the levels of capital and the composite input, rather than being a common parameter. Raval (2023a), in turn, assumes constant returns to scale, and that the cost-minimization conditions hold, at least on average, for all inputs (including capital). Further, with J^V flexible inputs, cost minimization imposes $J^V - 1$ independent restrictions at each observation, and a single unobservable such as A_{it}^L can satisfy only one of them—conflicting markups may still arise when there are more than two flexible inputs (as in the case of energy and non-energy materials considered by Raval (2023b)). Fernald et al. (2026) make a related observation:

the labor-augmenting view rests on assumptions about the nature of productivity that are difficult to test, technological bias having long been recognized as hard to distinguish from differential patterns of substitution.

In general, the gap between model-implied and observed cost shares has candidate sources beyond the inflexibility of the production function and labor-augmenting productivity, including unobserved differences in input prices across firms (Grieco et al., 2016), wedges in the input first-order conditions such as monopsony power or adjustment costs (Rubens, 2023; Kirov and Traina, 2023; Doraszelski and Jaumandreu, 2019; Azzam et al., 2026), and measurement error in inputs (McElroy, 1987; Fox and Smeets, 2011; Gollin and Udry, 2021). In the ratio of first-order conditions, each of these enters as a single scalar, so they are not separately identified from production data alone (Raval, 2023b; Rubens et al., 2026). Separating them requires additional structure or data, such as firm-level input prices (Doraszelski and Jaumandreu, 2018),[16] a model of input supply (Rubens et al., 2026), or restrictions on the wedge (Azzam et al., 2026; Dobbelaere et al., 2025; Jaumandreu, 2026). This limitation applies equally to the SNTL model. It takes observed cost shares as the split of scale elasticity across flexible inputs; if an input's price departs from its shadow cost by a wedge, the split is incorrect, and the markup estimates are biased. In such cases, the SNTL could be extended to incorporate a wedge that is constant within an industry, as in Dobbelaere et al. (2025) and Jaumandreu (2026).

For an applied researcher, the choice between these models depends on the specific research question and on the industry at hand. For the purposes of estimating markups, the SNTL offers a flexible resolution of the misspecification problem, and it does so without assumptions beyond those the standard approach already maintains. Nevertheless, a procedure for choosing among the competing explanations would be desirable. Some of them place

[16]Firm-level input prices remove only one of these sources. Even with observed prices, labor-augmenting productivity and a labor-market wedge enter the ratio of first-order conditions as a single scalar, so separating them still requires additional structure or data (Rubens et al., 2026).

different observables in the output equation. Under separability in capital and homogeneity in the flexible inputs, substituting labor-augmenting productivity out through the ratio of first-order conditions makes output depend on the materials quantity and the labor cost share, with the labor quantity dropping out, as in Section 6 (Doraszelski and Jaumandreu, 2019), whereas an input-market wedge adds a share-weighted labor term to the equation (Jaumandreu, 2026). Such exclusion restrictions may be testable. We leave this undertaking to future work.

8 Additional Applications: Chile and the US Manufacturing Data

To reinforce our central point, this section repeats the exercise of estimating markups using manufacturing data from two more countries—Chile and the United States—employing the three methods: the standard approach (Raval, 2023b), the SNTL, and the production function with labor-augmenting productivity (Doraszelski and Jaumandreu, 2019). Details on the data and construction of variables are provided in Appendix B. The estimates, summarized in Table 4 and Figure 4, reproduce the pattern established for Colombia. Under the standard approach—which places no requirement that the data and the estimated production function be jointly consistent with cost minimization—the markups recovered from labor and materials diverge. Once the cost-minimization conditions are embedded in estimation, this divergence disappears: both the SNTL and the labor-augmenting model deliver markups that coincide across the two flexible inputs at every data point, and their estimated trends move together. As before, the SNTL achieves this invariance while maintaining a Hicks-neutral specification, underscoring that it is the imposition of cost minimization, not non-neutral technical change, that resolves the conflict.

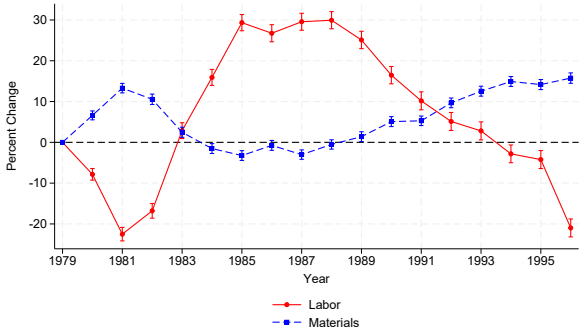
Table 4: Markup Estimates: Chile and the United States

	Mean	Median	SD	P10	P90
Panel A. ***Chile*** *(1979–1996)*					
Standard Approach					
Labor	2.15	1.69	1.75	0.82	3.91
Materials	1.28	1.20	0.46	0.89	1.69
SNTL					
Labor	1.30	1.24	0.42	0.92	1.74
Materials	1.30	1.24	0.42	0.92	1.74
Labor-Augmenting Productivity					
Labor	1.37	1.31	0.49	0.77	1.96
Materials	1.37	1.31	0.49	0.77	1.96
Panel B. ***United States*** *(1970–2010)*					
Standard Approach					
Labor	1.33	0.96	1.63	0.03	3.05
Materials	1.41	1.17	1.06	0.56	2.48
SNTL					
Labor	1.12	1.08	0.42	0.64	1.62
Materials	1.12	1.08	0.42	0.64	1.62
Labor-Augmenting Productivity					
Labor	1.40	1.38	0.36	0.99	1.84
Materials	1.40	1.38	0.36	0.99	1.84

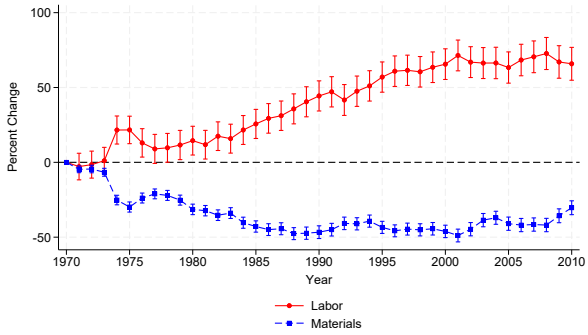
Notes: Markups are computed as the output elasticity divided by the revenue share for each input. Standard Approach refers to the translog production function estimated by the control-function procedure (Raval, 2023b); SNTL to the semi-nonparametric translog with the cost-minimization conditions imposed in estimation, and estimated using the control function approach; and Labor-Augmenting Productivity to the production function with labor-augmenting productivity estimated by the dynamic panel procedure of Doraszelski and Jaumandreu (2019). Under SNTL and Labor-Augmenting Productivity, markups are invariant across flexible inputs by construction, so the labor and materials rows coincide. Industries are defined at the three-digit level for Chile and at the two-digit level for the United States.

Figure 4: Markup Trends: Chile and the United States

Standard Approach

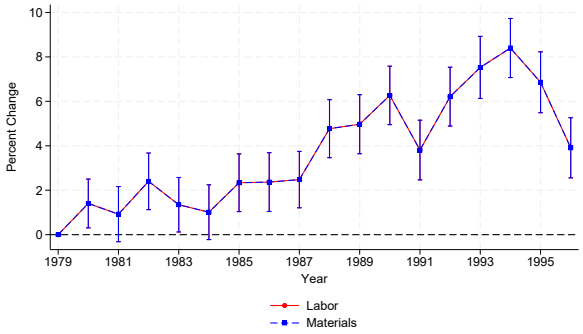


(a) Chile

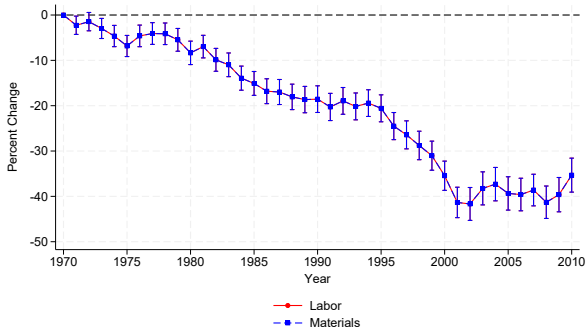


(b) United States

SNTL

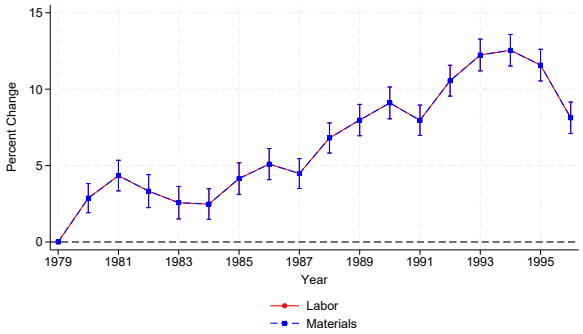


(c) Chile



(d) United States

Labor-Augmenting Productivity



(e) Chile



(f) United States

Notes: Each panel plots the estimated year effects in average markups based on equation (10). Vertical bars represent 95% confidence intervals based on clustering at the plant level. All estimates are relative to the first year, which is set to zero. Within each row, the left panel is Chile and the right panel is the United States. Standard Approach refers to the translog production function; SNTL to the semi-nonparametric translog; and Labor-Augmenting Productivity to the production function with labor-augmenting productivity.

9 A Transparency Diagnostic: Decomposing the Revenue Shares

The production approach recovers markups as a wedge between the output elasticity of a flexible input and its revenue share. As emphasized by Fernald et al. (2026), this makes production-based markups a residual—given the first-order condition, any variation in revenue share that is not accounted for by the output elasticity is attributed to the markups. Taking logs of (4) makes this explicit:

$$\ln R_{it}^j = \ln \theta_{it}^j - \ln \mu_{it}. \tag{43}$$

A general concern is how much of the observed variation of revenue shares is attributed to markups rather than to output elasticities. In the limiting case of a single output elasticity per industry-year, as in the Cobb-Douglas production function, the elasticity has no within-industry-year variation, and the markup absorbs all the variation by construction.

Quantifying the extent of share variation explained by output elasticities therefore is a useful diagnostic. Fernald et al. (2026) propose a simple transparency check, which we adopt here. We regress $\ln R_{it}^j$ on $\ln \hat{\theta}_{it}^j$ and report the R^2 as the proportion of variation that is linearly explained by the estimated output elasticity, and $1 - R^2$ as the residual share attributed to estimated markups. We first estimate a pooled regression, without fixed effects, and then add industry-year fixed effects, in which case the reported figure is the within R^2. The diagnostic is applied separately to labor and materials shares in revenue, and separately to each of the three estimators. Results are provided in Table 5.

Across all three countries, the estimators that impose cost minimization—the SNTL and the production function with labor-augmenting productivity—leave more of the revenue-share variation to be explained by output elasticities. The increase is present in nearly every comparison and is largest in the U.S. data, where the standard translog leaves a very small fraction of the variation to the estimated elasticities. These results indicate that the standard approach relies on markups to explain a large proportion of the variation in several

of these samples. Imposing the first-order conditions of cost minimization at the estimation stage—rather than only at the markup-recovery stage—materially reduces the explanatory importance of markups.

Table 5: An R^2 Decomposition of the Variation of Revenue Shares

	Standard Approach		SNTL		Labor-Augmenting Productivity	
	Elasticity	Markup	Elasticity	Markup	Elasticity	Markup
Panel A. ***Colombia*** *(1978–1991),* $N = 82{,}487$						
Pooled						
Labor	0.42	0.58	0.89	0.11	0.88	0.12
Materials	0.64	0.36	0.71	0.29	0.69	0.31
Industry-Year Fixed Effects						
Labor	0.41	0.59	0.88	0.12	0.97	0.03
Materials	0.74	0.26	0.71	0.29	0.92	0.08
Panel B. ***Chile*** *(1979–1996),* $N = 62{,}081$						
Pooled						
Labor	0.24	0.76	0.80	0.20	0.70	0.30
Materials	0.26	0.74	0.29	0.71	0.29	0.71
Industry-Year Fixed Effects						
Labor	0.26	0.74	0.78	0.22	0.85	0.15
Materials	0.24	0.76	0.20	0.80	0.54	0.46
Panel C. ***United States*** *(1970–2010),* $N = 62{,}377$						
Pooled						
Labor	0.01	0.99	0.63	0.37	0.80	0.20
Materials	0.16	0.84	0.54	0.46	0.82	0.18
Industry-Year Fixed Effects						
Labor	0.00	1.00	0.61	0.39	0.78	0.22
Materials	0.15	0.85	0.60	0.40	0.86	0.14

Notes: Each pair of cells reports the fit of a regression of the log revenue share on the log estimated output elasticity, $\ln R^j_{it} = b_1 + b_2 \ln \hat{\theta}^j_{it} + \varepsilon^j_{it}$, run separately by country, input, and estimator. The Elasticity columns report the R^2; the Markup columns report $1 - R^2$. Pooled rows report regressions without fixed effects; Industry-Year Fixed Effects rows add industry-year fixed effects and report the within R^2. Standard Approach refers to the translog production function estimated by the control-function procedure (Raval, 2023b); SNTL to the semi-nonparametric translog with the cost-minimization conditions imposed in estimation; and Labor-Augmenting Productivity to the production function with labor-augmenting productivity estimated by the dynamic panel procedure of Doraszelski and Jaumandreu (2019). Industries are defined at the three-digit level for Colombia and Chile and at the two-digit level for the United States. The diagnostic follows Fernald et al. (2026).

10 Conclusion

This paper revisits the puzzle of conflicting markup estimates, arising from the production function approach, documented by Raval (2023b) and others. We argue that the source of this conflict is that the standard estimation procedure, commonly used under Hicks-neutral productivity, imposes no requirement that the data and the parameterized production function be jointly consistent with the maintained assumption of cost minimization. The optimality conditions of cost minimization explicitly link input cost shares to the structure of the production function, but this relationship is not exploited in the estimation process. Allowing for non-neutral technical change resolves the conflict, but we have shown that this resolution rests on imposing the cost-minimization conditions that the standard approach excludes.

To support our main contention, we have developed a novel econometric framework that directly embeds the cost-minimization conditions—through cost shares of flexible inputs—into the production function structure. The resulting estimating equation delivers firm-year-specific markups that are invariant across flexible inputs by construction, while continuing to maintain *only* Hicks-neutral productivity. The framework bridges two traditions in the production approach to markup estimation: imposing cost-minimization conditions into estimation in the spirit of Hall (1988) and Klette (1999), and employing structural production function estimation procedures, similar to De Loecker and Warzynski (2012). Applied to plant- and firm-level manufacturing data from Colombia, Chile, and the United States, the transparency diagnostic proposed by Fernald et al. (2026) suggests that this novel estimation framework attributes a larger share of the observed variation of revenue shares to fundamentals of the production structure, rather than the markup hypothesis.

Our findings indicate that the puzzle of conflicting markup estimates may well reflect the failure to fully account for the maintained cost minimization assumption, rather than how technological change is modeled. The hypothesis of factor-augmenting productivity,

of course, matters for several interesting questions in the production-function literature—such as biased technical change, input-mix heterogeneity, and the declining labor share—but clearly it is not the only route to obtain markup estimates that are invariant across flexible inputs.

References

Ackerberg, Daniel A and Jan De Loecker, "Production Function Identification under Imperfect Competition," Technical Report, Mimeo, University of Leuven 2021.

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

Arellano, Manuel and Stephen Bond, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*, 1991, *58* (2), 277–297.

Azzam, Azzeddine, Jordi Jaumandreu, and Rigoberto Lopez, "Farmers, Workers, Neither, or Both: Who Is Exploited?," 2026.

Blundell, Richard and Stephen Bond, "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 1998, *87* (1), 115–143.

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

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data," *Journal of monetary economics*, 2021, *121*, 1–14.

Brand, James, "Estimating productivity and markups under imperfect competition," 2019.

Caves, Douglas W, Laurits R Christensen, and W Erwin Diewert, "The economic theory of index numbers and the measurement of input, output, and productivity," *Econometrica: journal of the Econometric Society*, 1982, pp. 1393–1414.

De Loecker, Jan and Chad Syverson, "An industrial organization perspective on productivity," in "Handbook of industrial organization," Vol. 4, Elsevier, 2021, pp. 141–223.

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

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

De Ridder, Maarten, Basile Grassi, and Giovanni Morzenti, "The hitchhiker's guide to markup estimation: assessing estimates from financial data," *Econometrica*, 2026, *94* (1), 137–168.

Demirer, Mert, "Production Function Estimation with Factor-Augmenting Technology: An Application to Markups," *Econometrica*, 2026. Forthcoming.

Dobbelaere, Sabien, Jordi Jaumandreu, and Jacques Mairesse, "Market Power and Technology in US Manufacturing," 2025.

Doraszelski, Ulrich and Jordi Jaumandreu, "R&D and productivity: Estimating endogenous productivity," *Review of economic studies*, 2013, *80* (4), 1338–1383.

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

_ **and** _ , "Using Cost Minimization to Estimate Markups," 2019.

_ **and** _ , "Reexamining the De Loecker & Warzynski (2012) method for estimating markups," 2021.

Fernald, John, Amit Gandhi, Dimitrije Ruzic, and James Traina, "Micro and Macro Perspectives on Production-Based Markups," *Journal of Economic Literature*, 2026. Forthcoming.

Flynn, Zach, James Traina, and Amit Gandhi, "Measuring markups with production data," *Available at SSRN 3358472*, 2019.

Foster, Lucia S, John C Haltiwanger, and Cody Tuttle, "Rising markups or changing technology?," Technical Report, National Bureau of Economic Research 2022.

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

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

Gollin, Douglas and Christopher Udry, "Heterogeneity, measurement error, and misallocation: Evidence from African agriculture," *Journal of Political Economy*, 2021, *129* (1), 1–80.

Gottardo, Giulio, "Production Functions with Noisy Data: A Flexible Cost Share Approach," 2026. University of Oxford and Pembroke College.

Grieco, Paul LE, Shengyu Li, and Hongsong Zhang, "Production function estimation with unobserved input price dispersion," *International Economic Review*, 2016, *57* (2), 665–690.

Hall, Robert E, "The Relation Between Price and Marginal Cost in US industry," *Journal of Political Economy*, 1988, *96* (5), 921–947.

Hashemi, Arshia, Ivan Kirov, and James Traina, "The production approach to markup estimation often measures input distortions," *Economics Letters*, 2022, *217*, 110673.

Jaumandreu, Jordi, "Robust Production Function Estimation when there is Market Power," 2024.

_ , "Semiparametric Estimation of Heterogeneous Firm-level Producton Functions," Technical Report, Centre for Economic Policy Research 2026.

Kirov, Ivan and James Traina, "Labor market power and technological change in US manufacturing," Technical Report, mimeo 2023.

_ , **Paolo Mengano, and James Traina**, "Measuring markups with revenue data," *International Journal of Industrial Organization*, 2026, p. 103263.

Klette, Tor Jakob, "Market power, scale economies and productivity: estimates from a panel of establishment data," *The Journal of Industrial Economics*, 1999, *47* (4), 451–476.

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

Marschak, Jacob and William H Andrews, "Random Simultaneous Equations and the Theory of Production," *Econometrica, Journal of the Econometric Society*, 1944, pp. 143–205.

McElroy, Marjorie B, "Additive General Error Models for Production, Cost, and Derived Demand or Share Systems," *Journal of Political Economy*, 1987, *95* (4), 737–757.

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

Raval, Devesh, "The Micro Elasticity of Substitution and Non-neutral Technology," *The RAND Journal of Economics*, 2019, *50* (1), 147–167.

_ , "A flexible cost share approach to markup estimation," *Economics Letters*, 2023, *230*, 111262.

_ , "Testing the Production Approach to Markup Estimation," *Review of Economic Studies*, 2023, *90* (5), 2592–2611.

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

_ , **Yingjie Wu, and Mingzhi Xu**, "Exploiting or augmenting labor?," *American Economic Review: Insights*, 2026, *8* (1), 72–89.

Shenoy, Ajay, "Estimating the Production Function under Input Market Frictions," *Review of Economics and Statistics*, 2021, *103* (4), 666–679.

Solow, Robert M, "Technical change and the aggregate production function," *The review of Economics and Statistics*, 1957, *39* (3), 312–320.

Traina, James, "Is aggregate market power increasing? production trends using financial statements," *Production Trends Using Financial Statements (February 8, 2018)*, 2018.

Wang, Ziyao, "Non-neutral Technological Change in Chinese Manufacturing," *Available at SSRN 5176447*, 2025.

Zhang, Hongsong, "Non-neutral Technology, Firm Heterogeneity, and Labor Demand," *Journal of Development Economics*, 2019, *140*, 145–168.

Appendices

A Estimation Procedures

The control function approach to production function estimation—introduced by Olley and Pakes (1996) and extended by Levinsohn and Petrin (2003) and Ackerberg et al. (2015)—makes the following assumptions. First, observed output differs from firms' planned output by an unobservable term (whereas this is not the case for inputs). Second, productivity is Hicks-neutral; the productivity term is known to firms when they make their production decision but it is unobserved by the econometrician. The production function is thus represented as:

$$Y_{it} = \underbrace{F(\mathbf{X}_{it}; \boldsymbol{\beta}) \exp(A_{it}^H)}_{\text{Planned Output}} \exp(\epsilon_{it}^Y), \tag{44}$$

where A_{it}^H denotes Hicks-neutral productivity, ϵ_{it}^Y collects measurement error and shocks to output that are unanticipated by the firm at time t, Y_{it} is the observed output, and $F(\mathbf{X}_{it}; \boldsymbol{\beta}) \exp(A_{it}^H)$ equals Q_{it}, the firm's planned output. Log-transforming equation (44) yields:

$$y_{it} = f(\mathbf{x}_{it}; \boldsymbol{\beta}) + A_{it}^H + \epsilon_{it}^Y, \tag{45}$$

where lower-case letters represent log-transformed values. We consider the standard scenario in which inputs include capital (K_{it}), labor (L_{it}), and materials (M_{it}), with capital treated as a predetermined input.

The primary challenge in estimating output elasticities from equation (5) is that the unobserved Hicks-neutral productivity is correlated with input choices. That is, firm i's information set contains current and past productivity terms (it is assumed that it does not include future productivity terms). To address the resulting simultaneity, Olley and Pakes (1996) propose using observed demand for a factor input as a proxy for the unobserved productivity. Following Levinsohn and Petrin (2003) and Ackerberg et al. (2015), we rely

on materials demand, which, from the postulated production function, can be written as:

$$m_{it} = m_t(k_{it}, l_{it}, A^H_{it}; \boldsymbol{\beta}). \tag{46}$$

Equation (46) embodies the scalar unobservability assumption of the control function approach to production function estimation: the only unobservable entering $m_t(\cdot)$ is the productivity term.[17] The control function procedure further assumes that $m_t(\cdot)$ is strictly monotonic in productivity. With these two assumptions, equation (46) can be inverted to proxy for the unobserved productivity as $A^H_{it} = m_t^{-1}(k_{it}, l_{it}, m_{it}; \boldsymbol{\beta})$. Substituting for A^H_{it} in equation (5) yields:

$$y_{it} = f(k_{it}, l_{it}, m_{it}; \boldsymbol{\beta}) + m_t^{-1}(k_{it}, l_{it}, m_{it}; \boldsymbol{\beta}) + \epsilon^Y_{it} \equiv \phi_t(k_{it}, l_{it}, m_{it}; \boldsymbol{\beta}) + \epsilon^Y_{it}, \tag{47}$$

The reduced-form expression $\phi_t(\cdot)$ is estimated non-parametrically as a flexible function of observed inputs. This step constitutes the first stage of the control function procedure and provides estimates of planned output ($\hat{\phi}_{it}$) and the unobserved term ($\hat{\epsilon}^Y_{it}$).[18]

The second stage of the control function procedure estimates the production function parameters, $\boldsymbol{\beta}$, by assuming that productivity follows a first-order Markov process:

$$A^H_{it} = g_t(A^H_{it-1}) + \eta^H_{it}, \tag{48}$$

where η^H_{it} is the innovation in productivity. Using the first-stage estimate $\hat{\phi}_{it}$, productivity is then recovered as a function of $\boldsymbol{\beta}$ as $\hat{A}^H_{it}(\boldsymbol{\beta}) = \hat{\phi}_{it} - f(k_{it}, l_{it}, m_{it}; \boldsymbol{\beta})$. Non-parametrically regressing $\hat{A}^H_{it}(\boldsymbol{\beta})$ on $\hat{A}^H_{it-1}(\boldsymbol{\beta})$ yields the estimated innovation $\hat{\eta}^H_{it}(\boldsymbol{\beta})$. The moment conditions

[17] It has been noted that this assumption implicitly restricts the form of output market competition, raising questions as to its implications for recovering markups. See Doraszelski and Jaumandreu (2019, 2021), Jaumandreu (2024), De Loecker and Syverson (2021), and Demirer (2026) for more discussion.

[18] Recall that an estimate of planned output is required to calculate the revenue shares used in equation (4).

used to estimate $\boldsymbol{\beta}$ are:

$$\mathrm{E}\big[\mathbf{Z}_{it} \cdot \eta_{it}^{H}\big] = \mathbf{0}, \tag{49}$$

where $\mathbf{Z}_{it}$ includes current and lagged values of the predetermined input (capital), lagged values of flexible inputs, and their combinations, with the exact set determined by the functional form of f.

An alternative estimation approach is the dynamic panel method (Arellano and Bond, 1991; Blundell and Bond, 1998, 2000), which requires the productivity evolution to follow a first-order *linear* Markov process:

$$A_{it}^{H} = \alpha_{H}^{0} + \alpha_{H}^{1} A_{it-1}^{H} + \eta_{it}^{H}, \tag{50}$$

where α_{H}^{1} captures the persistence of the productivity process.[19] Using the linearity of productivity evolution, this method differences out the persistent part of productivity. To illustrate, substitute (50) in (5):

$$y_{it} = f(\mathbf{x}_{it}; \boldsymbol{\beta}) + \alpha_{H}^{0} + \alpha_{H}^{1} A_{it-1}^{H} + \eta_{it}^{H} + \epsilon_{it}^{Y} \tag{51}$$

Hence, the unobserved productivity term A_{it-1}^{H} can be removed by implementing the so-called dynamic panel transformation:

$$y_{it} = \alpha_{H}^{1}\, y_{it-1} + f(\mathbf{x}_{it}; \boldsymbol{\beta}) - \alpha_{H}^{1}\, f(\mathbf{x}_{it-1}; \boldsymbol{\beta}) + \alpha_{H}^{0} + u_{it}, \tag{52}$$

where $u_{it} \equiv \eta_{it}^{H} + \epsilon_{it}^{Y} - \alpha_{H}^{1}\, \epsilon_{it-1}^{Y}$. Equation (52) can be estimated using similar moment conditions as earlier, with instruments such as current and lagged values of pre-determined inputs, lagged values of flexible inputs, and their combinations. Crucially, the dynamic panel method, relying entirely on the linearity of the productivity process, does not require

[19] We may replace the constant, α_{H}^{0}, with time fixed-effects.

inverting a factor demand function or assuming scalar unobservability.[20]

Estimation Details. We estimate equations (7) and (33) in Sections 3 and 5 following the control-function procedure of Raval (2023b) and De Loecker and Warzynski (2012). Specifically, in the first stage, y_{it} is regressed on a third-order polynomial in inputs and year fixed effects. For equation (33), the inputs also include the cost-share-weighted index z_{it}. The fitted values $\hat{\phi}_{it}$ are output net of measurement error, and the residuals $\hat{\epsilon}^Y_{it}$ are used to compute revenue shares at planned output. In the second stage, Hicks-neutral productivity is assumed to follow a linear first-order Markov process, $A^H_{it} = \alpha^0_H + \alpha^1_H A^H_{it-1} + \eta^H_{it}$. For a given parameter vector, productivity is recovered as $\hat{\phi}_{it}$ minus the production function, and the innovation η^H_{it} is the residual from regressing productivity on its lag. Any change in β_0 shifts recovered productivity by a constant, which the intercept of the productivity process absorbs while leaving the innovation unchanged. Hence β_0 and α^0_H are not separately identified. This has no bearing on the estimates of interest, as β_0 enters neither the output elasticities nor the markups. The parameters are estimated by GMM using the instruments specified in equations (8) and (34).

B Data

Colombian Manufacturing Survey Data

The Colombian data are drawn from the annual Manufacturing Census (*Encuesta Anual Manufacturera*) compiled by the *Departamento Administrativo Nacional de Estadística* (DANE), covering the years 1978–1991, and include approximately 7,000 plants per year. We obtain the data from the replication package accompanying Raval (2023b), and follow Raval (2023b) exactly in the construction and cleaning of the sample.

[20] See De Loecker and Syverson (2021), Shenoy (2021), and Jaumandreu (2024) for a comparison between these two estimation procedures.

Chilean Manufacturing Survey Data

The Chilean data are drawn from the Annual National Industrial Survey (*Encuesta Nacional Industrial Anual*, ENIA) compiled by the *Instituto Nacional de Estadística* (INE), covering the years 1979–1996, and include approximately 5,000 manufacturing plants per year. As with Colombia, we obtain the data from the replication package accompanying Raval (2023b), and follow Raval (2023b) exactly in the construction and cleaning of the sample.

U.S. Compustat Data

The U.S. data are an unbalanced panel of publicly listed manufacturing firms from Compustat (Standard & Poor's), accessed through Wharton Research Data Services (WRDS), covering the years 1970–2010. We follow Raval (2023b) in constructing the sample.

C Estimation of the Labor-Augmenting Translog Production Function

This appendix details the dynamic panel estimation of equation (41), following Doraszelski and Jaumandreu (2019). As in the text, Hicks-neutral productivity is assumed to follow a linear first-order Markov process:

$$A_{it}^{H} = \alpha_{H}^{0} + \alpha_{H}^{1} A_{it-1}^{H} + \eta_{it}^{H}, \tag{53}$$

where α_H^0 is a constant, α_H^1 is the persistence parameter, and η_{it}^H is the innovation shock. Substituting (53) into (41) and differencing out the persistent component A_{it-1}^H using the lagged version of (41) yields:

$$\begin{aligned} y_{it} = \gamma_0 + \alpha_H^1 y_{it-1} + \beta_K(k_{it} - \alpha_H^1 k_{it-1}) + \tfrac{1}{2}\beta_{KK}(k_{it}^2 - \alpha_H^1 k_{it-1}^2) \\ + (\beta_L + \beta_M)(m_{it} - \alpha_H^1 m_{it-1}) - \frac{(\beta_L + \beta_M)^2}{2\beta}\left[(S_{it}^L)^2 - \alpha_H^1 (S_{it-1}^L)^2\right] + u_{it}, \end{aligned} \tag{54}$$

where $\gamma_0 = \big(\beta_0 + \frac{\beta_L^2}{2\beta}\big)(1-\alpha_H^1) + \alpha_H^0$ and the composite error is $u_{it} = \eta_{it}^H + \epsilon_{it}^Y - \alpha_H^1\,\epsilon_{it-1}^Y$. The parameters to be estimated are:

$$\boldsymbol{\beta} = \big(\gamma_0,\ \alpha_H^1,\ \beta_K,\ \beta_{KK},\ \gamma^V,\ \delta\big)', \tag{55}$$

where $\gamma^V = \beta_L + \beta_M$ and $\delta = (\beta_L + \beta_M)^2/(2\beta)$, giving six free parameters. The individual coefficients β_0, α_H^0, β_L, and β_M are not separately identified. The instrument vector is:

$$Z_{it} = \big(1,\ k_{it},\ k_{it}^2,\ m_{it-1},\ l_{it-1},\ (S_{it-1}^L)^2\big)'. \tag{56}$$

Capital is treated as predetermined, so k_{it} and k_{it}^2 enter contemporaneously; labor and materials are flexible, so the lags m_{it-1}, l_{it-1}, and $(S_{it-1}^L)^2$ are used. To obtain an estimate of planned output, we run an auxiliary regression of y_{it} on a third-order polynomial in $(k_{it}, l_{it}, m_{it}, S_{it}^L)$ and year fixed effects. Including S_{it}^L in the polynomial introduces input price ratio variation that the standard (k, l, m) polynomial would not capture in the presence of labor-augmenting productivity, in the spirit of Doraszelski and Jaumandreu (2019).