Factor Bias and Heterogeneous Output Quality in Multiproduct Production*

Yi Lee†, National Tsing Hua University
Shengyu Li‡, The University of New South Wales
Mark J. Roberts§, The Pennsylvania State University and NBER

May 2026

Abstract

This paper develops and estimates a structural model of multiproduct production that allows plant-product quality and factor-augmenting input efficiency. Using plant-level data from Taiwan's textile industry, we specify a translog multiproduct cost function with input share and output supply equations and estimate the system of equations via GMM. The model recovers plant-specific input efficiencies, output qualities, economies of scale and scope, and substitution elasticities. It does not impose restrictions for non-joint production or input-output separability and does not require imputing product-level input allocations. We document increasing returns to scale, substitutability between labor and materials, and cost complementarities from joint production. Variation in labor shares and product mixes are jointly driven by large and persistent heterogeneity in both input efficiencies and output qualities. A multilateral productivity index shows that variation in input efficiency and output quality is the dominant source of productivity differences across plants.

Keywords: *multiproduct production, translog cost function, factor bias technology, output quality, productivity*

JEL classification: *D24, L11, L15, O47.*

*The authors thank Mauro Caselli, Erwin Diewert, Uli Doraszelski, Paul Grieco, Kevin Fox, Joonkyo Hong, Davide Luparello, Scott Orr, Amil Petrin, Chad Syverson, Daniel Xu, Hongsong Zhang and many seminar and conference participants for very helpful comments. All errors are the authors' responsibility.

†Department of Economics, National Tsing Hua University. Email: ylee@mx.nthu.edu.tw.

‡School of Economics & Centre for Applied Economic Research, Business School, the University of New South Wales, Australia. Email: shengyu.li@unsw.edu.au.

§Department of Economics, the Pennsylvania State University. Email: mroberts@psu.edu.

1 Introduction

Two broad empirical patterns are commonly observed across manufacturing industries and countries. First, there has been a persistent decline in labor's share of revenue over time. Second, large multiproduct firms increasingly dominate aggregate output and sales. While many economic forces may contribute to these trends, some have been identified as particularly important. Biased technological change—specifically, labor-augmenting improvement—raises labor productivity. When combined with a low elasticity of substitution between inputs, such changes lead to a declining labor share. Also, economies of scale and scope give cost advantages to firms that are both large and diversified, allowing them to produce more efficiently across a range of products. In addition, cost-driven concentration is amplified when larger firms also produce higher-quality products, further increasing their market share. Although each of these forces has been studied individually, they are inherently interconnected, as they reflect different dimensions of a firm's underlying production technology.

In this paper, we develop an empirical model to jointly estimate these key forces shaping plant behavior and performance, using data on multiproduct plants in Taiwan's textile industry. The model is centered on a flexible translog multiproduct cost function, complemented by input cost share and output supply equations. It allows for plant heterogeneity in both output quality and factor-augmenting input efficiency, treating these as latent plant characteristics that influence endogenous input and output choices, while preserving joint production and input-output non-separability and avoiding product-level input allocation.

The production framework is highly flexible. It allows key features of the technology, such as input substitution elasticities, economies of scale and scope, marginal rates of transformation, marginal costs, and input cost shares, to depend on the full set of outputs, variable input prices, and fixed factors, rather than being fixed parameters. The model does not impose separability between inputs and outputs, thereby capturing the interdependencies between production decisions across products. In contrast to recent approaches that estimate multiproduct technologies by allocating measured input totals to individual products, our method does not require detailed product-level input data. Furthermore, by allowing input-specific efficiencies and product-specific quality to vary across plants and evolve over time, the model captures persistence and heterogeneity in these terms. We use the estimated translog multiproduct cost function to construct a plant-level productivity index that can be decomposed into the contributions of input efficiency, product quality, capital stock, and scale economies.

The structural model of multiproduct cost assumes that plants choose labor and material

inputs to minimize variable costs, conditional on fixed capital and output levels. Inputs and input prices are expressed in efficiency units, while output levels and prices are expressed in quality-adjusted units. The variable cost function is specified as a translog function of input prices, output levels, and capital stock. As a part of the structural model, input cost share equations are derived from Shephard's lemma, stating that the cost elasticity with respect to an input price yields the cost-minimizing share of expenditure on that input. At the same time, short-run output supply equations follow from the plant's profit-maximization conditions, which implies that the ratio of revenue to cost equals the elasticity of variable cost with respect to each output adjusted by a markup. Unobserved output quality and input efficiency terms are treated as structural errors in the empirical model and are modeled as evolving according to first-order Markov processes.

Estimation proceeds in two parts. First, product-specific demand functions are estimated to recover the demand elasticity for each output, which, together with the assumed conduct model, implies output-specific markups. In the application, the demand functions are specified in constant-elasticity form, and instrumental variables are employed to address potential endogeneity of prices, but more general demand systems can be incorporated. Second, we estimate the cost-function parameters. We begin with the subset of parameters that appear in the input share and output supply equations, using GMM with lagged plant characteristics and characteristics of rival plants as instruments. We then estimate the remaining parameters associated with the fixed factor of production, namely capital stock, using the cost equation. After estimation, the unobserved output quality and input efficiency terms are recovered as linear functions of observable variables using the plant's first-order conditions.

We estimate the model using plant-level panel data from the Taiwanese manufacturing census, focusing on two major product categories: cotton fiber and man-made fiber, over the period 1992 to 2004. The dataset provides information on each plant's sales revenue and output quantity for both products, as well as plant-level inputs, including employment, capital stock, and expenditures on labor and materials.

The data reject both input-output separability and non-joint production, indicating that the flexible multiproduct cost structure is not merely a theoretical refinement but is empirically important. First, because of input-output non-separability, the same change in factor prices or factor-biased technology can affect the supply of cotton fiber and man-made fiber by different magnitudes, thereby generating variation in product mix across plants. At the same time, the results show that labor shares (i.e., labor cost to variable cost ratio) are

shaped not only by factor-biased technology but also by output quality.[1] Importantly, this output-quality channel is absent from much of the existing literature on biased technological change. Second, we find evidence of cost complementarities across outputs arising from joint production. Specifically, a 1 percent increase in the output of man-made fiber lowers the marginal cost of cotton fiber by approximately 0.11 percent, while a 1 percent increase in the output of cotton fiber reduces the marginal cost of man-made fiber by about 0.10 percent. Short-run economies of scope are evident as well. On average, the estimated index is 1.255, implying that producing the two products separately increases short-run variable costs by about 25.5 percent relative to producing them jointly. These results imply that ignoring joint production or non-separability would mismeasure marginal cost, scope economies, and sources of labor-share variation in multiproduct plants.

Moreover, estimates of multiproduct short-run economies of scale, defined as the inverse of the sum of cost elasticities with respect to outputs, average 1.25, indicating increasing returns to scale. This measure declines with output levels but remains above one for approximately 82 percent of plant observations. The estimated elasticity of substitution between labor and materials averages 2.19, suggesting a high degree of substitutability between these inputs.

We recover plant- and year-specific measures of labor efficiency, materials efficiency, and the qualities of the two outputs, and find that all four exhibit statistically significant upward trends over time, ranging from 1.5 percent to 6.2 percent per year. Plant-level heterogeneity is also substantial and persistent, accounting for between 62.1 percent and 78.7 percent of the total variation in these measures. Among these factors, heterogeneity in output quality emerges as the primary driver of variation in both labor cost shares and revenue-to-cost ratios, surpassing the role of labor efficiency, which has been the central focus of much of the literature on biased technological change.

To assess the importance of each of these factors, we construct a multilateral multiproduct productivity index (MPP) that measures differences in the plants' variable cost after netting out the effects of factor price variation. Capital stock contributes positively to MPP by lowering variable costs. Output expansions increase (decrease) productivity when there are increasing (decreasing) returns to scale. Improvements in input efficiency raise MPP by reducing effective input costs, while higher output quality increases production costs and therefore lowers MPP. Among all components, variations in input efficiencies and output qualities are the dominant sources of heterogeneity in MPP across plants.

[1] In our context, we refer to labor share as the labor cost to variable cost ratio throughout the paper unless stated otherwise.

The remainder of this paper is organized as follows. Section 2 summarizes the literature. We introduce the theoretical model of multiproduct production in Section 3. Section 4 presents the empirical framework and estimation strategy. Section 5 describes the data. The estimation results are presented in Section 6. In Section 7, we analyze the output quality and input efficiency terms as well as the heterogeneity of cost structure among plants. Section 8 constructs the multilateral productivity index and examines its components. We conclude in Section 9.

2 Literature Review

2.1 Multiproduct Production Estimation

The literature on multiproduct production originated with the specification of transformation functions that characterize a firm's capacity to produce a vector of outputs from a vector of inputs. This foundational work was extended through the development of duality theory, which established formal relationships between transformation functions and multiproduct cost and profit functions (e.g., Diewert, 1973; McFadden, 1978). Represent the technology as the transformation function $T(\boldsymbol{Q}, \boldsymbol{X}) = 0$, where $\boldsymbol{Q}$ is a vector of outputs and $\boldsymbol{X}$ is a vector of inputs. If the vector of inputs is variable with factor price vector $\boldsymbol{W}$, the multiproduct total cost function is $C(\boldsymbol{W}, \boldsymbol{Q})$, representing the least cost combination of inputs for producing $\boldsymbol{Q}$ given $\boldsymbol{W}$; if a subset of the inputs (e.g., K) is fixed, the multiproduct variable cost function is $C(\boldsymbol{W}, K, \boldsymbol{Q})$ where $\boldsymbol{W}$ is the vector of variable input prices. Hall (1973) developed the relationship between restricted forms of the transformation function and cost function. Non-joint production is represented by separate production functions for each output, $T(Q_1(\boldsymbol{X}_1), \ldots, Q_N(\boldsymbol{X}_N)) = 0$. The corresponding cost function is the sum of separate product-specific cost functions $C(\boldsymbol{W}, \boldsymbol{Q}) = C_1(W, Q_1) + \cdots + C_N(W, Q_N)$. The implication is that the marginal cost of each output is independent of the level of any other output. Technology with input-output separability has the form $T = -g(\boldsymbol{Q}) + f(\boldsymbol{X})$ with corresponding cost function $C(\boldsymbol{W}, \boldsymbol{Q}) = \phi(\boldsymbol{W})h(\boldsymbol{Q})$. The implication is that the ratio of marginal costs of any two outputs, which equals the marginal rate of transformation between them, is not a function of factor prices.

These theoretical advances spurred a wave of empirical studies estimating multiproduct cost functions, with a focus on flexible functional forms that relaxed restrictive assumptions such as non-joint production, input-output separability, constant elasticities of substitution, and fixed returns to scale. A key early contribution was Burgess (1974), who applied a translog multiproduct cost function to aggregate time-series data, modeling the relationship between inputs (capital, labor, and imported materials) and outputs (consumption and investment

goods). His analysis emphasized that the translog form does not impose separability between inputs and outputs, a restriction that his empirical results reject.

Subsequent empirical applications of translog multiproduct cost functions proliferated across a range of industries using micro-level panel data. Notable examples include studies of trucking (Spady and Friedlaender, 1978), airlines (Caves et al., 1980b, 1984), railroads (Brown et al., 1979; Caves et al., 1980a, 1981a,b), electric utilities (Fuss and Waverman, 1981; Nelson et al., 1987), and telecommunications (Denny et al., 1981; Evans and Heckman, 1984). These empirical developments were accompanied by important theoretical contributions, such as the formalization of economies of scale and scope in multiproduct settings by Baumol et al. (1982).

By the late 1980s, the literature on empirical models of multiproduct technologies, particularly those based on the multiproduct cost function, was well developed. Empirical applications typically relied on plant- or firm-level data comprising vectors of physical outputs (such as kilowatt hours or ton-miles of shipments) and vectors of inputs and input prices, commonly including capital, labor, materials, and energy. Factor demand equations, most often specified as input cost share equations, played a central role in the estimation strategy. Productivity change was typically incorporated through a time trend or year fixed effects, with the elasticity of cost with respect to time interpreted as a measure of technical progress. When a flexible functional form, such as the translog, was used for the cost function, key economic measures such as factor demands, substitution elasticities, productivity growth, economies of scale, and non-Hick's neutral factor biases were allowed to vary between observations, depending on the output mix, factor prices, and time.

A key limitation of this literature, however, was the underdevelopment of the error structure in the estimating equations. The residuals in the cost and factor demand equations were often attributed to "errors in optimization" by the firm, rather than modeled as arising from economically meaningful sources such as heterogeneity in input efficiency or output quality, or technology shocks that are known to the firm. Factor prices were commonly treated as exogenous, and while instrumental variable methods were occasionally used to address the endogeneity of output quantities, a coherent structural interpretation of the error terms was often lacking.

Research focusing on production function estimation has emphasized the endogeneity of inputs and outputs that arises from firm-level heterogeneity, specifically, characteristics that are observed by firms but unobserved by researchers. Griliches and Mairesse (1995) emphasized that productivity shocks, though unobserved by the econometrician, are typically

known to firms and influence their input decisions, leading to simultaneity bias in standard estimation approaches. They advocated for the use of panel data and instrumental variable techniques to achieve credible identification.

Building on this insight, a number of structural econometric methods have been developed for production function estimation to control for unobserved productivity that varies not only over time but also across firms. Notable contributions include Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg et al. (2015), and Gandhi et al. (2020), who propose methodologies for identifying productivity using proxy variables and control functions. This includes assumptions such as the firm's information set to justify timing restrictions and cost minimization (or profit maximization) to justify incorporating factor demand equations in the estimation.[2] These approaches have become standard in estimating single-output production functions, particularly in empirical studies that adopt a Cobb-Douglas specification and model productivity as a Hicks-neutral technology shifter.[3]

The growing availability of micro-level datasets containing information on both output quantities and prices for multiple products produced by individual firms has spurred renewed interest in estimating models of multiproduct production, particularly to recover product-specific productivity indices and marginal costs. A common starting point in this recent literature is to assume a non-joint transformation function, treating multiproduct production as a collection of independent single-product technologies. In addition to this restriction on the multiproduct technology, this approach introduces additional complications by requiring data on the inputs allocated to each product. However, micro datasets on producers typically report the total use of each input (capital, labor, and intermediates) at the plant or firm level.

To address this data limitation, De Loecker et al. (2016), Orr (2022), and Valmari (2023) assume a non-joint production structure and develop methods to recover unobserved, mutually exclusive, and exhaustive input shares for each product. Expressing their framework in terms of the underlying transformation function, De Loecker et al. (2016) implicitly begin with a transformation function $T\big(\exp(\omega)Q_1(\boldsymbol{X}_1),\ \ldots,\ \exp(\omega)Q_N(\boldsymbol{X}_N)\big) = 0$, where each individual production function $Q_n(\boldsymbol{X}_n)$ is specified as a translog function, and $\exp(\omega)$ represents a technology index that varies across firms and time but not across products. They estimate

[2] Jaumandreu (2025) discusses how to extend the production models to environments where factor demand is influenced by market power. He compares control function and dynamic panel estimation methods and develops a production function estimator that is robust to market power.

[3] De Loecker and Syverson (2021) summarizes this literature and discusses the differences between the production function and cost function estimation of technology.

these separate production functions using data on single-product firms and infer the corresponding input bundles $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ and the firm-level technology term ω for multiproduct firms by exploiting a system of equations for input shares implied by the non-joint production structure. Orr (2022) and Valmari (2023) generalize the transformation function to allow for product-level productivity terms, $T\big(\exp(\omega_1)Q_1(\boldsymbol{X}_1), \ldots, \exp(\omega_N)Q_N(\boldsymbol{X}_N)\big) = 0$, while maintaining the assumption of non-joint production. They incorporate information from product demand to recover the input allocations $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ for each firm from the first-order conditions of profit maximization, using observed prices, quantities, and estimated demand elasticities.

The translog cost function applications summarized above do not impose the assumption of non-joint production and do not need to allocate inputs uniquely across outputs. Several recent studies also estimate models of multiproduct technology without assuming non-joint production. Maican and Orth (2021) estimate a model for multiproduct retail firms under the assumption that outputs and inputs are separable, and their model incorporates firm-level productivity, demand shocks, and a parameter capturing economies of scope in the number of products sold. Caselli et al. (2026) estimate a transformation function model with product-specific productivity of the form: $T\big(G(\exp(\omega_1)Q_1, \ldots, \exp(\omega_n)Q_n), F(\boldsymbol{X})\big) = 0$. While this specification allows for joint production, it maintains input–output separability.[4] Their empirical strategy exploits firms' profit maximization conditions to establish a one-to-one mapping between the observed data and the unobserved product-specific productivity terms $(\omega_1, \ldots, \omega_n)$ without relying on assumptions on how productivity evolves. Dhyne et al. (2024) develop a methodology to estimate a general transformation function with product-specific productivities: $T\big(\exp(\omega_1)Q_1, \ldots, \exp(\omega_n)Q_n, \boldsymbol{X}\big) = 0$. The framework allows for both joint-production and input–output non-separability. Building on theoretical work by Diewert (1973) and Lau (1976), they show that the general production structure can be estimated from a system of equations where each output Q_n is specified as a function of the vector of firm-level inputs $\boldsymbol{X}$, the vector of other outputs $\boldsymbol{Q}_{-n}$, and a product-specific productivity term, $Q_n = \exp(\omega_n)F_n(\boldsymbol{Q}_{-n}, \boldsymbol{X})$. They show how to use the estimates to recover the marginal cost of each output.

The multiproduct cost function model we develop in this paper incorporates joint produc-

[4]Separability assumptions have been used to simplify the data requirements for estimating production models. For example, Khmelnitskaya et al. (forthcoming) demonstrate how demand-side data alone (i.e., output prices, quantities, and exogenous demand shifters) can be used to estimate parameters related to economies of scale and scope in a cost model with input-output separability. Their approach relies on first-order conditions for output choices and the property that, under separability, the marginal rate of transformation is independent of input prices.

tion, input–output non-separability, and product-specific productivities, but also incorporates input-biased technology (as discussed in the following subsection).

2.2 Input-biased Technical Change

For problems in many fields of economics, whether technical change is biased towards particular factors is important. For example, Acemoglu (2002a,b) argues that skill-biased technical change is a major source of growing wage inequality in developing countries in the twentieth century. A key goal of empirical work in the applied production literature is to disentangle the effects of biased technical change from those of input substitutability in explaining variation in input use and cost shares. One widely used method for incorporating biased technical change into empirical models is to include a time trend as an argument in a flexible production function, allowing input shares to evolve over time. This approach is well represented in the work of Dale Jorgenson and collaborators (e.g., Gollop et al., 1987), as well as in many of the multiproduct translog cost function studies discussed earlier.[5]

A complementary approach was introduced by Binswanger (1974), who modeled technical change using factor-augmentation coefficients, one for each input. These coefficients scale input prices in the firm's cost function, thereby altering the effective price the firm faces for each input. Higher augmentation coefficients reflect greater input-specific productivity, thus reducing the effective input price. He estimated a translog cost function with each input's price modeled as the product of an augmentation coefficient and time. In his application to post-WWII U.S. agriculture, Binswanger found that technical change was fertilizer- and machinery-using and labor-saving. He concluded that approximately two-thirds of the observed decline in labor's cost share was attributable to biased technical change, while only one-third was due to price substitution effects. Gollop and Roberts (1981) applied Binswanger's framework to panel data from electric utilities, using a translog cost function. Their findings indicated labor- and capital-using, but fuel-saving technical change, and they rejected the hypothesis of Hicks-neutral technical progress. These studies illustrate how allowing for input-specific augmentation provides a more comprehensive view of how technological progress influences input choices and cost structure.

Despite the advances in structural production estimation following Olley and Pakes (1996), the applied literature has been slow to incorporate input-biased technological differences

[5] A recent example is Diewert et al. (2025). The authors compare the estimation of multiproduct cost functions and gross output functions using aggregate U.S. data with six inputs and four outputs. They incorporate piecewise linear splines in time to allow for biased technical change. They conclude that, at least for aggregate time series data, it is preferable to estimate multiproduct cost functions.

across firms or over time. Instead, productivity is often specified as a Hicks-neutral shifter, implicitly assuming that technological variation does not affect input cost shares. This limits the ability of these models to capture the full range of technological heterogeneity observed in practice. A small but growing number of recent studies have begun to incorporate input-augmenting productivity into production models to explain firm-level differences in size, markups, and input usage. Doraszelski and Jaumandreu (2018) estimate a CES production function with two separate components: labor-augmenting productivity and Hicks-neutral productivity. They identify the labor-augmenting component from the cost-minimizing first-order conditions between labor and materials, and recover the Hicks-neutral component from the production function. In their application to Spanish manufacturing, they find that both productivity components contribute equally to output growth.

Several other studies adopt the CES production function to explore input-biased technological change. Raval (2019) shows that labor-augmenting productivity plays a key role in determining firm size and export participation. Zhang (2019) finds that non-neutral technological change accounts for roughly 50 percent of the decline in the labor share in China's steel industry. Rubens et al. (2026) propose a framework in which both labor-augmenting productivity and monopsony power influence relative labor demand. Estimating the CES production function jointly with a labor supply equation, they identify monopsony distortions and find that private firms in China's nonferrous metals industry tend to exhibit both higher labor-augmenting productivity and stronger monopsony power relative to their state-owned counterparts. Aw and Lee (2025) study Taiwan electronics firms and find that R&D investments have a large and positive effect on labor-augmenting productivity while an increase in the number of firm affiliates increases capital-augmenting productivity.

While the CES production framework and the use of input first-order conditions have enabled researchers to disentangle input substitution elasticities from biased technological differences, this approach imposes strong functional form restrictions. In particular, it assumes that substitution elasticities are identical across all input pairs and constant across firms and time. These assumptions can be quite restrictive when the goal is to distinguish the effects of biased technical change from input substitution driven by factor prices. Two recent papers generalize the CES specification and relax the assumptions of constant elasticity of substitution and homogeneous returns to scale while distinguishing labor-augmenting and Hicks-neutral productivity. Demirer (2025) develops a nonparametric estimation method that relies on the assumption of homothetic separability. Zhao et al. (2025) develop an estimation method that utilizes both the labor and material share equations (rather than their ratio) in order to identify a translog production function that does not impose homothetic separability.

In the following sections, we develop and estimate a cost function model that combine the insights from the multiproduct production and factor-biased technology literature.

3 Theoretical Model of Multiproduct Production

This section introduces a theoretical framework for modeling multiproduct plants. We begin by describing the plant's production technology using a transformation function and recasting it as a flexible variable cost function. A key element of our model is that we incorporate efficiency differences in inputs, which correspond to biased technical progress, and unobserved quality differences in outputs, which contribute to output revenue heterogeneity. We then derive the input share equations and output supply conditions from the plant's short-run cost minimization and profit maximization problems, respectively. These structural relationships serve as the foundation for the empirical estimation strategy presented in subsequent sections. Our theoretical framework shares many characteristics with the model by Diewert and Fox (2008) but our approach to estimation using microdata differs from their focus on characterizing sectoral-level time series data.

3.1 Production

We model plant j's technology in time period t with a transformation function that converts inputs of labor, materials, and capital stock into a vector of N outputs. The production relationship is represented by the transformation function:

$$T(\boldsymbol{Q}^*_{jt}, L^*_{jt}, M^*_{jt}, K^*_{jt}) = 0, \tag{1}$$

where $\boldsymbol{Q}^*_{jt} = (Q^*_{1jt}, \ldots, Q^*_{Njt})$ denotes the vector of outputs measured in quality-adjusted units. Specifically, each output Q^*_{njt} is defined as the product of its physical quantity and an output-specific quality term:

$$Q^*_{njt} = Q_{njt} \exp(\nu_{njt}),$$

where Q_{njt} is the observed physical quantity of product n produced by plant j at time t, and ν_{njt} captures the quality of that output.[6]

[6]The term ν_{njt} reflects differences in product quality across plants, products, and time that affect plants' input choices and costs. It is analogous to the product-specific productivity in Dhyne et al. (2024), Caselli et al. (2026), and Cairncross et al. (2025) but we will use output quality to refer to these terms. We expect that products with larger values of ν_{njt} will require more (or more expensive) inputs and will be more costly to produce.

Similarly, the inputs (labor, materials, and capital) are also expressed in efficiency units:

$$L^*_{jt} = L_{jt}\exp(\mu_{Ljt}), \quad M^*_{jt} = M_{jt}\exp(\mu_{Mjt}), \quad K^*_{jt} = K_{jt}\exp(\mu_{Kjt}),$$

where L_{jt}, M_{jt}, K_{jt} are the observed physical quantities of labor, materials, and capital, respectively, and μ_{mjt} ($m \in \{L, M, K\}$) represents the efficiency of input m for plant j in period t.[7]

We assume that these quality and efficiency terms are known to plants at the time of their input-output decisions but are unobservable to the researcher. They are not choice variables for the plant, but rather characteristics of the plant that affect the endogenous choice of inputs and outputs at each point in time. A key goal of our empirical strategy is to recover these unobserved plant-level latent factors and quantify their contribution to the cost structure. We use them to summarize patterns of factor-augmenting technical differences and quality heterogeneity across multiproduct producers.

3.2 Short-run Cost Minimization

We follow Binswanger (1974) to develop a cost function in which input prices are expressed in efficiency units. We assume that labor and materials are variable inputs, while capital stock is fixed in the short run.

Let W_{Ljt} and W_{Mjt} denote the prices (wage rate and materials price, respectively) per unit of physical input. Each unit of labor L_{jt} provides $\exp(\mu_{Ljt})$ efficiency units; each unit of materials M_{jt} provides $\exp(\mu_{Mjt})$ efficiency units. Thus, the prices per efficiency unit of labor and materials are, respectively,

$$W^*_{Ljt} = \frac{W_{Ljt}}{\exp(\mu_{Ljt})}, \qquad W^*_{Mjt} = \frac{W_{Mjt}}{\exp(\mu_{Mjt})}.$$

We refer to W^*_{Ljt} and W^*_{Mjt} as the shadow (efficiency-adjusted) prices of labor and materials, respectively.

Given the fixed capital stock K^*_{jt}, the quality-adjusted output vector $\boldsymbol{Q}^*_{jt}$, and the shadow

[7] By modeling inputs in efficiency units and outputs in quality-adjusted units, our framework addresses the concern raised by De Roux et al. (2024), who show that production estimation based on firm-level physical quantities of inputs and outputs (aggregated from the firm–product level) may suffer from quality and variety biases when firms differ in the unobservable quality and mix of their inputs and outputs. In our framework, this concern is accommodated by explicitly estimating a multiproduct production structure and allowing the quality terms, ν_{njt} for each output, and the efficiency terms, μ_{mjt} for each input, to absorb the quality and variety effects.

prices of inputs, the short-run variable cost function is defined as:

$$C_{jt}(W^*_{Ljt}, W^*_{Mjt}, K^*_{jt}, \boldsymbol{Q}^*_{jt}) \equiv \min_{L^*_{jt}, M^*_{jt}} \left\{ W^*_{Ljt} L^*_{jt} + W^*_{Mjt} M^*_{jt} \right\}$$
$$\text{s.t.} \quad T(\boldsymbol{Q}^*_{jt}, L^*_{jt}, M^*_{jt}, K^*_{jt}) = 0. \tag{2}$$

This formulation represents a multiproduct shadow price variable cost function, where the technology is expressed as a cost function in terms of efficiency-adjusted inputs, quality-adjusted output levels, and fixed capital.

By construction, the total expenditures in efficiency units are equal to those measured in physical units:

$$E_{Ljt} \equiv L_{jt} W_{Ljt} = L^*_{jt} W^*_{Ljt}, \quad E_{Mjt} \equiv M_{jt} W_{Mjt} = M^*_{jt} W^*_{Mjt},$$

and therefore, the observed total expenditure on variable inputs is:

$$C_{jt} \equiv E_{Ljt} + E_{Mjt} = L_{jt} W_{Ljt} + M_{jt} W_{Mjt} = L^*_{jt} W^*_{Ljt} + M^*_{jt} W^*_{Mjt}.$$

We assume that the cost function takes a translog functional form, which is flexible and allows for heterogeneous substitution elasticities and returns to scale across plants. In logarithmic form, the cost function is specified as:

$$\begin{aligned} c_{jt} = \alpha_0 &+ \sum_{m \in \{L,M\}} \alpha_m w^*_{mjt} + \alpha_K k^*_{jt} + \sum_{n=1}^{N} \beta_n q^*_{njt} \\ &+ \frac{1}{2} \sum_{m \in \{L,M\}} \sum_{i \in \{L,M\}} \delta_{mi} w^*_{mjt} w^*_{ijt} + \sum_{m \in \{L,M\}} \delta_{Km} w^*_{mjt} k^*_{jt} + \frac{1}{2} \delta_{KK} (k^*_{jt})^2 \\ &+ \sum_{m \in \{L,M\}} \sum_{n=1}^{N} \phi_{mn} w^*_{mjt} q^*_{njt} + \sum_{n=1}^{N} \phi_{Kn} k^*_{jt} q^*_{njt} + \frac{1}{2} \sum_{n=1}^{N} \sum_{s=1}^{N} \gamma_{ns} q^*_{njt} q^*_{sjt}, \end{aligned} \tag{3}$$

where lowercase variables denote logarithms of the variables, and the elements of the vector $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$ are the parameters of the variable cost function.

Applying Shephard's Lemma, we obtain the input cost share equations for labor and mate-

rials as:

$$\frac{E_{Ljt}}{C_{jt}} = \frac{L^*_{jt}W^*_{Ljt}}{C_{jt}} = \frac{\partial c_{jt}}{\partial w^*_{Ljt}} = \alpha_L + \sum_{m\in\{L,M\}} \delta_{Lm} w^*_{mjt} + \delta_{LK} k^*_{jt} + \sum_{n=1}^{N} \phi_{Ln} q^*_{njt}, \tag{4}$$

$$\frac{E_{Mjt}}{C_{jt}} = \frac{M^*_{jt}W^*_{Mjt}}{C_{jt}} = \frac{\partial c_{jt}}{\partial w^*_{Mjt}} = \alpha_M + \sum_{m\in\{L,M\}} \delta_{Mm} w^*_{mjt} + \delta_{MK} k^*_{jt} + \sum_{n=1}^{N} \phi_{Mn} q^*_{njt}. \tag{5}$$

These input share equations contain the subset of cost function parameters related to labor and material inputs and their interactions with capital and outputs.

3.3 Short-run Profit Maximization

We assume that plants behave as short-run profit maximizers in choosing their output levels. Let the price of product n in physical units be denoted P_{njt}. Since product quality affects the effective output delivered per physical unit, we define the quality-adjusted price as:

$$P^*_{njt} = \frac{P_{njt}}{\exp(\nu_{njt})}, \quad \text{for } n = 1, 2, \ldots, N.$$

For each product there is a demand function:

$$Q^*_{njt} = D_n(\mathbf{P}^*_{jt}, \mathbf{Y}_{jt}), \tag{6}$$

where $\mathbf{P}^*_{jt}$ is a vector of quality-adjusted prices for all relevant products in the market, including those produced by other plants, and $\mathbf{Y}_{jt}$ denotes a vector of demand shifters such as product characteristics, market size, consumer income, and competitive structure. The demand for product n produced by plant j may depend on its own price P^*_{njt}, the prices of other products produced by the same plant, and the prices of competing products offered by rival plants.

The plant chooses output quantities $\boldsymbol{Q}^*_{jt}$ to maximize short-run profits, taking input shadow prices and capital as given:

$$\max_{\boldsymbol{Q}^*_{jt}} \sum_{n=1}^{N} P^*_{njt} Q^*_{njt} - C_{jt}(W^*_{Ljt}, W^*_{Mjt}, K^*_{jt}, \boldsymbol{Q}^*_{jt}) \tag{7}$$

subject to: inverse demand $P^*_{njt} = P_n(\mathbf{Q}^*_{jt}, \mathbf{Y}_{jt})$ implied by (6), for all n.

Taking the first-order condition with respect to each output Q^*_{njt}, and using the translog cost function (3), we obtain, for $\forall n \in \{1, \ldots, N\}$:

$$\frac{1}{\theta_{njt}} \frac{R_{njt}}{C_{jt}} = \frac{\partial c_{jt}}{\partial q^*_{njt}} = \beta_n + \sum_{m \in \{L,M\}} \phi_{mn} w^*_{mjt} + \phi_{Kn} k^*_{jt} + \sum_{s=1}^{N} \gamma_{ns} q^*_{sjt}, \tag{8}$$

where $R_{njt} \equiv P_{njt} Q_{njt}$ is the revenue from product n, and $\theta_{njt} \equiv 1/\left[1 - \sum_{i=1}^{N} \eta^{inv}_{injt} \frac{R_{ijt}}{R_{njt}}\right]$ is the markup for product n under multiproduct pricing, where $\eta^{inv}_{injt} = -\frac{\partial P^*_{njt}}{\partial Q^*_{ijt}} \frac{Q^*_{ijt}}{P^*_{njt}}$ is the elasticity of inverse demand for price of product n with respect to the quantity of product i.[8]

This condition equates the plant's markup-adjusted revenue-cost ratio to the cost elasticity of producing output n. These revenue-cost equations contain the subset of cost function parameters related to the outputs and their interactions with capital and input prices. This equation highlights how the product revenue–cost ratio, $\frac{R_{njt}}{C_{jt}}$, is determined by heterogeneity in product-specific markups (θ_{njt}), product-specific quality (ν_{njt}), and factor-augmenting efficiency (μ_{mjt}) within a framework that allows for joint production and non-separability between inputs and outputs.

Equations (3), (4), (5), and (8) characterize the general multiproduct production model.[9] Imposing additional restrictions on the technology or efficiency terms yields several special cases of interest:

a. Hicks neutral technical change: $\mu_{mjt} = \mu_{jt}$, for $m \in \{L, M, K\}$.

b. No factor-augmenting technical change: $\mu_{mjt} = 0$, for $m \in \{L, M, K\}$.

c. No product-specific quality: $\nu_{njt} = \nu_{jt}$, for $n = 1, 2, \ldots, N$.

d. No plant-level output quality: $\nu_{njt} = 0$, for $n = 1, 2, \ldots, N$.

e. Input-output separability: $\phi_{mn} = 0$, for $m \in \{L, M, K\}$ and $n = 1, 2, \ldots, N$.

f. Non-joint production: $\gamma_{ns} = 0$, for $s \neq n$.

[8] To see this, $\theta_{njt} = \frac{P_{njt}}{mc(Q_{njt})} = \frac{P_{njt}}{\sum_{i \in N} \frac{\partial P_{ijt}(\boldsymbol{Q}_{jt}, \boldsymbol{Q}_{-jt}; \boldsymbol{Y}_{jt})}{\partial Q_{njt}} Q_{ijt} + P_{njt}} = \frac{1}{\sum_{i \in N} \frac{\partial P_{ijt}(\boldsymbol{Q}_{jt}, \boldsymbol{Q}_{-jt}; \boldsymbol{Y}_{jt})}{\partial Q_{njt}} \frac{Q_{ijt}}{P_{njt}} + 1} = \frac{1}{\sum_{i \in N} \frac{\partial P_{ijt}(\boldsymbol{Q}_{jt}, \boldsymbol{Q}_{-jt}; \boldsymbol{Y}_{jt})}{\partial Q_{njt}} \frac{Q_{njt}}{P_{ijt}} \frac{P_{ijt} Q_{ijt}}{P_{njt} Q_{njt}} + 1} = \frac{1}{1 - \sum_{i \in N} \eta^{inv}_{injt} \frac{R_{ijt}}{R_{njt}}}$, where the second equality comes from the definition of marginal cost and the last equality comes from the definition of revenue and inverse demand elasticity.

[9] If there are no fixed factors then the capital parameters $\alpha_K = \delta_{KK} = \delta_{Km} = \phi_{Kn} = 0$. In this case, C_{jt} becomes total cost rather than variable cost.

A combination of conditions (b) and (e) is examined by Cairncross et al. (2025). Under these conditions, in our model, the relative markup between two products, k and n, can be expressed as

$$\frac{\theta_{kjt}}{\theta_{njt}} = \frac{\beta_n + \sum_{s=1}^{N} \gamma_{ns}(q_{sjt} + \nu_{sjt})}{\beta_k + \sum_{s=1}^{N} \gamma_{ks}(q_{sjt} + \nu_{sjt})} \frac{R_{kjt}}{R_{njt}}. \tag{9}$$

This expression shows that the relative markup between the two products can be inferred from their revenue ratio, together with the quantities and qualities of all products, given the cost function parameters. This relationship is similar to Cairncross et al. (2025, eq. 16).[10] They use it to demonstrate that a model using production data alone is insufficient to identify both relative product markups and product productivities. They emphasize that additional assumptions on demand and firm conduct are required to identify heterogeneity in both dimensions.

Our empirical framework, developed in Section 4, incorporates a specification of output demand, input demand, plants' first-order conditions for profit maximization, and a flexible representation of production technology to identify all cost function parameters, product quality, and input-augmenting efficiencies.

3.4 Theoretical Constraints on the Cost Function

In theory, the cost function parameters are subject to symmetry, linear homogeneity, and concavity constraints.[11] In our translog specification, symmetry implies $\delta_{mi} = \delta_{im}$ for all variable input price pairs and $\gamma_{ns} = \gamma_{sn}$ for all output pairs. Linear homogeneity in input prices implies three sets of restrictions: (i) $\alpha_L + \alpha_M = 1$, (ii) $\sum_{m\in\{L,M\}} \delta_{mi} = 0$ for all $i \in \{L, M, K\}$, and (iii) $\sum_{m\in\{L,M\}} \phi_{mn} = 0$ for all $n = 1, \ldots, N$. Imposing these theoretical constraints is equivalent to normalizing both the variable cost and one of the input prices by the other input price. In our application, we will normalize by the material price. Specifically, we define the normalized variable cost and wage rate as $\tilde{c}_{jt} \equiv c_{jt} - w^*_{Mjt}$ and $\tilde{w}^*_{Ljt} \equiv w^*_{Ljt} - w^*_{Mjt}$, where all variables are in logarithmic form. As a result of the linear homogeneity restriction, the material share equation becomes redundant and drops out of the estimating system.

[10] It is actually more general than the specific separable parameterization of Cairncross et al. (2025), as it depends on the entire vector of output quantities and productivities. In their formulation, the ratio depends only on products n and k. To obtain their specification from our translog cost function, one must impose the assumption of non-joint production (condition (f) discussed above).

[11] The cost function must be concave in variable input prices, meaning that the Hessian matrix of the cost function with respect to the input prices (w^*_{Ljt}, w^*_{Mjt}) should be negative semi-definite. We do not impose this concavity condition during estimation but assess it post-estimation to ensure empirical validity.

The normalized translog variable cost function is given by:

$$\begin{aligned}\tilde{c}_{jt} = \alpha_0 + \alpha_L \tilde{w}^*_{Ljt} + \alpha_K k^*_{jt} + \sum_{n=1}^{N} \beta_n q^*_{njt} \\ + \frac{1}{2}\delta_{LL}(\tilde{w}^*_{Ljt})^2 + \delta_{KL}\tilde{w}^*_{Ljt}k^*_{jt} + \frac{1}{2}\delta_{KK}(k^*_{jt})^2 \\ + \sum_{n=1}^{N} \phi_{Ln}\tilde{w}^*_{Ljt}q^*_{njt} + \sum_{n=1}^{N} \phi_{Kn}k^*_{jt}q^*_{njt} + \frac{1}{2}\sum_{n=1}^{N}\sum_{s=1}^{N} \gamma_{ns}q^*_{njt}q^*_{sjt}. \end{aligned} \tag{10}$$

Under the same normalization, the labor cost share equation (4) simplifies to:

$$\frac{E_{Ljt}}{C_{jt}} = \alpha_L + \delta_{LL}\tilde{w}^*_{Ljt} + \delta_{LK}k^*_{jt} + \sum_{n=1}^{N} \phi_{Ln}q^*_{njt}. \tag{11}$$

Similarly, the markup-adjusted revenue–cost ratio (8) becomes:

$$\frac{1}{\theta_{njt}}\frac{R_{njt}}{C_{jt}} = \beta_n + \phi_{Ln}\tilde{w}^*_{Ljt} + \phi_{Kn}k^*_{jt} + \sum_{s=1}^{N} \gamma_{ns}q^*_{sjt}, \quad \forall n \in \{1, \ldots, N\}. \tag{12}$$

Together, equations (6), (10), (11), (12) form the empirical core of our structural model of production and demand. The model expresses factor prices and capital in efficiency units and output levels and prices in quality-adjusted units. The key unobserved components, input efficiencies μ_{mjt} and output qualities ν_{njt}, are not directly observed but are embedded in the plant's behavior through their impact on shadow prices and quality-adjusted output units. Recovering these latent variables is central to our empirical objective.

This framework offers several methodological advantages. First, it does not impose separability between inputs and outputs, so the estimated marginal costs and, therefore, marginal rates of transformation (MRT) will depend on input prices and capital stock in efficiency units and output levels in quality-adjusted units. Second, it allows joint production, so that the marginal cost of one product is affected by the output level of other products. It also allows potential economies of scope. Third, it does not require allocating inputs to individual product lines, a task often infeasible in standard micro datasets. Fourth, the model allows variable markups across products and plants, with plant-specific markup adjustments entering directly into the output supply conditions. Finally, the flexibility of the translog functional form permits substitution elasticities and returns to scale to vary across plants and over time, depending on input prices, capital stock, and output composition. All these

advantages are potentially important for rationalizing the observed data on input and output choices when the technology is characterized by biased technical change, product quality differences, and scale and scope economies.

4 Empirical Model and Estimation Method

Estimation of the model requires data on the prices, quantities, and revenues of all products, input prices and expenditures for variable inputs, and quantities of fixed factors. It also requires a specification of product demand functions, which in turn requires data on exogenous demand characteristics or demand shifters. Our estimation algorithm contains two parts. First, we estimate the demand functions (6) to obtain the product markups θ_{njt}. Second, given the markup estimates, we estimate the production system, (10), (11), and (12) for cost function parameters and unobserved input efficiencies μ_{mjt} and output qualities ν_{njt}.[12]

While we have tightly specified the production side of the model, researchers have flexibility in the specification of the product demand functions. Depending on the data availability and industry context, a flexible form of demand function can be adopted to allow for variable product markups arising from demand interdependence across products and plants. For example, if product characteristics are available, Logit-based demand models can be used. Alternatively, if only market-level demand shifters are available and plants are relatively small, then more restrictive specifications such as constant-elasticity of substitution (CES) demand may be plausible.[13]

[12]Popular production-function estimators, such as Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2015), typically include a first stage that purges the dependent variable, output, which is often inferred from deflated revenue, with measurement error arising from imperfectly measured output-price deflators. By contrast, the dependent variables in our system are revenues and expenditures, which are reported directly in production surveys and are therefore less likely to suffer from this source of measurement error. We therefore do not implement an explicit correction for measurement error in the dependent variables. Measurement error may nevertheless affect some explanatory variables, in particular materials prices and physical output quantities. In our framework, such errors are at least partly absorbed by the latent input-efficiency and output-quality terms, μ_{jt} and ν_{jt}, respectively.

[13]The choice of demand function specification determines the degree of heterogeneity allowed in markups, which in turn is likely to affect the input efficiencies μ_{mjt} and output qualities ν_{njt} recovered from the cost model. Moreover, the demand model and the cost model (i.e., the input cost shares, output revenue–cost ratios, and cost function) can, in principle, be estimated jointly, imposing cross-equation restrictions on the demand elasticity parameters. In this paper, however, our primary objective is to develop a methodology for recovering input efficiencies and output qualities for multiproduct plants via a production cost model. We therefore adopt a relatively simple demand specification and estimation strategy, in order to concentrate on the cost function estimation methodology.

4.1 Empirical Demand Model

In the empirical implementation, we adopt a CES demand specification in quality-adjusted terms. Specifically, for each product n, we assume the following log-linear demand function:

$$\ln(Q^*_{njt}) = -\eta_n \ln(P^*_{njt}) + \lambda_n Y_{njt} + \omega^*_{njt}, \quad \forall n \in \{1, \ldots, N\},$$

where Q^*_{njt} and P^*_{njt} are the quality-adjusted quantity and price of product n of plant j in period t, $\eta_n > 0$ is the price elasticity of demand, Y_{njt} is a vector of observed demand shifters (e.g., market size), and ω^*_{njt} is an unobserved demand shock.

Substituting in the definitions of quality-adjusted variables, we obtain a demand equation in observable terms:

$$\ln(Q_{njt}) = -\eta_n \ln(P_{njt}) + \lambda_n Y_{njt} + \omega_{njt}, \quad \forall n \in \{1, \ldots, N\}, \tag{13}$$

where the composite unobservable $\omega_{njt} = (\eta_n - 1)\nu_{njt} + \omega^*_{njt}$ contains both the output quality term ν_{njt}, which enters via the transformation from quality-adjusted output, and the original demand shock ω^*_{njt}. This formulation highlights the challenge of estimating demand elasticities in the presence of unobserved product quality, which is correlated with prices and may bias OLS estimates.

To address the potential endogeneity of prices in equation (13), we use an instrumental variable estimator. Valid instruments must be correlated with the observed product price P_{njt} but uncorrelated with the composite demand shock ω_{njt}. We use lagged plant-level cost shifters. We provide more details on the demand variables and instrumental variables in Section 6.1.

Once the demand elasticities η_n are estimated, we can compute the inverse markup term that enters the first-order condition for profit maximization in equation (8). Under the assumption of constant-elasticity demand and no cross-price effects within the plant, the inverse markup term simplifies to:

$$\frac{1}{\theta_{njt}} = \left[1 - \sum_{i=1}^{N} \frac{1}{\eta_{injt}} \frac{R_{ijt}}{R_{njt}}\right] = \frac{\eta_n - 1}{\eta_n}, \quad \forall n \in \{1, \ldots, N\}. \tag{14}$$

4.2 Empirical Cost Model

The key idea behind estimation of the cost system is to divide it into observable variables and unobserved plant-level quality and efficiency terms, which enter the system as structural errors.[14]

The labor share (i.e., labor cost to variable cost ratio) equation (11) is rewritten as:

$$\begin{aligned}\frac{E_{Ljt}}{C_{jt}} &= \alpha_L + \delta_{LL}\tilde{w}_{Ljt} + \delta_{LK}k_{jt} + \sum_{n=1}^{N}\phi_{Ln}q_{njt} \\ &+ \underbrace{\left(-\delta_{LL}\tilde{\mu}_{Ljt} + \delta_{LK}\mu_{Kjt} + \sum_{n=1}^{N}\phi_{Ln}\nu_{njt}\right)}_{\text{Composite error: } \epsilon_{Ljt}},\end{aligned} \tag{15}$$

where $\tilde{w}_{Ljt} = w_{Ljt} - w_{Mjt}$ and $\tilde{\mu}_{Ljt} = \mu_{Ljt} - \mu_{Mjt}$. The first line contains the observable variables. The second line defines a composite error term for the labor share equation (ϵ_{Ljt}). It is a linear combination of the unobserved quality terms for outputs (ν_{njt}) and efficiency terms for labor ($\tilde{\mu}_{Ljt}$) and capital (μ_{Kjt}), weighted by second-order cost function parameters associated with labor. This provides a structural interpretation of the error term in the labor share equation.

Similarly, the output supply equations (12) become:

$$\begin{aligned}\frac{\eta_n - 1}{\eta_n}\frac{R_{njt}}{C_{jt}} &= \beta_n + \phi_{Ln}\tilde{w}_{Ljt} + \phi_{Kn}k_{jt} + \sum_{s=1}^{N}\gamma_{ns}q_{sjt} \\ &+ \underbrace{\left\{-\phi_{Ln}\tilde{\mu}_{Ljt} + \phi_{Kn}\mu_{Kjt} + \sum_{s=1}^{N}\gamma_{ns}\nu_{sjt}\right\}}_{\text{Composite error: } \epsilon_{njt}}, \quad \forall n \in \{1,\ldots,N\},\end{aligned} \tag{16}$$

where the composite error term is also a linear combination of unobserved quality and efficiency terms, with weights determined by second-order cost function parameters associated with output n.

[14] The goal of Binswanger (1974) and Diewert and Fox (2008) is to quantify the input efficiency terms in sectoral-level time series data. Therefore, they specify the efficiency terms as parametric functions of time. This leads to including a time trend as an explanatory variable in the input cost share equations. In contrast, our analysis emphasizes efficiency differences across plants as well as over time, and we do not impose a parametric time trend on the efficiency terms.

Finally, the variable cost, normalized by the material price, can be rewritten as:

$$c_{jt} - w_{Mjt} = \tilde{c}_{jt} - \mu_{Mjt}, \tag{17}$$

where the left-hand side, $c_{jt} - w_{Mjt}$, is observable, while $\tilde{c}_{jt}$ is defined by (10), incorporating both observable variables and unobserved quality and efficiency terms. The term μ_{Mjt} represents the efficiency term for material input.

The full empirical model incorporates $N+3$ sources of heterogeneity on the production side, including one for each of the N outputs, two variable inputs, and one fixed input. However, there are only $N+2$ equations, N output supply equations, one variable input share equation, and the cost function, available to disentangle the sources of heterogeneity.

The quality terms associated with any number of outputs can be identified from the corresponding set of output supply equations. The model can accommodate additional variable inputs, each contributing an efficiency variable and an input share equation to the system. The efficiency of the normalizing input, such as materials, can be separately identified using the cost function. The main challenge in fully identifying the parameters arises because the efficiency term associated with the fixed input (e.g., capital) cannot be identified without additional information. If no fixed input is present, as in the application by Binswanger (1974), then all input efficiency terms can be recovered from the input share equations and the cost function. Alternatively, if a valid proxy for the efficiency of the fixed input is available, it may be possible to identify this term directly. In the absence of such data or instruments, we impose a normalization by setting the efficiency of the fixed input to zero, i.e., $\mu_{Kjt} = 0$. If, in reality, $\mu_{Kjt} \neq 0$, the recovered quality and efficiency terms for other outputs and inputs reflect this deviation and become linear functions of μ_{Kjt}, as discussed below.

For the remaining efficiency and quality terms, we assume that each evolves independently following a first-order Markov process:

$$\nu_{njt} = g_n \nu_{njt-1} + \xi_{njt}, \quad n = 1, \ldots, N, \tag{18}$$

$$\tilde{\mu}_{Ljt} = g_L \tilde{\mu}_{Ljt-1} + \xi_{Ljt}, \tag{19}$$

$$\mu_{Mjt} = g_M \mu_{Mjt-1} + \xi_{Mjt}, \tag{20}$$

where ξ_{njt}, ξ_{Ljt}, and ξ_{Mjt} are i.i.d. innovation shocks and g_n, g_L, and g_M are the persistence parameters associated with the first-order Markov processes.

We estimate an empirical cost model for plants with two outputs (i.e., $N = 2$) in the following

two steps.

Step 1: estimate the equation system of input cost shares and output supply functions.

We express (15) and (16) in matrix notation:

$$\begin{bmatrix} \frac{\eta_1-1}{\eta_1} R_{1jt}/C_{jt} \\ \frac{\eta_2-1}{\eta_2} R_{2jt}/C_{jt} \\ E_{Ljt}/C_{jt} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \alpha_L \end{bmatrix} + \Omega \begin{bmatrix} q_{1jt} \\ q_{2jt} \\ -\tilde{w}_{Ljt} \end{bmatrix} + \begin{bmatrix} \phi_{K1} \\ \phi_{K2} \\ \delta_{KL} \end{bmatrix} k_{jt} + \begin{bmatrix} \epsilon_{1jt} \\ \epsilon_{2jt} \\ \epsilon_{Ljt} \end{bmatrix}, \tag{21}$$

where[15]

$$\Omega \equiv \begin{bmatrix} \gamma_{11} & \gamma_{12} & -\phi_{L1} \\ \gamma_{12} & \gamma_{22} & -\phi_{L2} \\ \phi_{L1} & \phi_{L2} & -\delta_{LL} \end{bmatrix}. \tag{22}$$

As shown above, the composite errors in the input share and output supply equations are linear functions of quality and efficiency terms, which are the structural errors in the model, with weights given by the second-order parameters of the cost function:

$$\begin{bmatrix} \epsilon_{1jt} \\ \epsilon_{2jt} \\ \epsilon_{Ljt} \end{bmatrix} = \Omega \begin{bmatrix} \nu_{1jt} \\ \nu_{2jt} \\ \tilde{\mu}_{Ljt} \end{bmatrix}. \tag{23}$$

Substitute (23) into (21) to solve for the structural errors in terms of observable data and cost function parameters:

$$\begin{bmatrix} \nu_{1jt} \\ \nu_{2jt} \\ \tilde{\mu}_{Ljt} \end{bmatrix} = \Omega^{-1} \left\{ \begin{bmatrix} \frac{\eta_1-1}{\eta_1} R_{1jt}/C_{jt} \\ \frac{\eta_2-1}{\eta_2} R_{2jt}/C_{jt} \\ E_{Ljt}/C_{jt} \end{bmatrix} - \begin{bmatrix} \beta_1 \\ \beta_2 \\ \alpha_L \end{bmatrix} - \begin{bmatrix} \phi_{K1} \\ \phi_{K2} \\ \delta_{KL} \end{bmatrix} k_{jt} \right\} - \begin{bmatrix} q_{1jt} \\ q_{2jt} \\ -\tilde{w}_{Ljt} \end{bmatrix}. \tag{24}$$

[15] We assume that $\det(\Omega) \neq 0$, so Ω is invertible. In our empirical implementation, $\det(\hat{\Omega}) = 0.037$.

As a result, we can write innovation terms in (18) and (19) as:

$$\xi_{1jt} = \nu_{1jt} - g_1 \nu_{1jt-1}, \tag{25}$$

$$\xi_{2jt} = \nu_{2jt} - g_2 \nu_{2jt-1}, \tag{26}$$

$$\xi_{Ljt} = \tilde{\mu}_{Ljt} - g_L \tilde{\mu}_{Ljt-1}, \tag{27}$$

where the efficiency and quality terms are as specified in (24).

We estimate the model parameters in (21), as well as g_1, g_2, and g_L, using GMM based on the moment conditions:

$$E[Z_{1jt}\xi_{1jt}] = 0, \quad E[Z_{2jt}\xi_{2jt}] = 0, \quad E[Z_{Ljt}\xi_{Ljt}] = 0,$$

where Z represents a set of instrumental variables uncorrelated with the innovation shocks to the evolution of the quality and efficiency terms.

The base set of instruments is defined as $Z^0_{jt} = (1, \frac{R_{1jt-1}}{C_{jt-1}}, \frac{R_{2jt-1}}{C_{jt-1}}, \frac{E_{Ljt-1}}{C_{jt-1}}, k_{jt-1})$. The instrument sets are extended to control for the exogenous aspects of the output markets. The plants' outputs and revenues may be impacted by rival plants' costs due to competition. Specifically, Z_{1jt} includes Z^0_{jt}, q_{1jt-1}, and lagged average rival producers' labor productivity, lagged wage rate, and lagged capital stock. Z_{2jt} includes Z^0_{jt}, q_{2jt-1}, and lagged average rival producers' labor productivity, lagged wage rate, and lagged capital stock. Z_{Ljt} includes Z^0_{jt}, w_{Ljt-1}, and lagged average rival producers' wage rate and lagged capital stock.

This step yields the quality terms ν_{1jt} and ν_{2jt}, the relative labor efficiency term $\tilde{\mu}_{Ljt}$, cost function parameters in (21), as well as persistence parameters g_1, g_2, and g_L.[16] These estimates do not include the intercept of the cost function (α_0), the linear and quadratic terms associated with the fixed input (α_K and δ_{KK}), and the efficiency level of the normalizing

[16] If $\mu_{Kjt} \neq 0$, (24) becomes:

$$\begin{bmatrix} \nu_{1jt} \\ \nu_{2jt} \\ \tilde{\mu}_{Ljt} \end{bmatrix} + \Omega^{-1} \begin{bmatrix} \phi_{K1} \\ \phi_{K2} \\ \delta_{KL} \end{bmatrix} \mu_{Kjt} = \Omega^{-1} \left\{ \begin{bmatrix} \frac{\eta_1 - 1}{\eta_1} R_{1jt}/C_{jt} \\ \frac{\eta_2 - 1}{\eta_2} R_{2jt}/C_{jt} \\ E_{Ljt}/C_{jt} \end{bmatrix} - \begin{bmatrix} \beta_1 \\ \beta_2 \\ \alpha_L \end{bmatrix} - \begin{bmatrix} \phi_{K1} \\ \phi_{K2} \\ \delta_{KL} \end{bmatrix} k_{jt} \right\} - \begin{bmatrix} q_{1jt} \\ q_{2jt} \\ -\tilde{w}_{Ljt} \end{bmatrix}.$$

Defining $\begin{bmatrix} \tilde{\phi}_{K1} \\ \tilde{\phi}_{K2} \\ \tilde{\phi}_{KL} \end{bmatrix} \equiv \Omega^{-1} \begin{bmatrix} \phi_{K1} \\ \phi_{K2} \\ \delta_{KL} \end{bmatrix}$, which is a vector of constant parameters, this relationship implies that the quality and efficiency terms we recover contain a linear component of the capital efficiency: $\begin{bmatrix} \nu_{1jt} + \tilde{\phi}_{K1}\mu_{Kjt} \\ \nu_{2jt} + \tilde{\phi}_{K2}\mu_{Kjt} \\ \tilde{\mu}_{Ljt} + \tilde{\phi}_{KL}\mu_{Kjt} \end{bmatrix}$.

input price (μ_{Mjt}). These remaining parameters are estimated in the next step using the cost function.

Step 2: use the cost function to estimate α_0, α_K, and δ_{KK} and recover μ_{Mjt}.

Given the estimates from Step 1, we can predict a portion of the observed cost as

$$\begin{aligned}\hat{c}_{jt} &= \alpha_L \tilde{w}^*_{Ljt} + \sum_{n=1}^{2} \beta_n q^*_{njt} + \frac{1}{2}\delta_{LL}(\tilde{w}^*_{Ljt})^2 + \delta_{KL}\tilde{w}^*_{Ljt}k_{jt} \\ &\quad + \sum_{n=1}^{2} \phi_{Ln}\tilde{w}^*_{Ljt}q^*_{njt} + \sum_{n=1}^{2} \phi_{Kn}k_{jt}q^*_{njt} + \frac{1}{2}\sum_{n=1}^{2}\sum_{s=1}^{2} \gamma_{ns}q^*_{njt}q^*_{sjt}. \end{aligned} \tag{28}$$

Therefore, the remaining portion of the observed cost, according to (10), is

$$c_{jt} - w_{Mjt} - \hat{c}_{jt} = \alpha_0 + \alpha_K k_{jt} + \frac{1}{2}\delta_{KK}k^2_{jt} - \mu_{Mjt}, \tag{29}$$

where the materials efficiency term, μ_{Mjt}, is the structural error in (29) and the only unknown parameters are $\alpha_0, \alpha_K, \delta_{KK}$.

Rewriting this equation in terms of the structural error μ_{Mjt} gives:

$$\mu_{Mjt} = \alpha_0 + \alpha_K k_{jt} + \frac{1}{2}\delta_{KK}k^2_{jt} - (c_{jt} - w_{Mjt} - \hat{c}_{jt}). \tag{30}$$

Using the Markov assumption of μ_{Mjt} in (20), the innovation term is:

$$\xi_{Mjt} = \mu_{Mjt} - g_M \mu_{Mjt-1}. \tag{31}$$

Thus, we can estimate the parameters $(\alpha_0, \alpha_K, \delta_{KK}, g_M)$ via GMM using moment conditions:

$$E[Z_{Mjt}\xi_{Mjt}] = 0, \tag{32}$$

where $Z_{Mjt} = (1, c_{jt-1} - w_{Mjt-1} - \hat{c}_{jt-1}, k_{jt-1}, k^2_{jt-1}, k_{jt}, k^2_{jt})$ is a set of instrumental variables. These lagged variables are orthogonal to ξ_{Mjt} because ξ_{Mjt} is not in the plant information set in period $t-1$.

Once the parameters are estimated, we can recover the quality terms (ν_{1jt}, ν_{2jt}) and the relative labor efficiency term $\tilde{\mu}_{Ljt}$ from (24), recover μ_{Mjt} from (30), and compute $\mu_{Ljt} = \tilde{\mu}_{Ljt} + \mu_{Mjt}$.

Our estimation approach is based on constructing moment conditions that exploit the orthogonality between the innovation shocks to the structural errors $(\xi_{1jt}, \xi_{2jt}, \xi_{Ljt}, \xi_{Mjt})$ and instrumental variables that are lagged levels of relevant variables. An alternative is to use a dynamic panel estimator (Blundell and Bond, 2000). In this approach (usually referred to as system GMM), the moment conditions include: (i) difference moments that set first-differences of the auto-regressive-filtered (AR-filtered) residuals of the structural errors to be orthogonal to lagged levels of the internal instruments, and (ii) level moments that set the AR-filtered residuals of the structural errors in levels to be orthogonal to lagged first differences of the instruments. We describe how this dynamic panel estimator can be implemented to estimate our model in Online Appendix A. Through a Monte Carlo study reported in Online Appendix B, we show that the dynamic panel approach performs comparably well in large panels, but is less precise than our implementation in small panels (i.e., the sample in Section 5).

5 Data

The data used in this paper come from the Taiwanese Survey and Census of Manufacturing Operations (TSCMO), conducted by the Ministry of Economic Affairs (MOEA) during the period 1992–2004.[17] The TSCMO collects detailed information from manufacturing establishments, including sales revenue, output quantities, number of employees, capital stock, and expenditures on labor and materials. Additionally, the survey provides product-specific data at the 7-digit Standard Industry Classification (SIC) level, including sales revenue and quantity sold for each product produced by a plant. Using these data, we construct plant-level prices for each product by dividing sales by physical quantity.

Although we observe plant-level output prices and wage rates, we do not observe plant-specific prices for material inputs. Instead, we proxy material input prices w_{Mjt} with industry-level material price indices w_{Mt}. De Loecker et al. (2016) and Grieco et al. (2016) highlight the importance of accounting for heterogeneity in material prices, which, if ignored, can lead to bias in production function estimation. In line with their approach of explicitly allowing for input price heterogeneity, we let any deviation between the industry material price index and a plant's actual (but unobserved) material price be absorbed by the efficiency term μ_{Mjt}. In effect, the model treats plant-level variation in material prices as a source of serially-correlated structural error (i.e., heterogeneity) in the normalized cost function (29). The estimates of μ_{Mjt} recovered in the next section therefore reflect all sources of plant-level material price heterogeneity. If plant-level material prices were observed, they would enter

[17] The survey was not conducted in 1996 and 2001, as these were Census years.

Table 1: Number of plants, by product mix

Year	Cotton fiber only	Man-made fiber only	Both	Total
1992	163	308	38	509
1993	159	280	32	471
1994	55	171	24	250
1995	63	147	26	236
1997	133	278	47	458
1998	149	273	46	468
1999	183	270	39	492
2000	154	319	38	511
2002	114	273	38	425
2003	78	284	35	397
2004	132	359	39	530
Total	1,383	2,962	402	4,747

directly as w_{Mjt} in the estimating equations, and the estimates of μ_{Mjt} would then capture only efficiency differences relative to these observed prices.

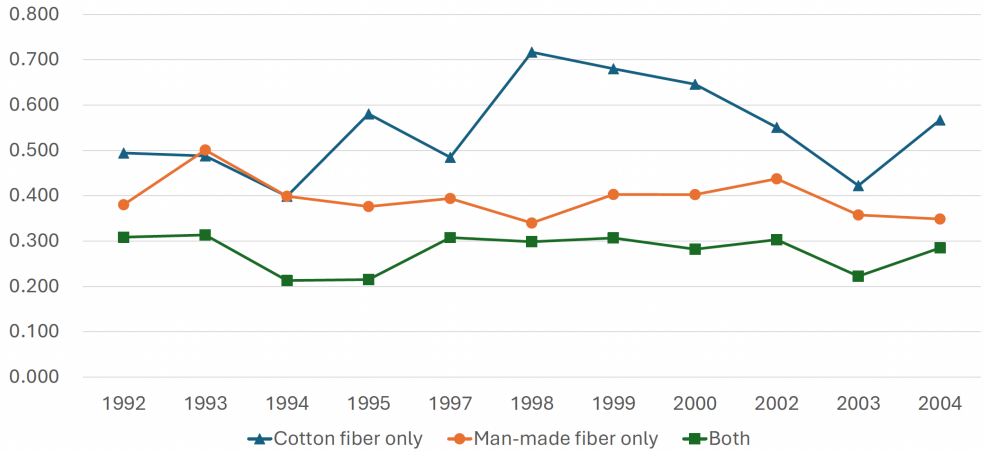
Table 2: Average plant characteristics, by product mix

Product mix	Employees (L)	Capital stock (K) (1,000 NTD)	Wage rate (W) (1,000 NTD)	K/L (1,000 NTD)
Cotton fiber only	61	194,677	280	1,961
Man-made fiber only	81	280,404	343	2,382
Both	191	788,705	395	4,649

To estimate the cost function for multiproduct plants, we focus on two product categories: Cotton Fiber (1310) and Man-Made Fiber (1360), which represent a substantial share of multiproduct producers in Taiwanese manufacturing. The data report both total revenue and physical quantity for each product, allowing us to construct output prices as the ratio of revenue to quantity.

This industry displays distinct patterns that reflect differences in production technologies between single-product and multiproduct plants. First, as shown in Table 1, approximately 431 plants produce either cotton fiber or man-made fiber in each year on average, with around 8.6 percent producing both. Table 2 compares plant characteristics across the three groups. On average, single-product plants employ 61 workers in the cotton fiber sector and 81 in the man-made fiber sector, while multiproduct plants are substantially larger, employing an average of 191 workers—roughly 2.3 times more. Multiproduct plants tend

Figure 1: Labor-to-materials expenditure over time, by plant type



to have higher capital stock, where the average capital stock of multiproduct plants is 2.8 times as large as that of single-product plants. Multiproduct plants also pay higher wages and exhibit greater capital intensity than their single-product counterparts. Second, Figure 1 shows the ratio of labor to materials expenditure from 1992 to 2004. Single-product plants consistently allocate a higher share of spending to labor compared to multiproduct plants, suggesting differences in input substitution elasticities or relative factor prices. Over time, the labor-to-materials expenditure ratio declines slightly, particularly among producers of man-made fiber and multiproduct producers, consistent with the adoption of labor-saving technologies during the sample period. Finally, within multiproduct plants, the composition of output varies widely. For instance, the revenue share of cotton fiber among these plants is just 6 percent at the 10th percentile, 45 percent at the median, and 84 percent at the 90th percentile, indicating considerable dispersion in product mix.

This industry provides an opportunity to compare the production technology of multiproduct and single-product producers. Our data allow us to estimate the production model separately for single-product producers and multiproduct producers and compare characteristics such as returns to scale and measures of economies of scope.[18]

[18]Prior studies estimating multiproduct translog models have had to deal with plants that produce zero amounts of some of the outputs. For example, instead of using logarithms, Caves et al. (1980b) adopt a Box-Cox transformation of output variables which can accommodate zero outputs. Another common approach is to impute small values of the outputs to approximate zero outputs (e.g., Cowing and Holtmann, 1983). Alternatively, De Loecker et al. (2016) estimate the production functions for single-product plants and assume that multiproduct plants are modeled as the combination of the single-product production.

6 Estimation Results

6.1 Demand Parameters

In this subsection, we report the estimates of the demand function (13), for two textile products: cotton fiber and man-made fiber. The demand function incorporates plant and time fixed effects in Y_{jt} to control for unobserved heterogeneity across plants and over time. To account for variation in local market demand, we include the total sales of the downstream clothing industry in the same city c in period $t-1$, $\ln(\text{DMS}_{ct-1})$, as an additional control variable. The demand function is estimated separately for the two products. The demand functions are estimated using the data for both single-product and multiproduct plants.

Table 3: First-stage estimates for $\ln(P_{njt})$

	Cotton Fiber	Man-made Fiber
$\ln(\text{DMS}_{ct-1})$	-0.275^{***}	-0.174^{***}
	(0.103)	(0.060)
$\ln(LP_{jt-1})$	-0.253^{***}	-0.133^{***}
	(0.035)	(0.025)
$\ln(W_{jt-1})$	0.075	0.075^{*}
	(0.080)	(0.041)
Year FE	Yes	Yes
Plant FE	Yes	Yes
Observations	1,017	2,002
Number of Plants	294	546
F-statistic (Relevance Test)	26.22	14.78

Note: Standard errors are in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The instrumental variables for product prices $\ln(P_{njt})$ are lagged labor productivity $\ln(LP_{jt-1})$, defined as the total plant output quantity per worker in logarithms, and lagged plant wage rate $\ln(W_{jt-1})$. These variables influence plants' marginal costs and are thus predictive of product prices, but they are unlikely to be influenced by contemporaneous demand shocks affecting product-level sales. Table 3 reports the first stage estimates of $\ln(P_{njt})$ on the set of instrumental variables.[19] Labor productivity is significantly correlated with the prices for both products. Higher-productivity plants have lower output prices. Plants with higher wage rates have higher output prices, although the statistical significance is weak. The F-statistics are greater than 10 for both products, indicating that the instrumental variables are relevant to the endogenous variable, $\ln(P_{njt})$.

[19]The number of observations is different from those reported in Table 1 because the lags of variables are

Table 4: Demand function estimates

	Cotton Fiber			Man-made Fiber		
	OLS	IV(1)	IV(2)	OLS	IV(1)	IV(2)
$\ln(P_{njt})$	-0.312^{***}	-2.389^{***}	-2.365^{***}	-0.161^{***}	-2.988^{**}	-2.738^{***}
	(0.035)	(0.326)	(0.320)	(0.026)	(0.589)	(0.511)
$\ln(\text{DMS}_{ct})$	-0.027	-0.615^{**}	-0.608^{**}	-0.029	-0.502^{**}	-0.461^{**}
	(0.100)	(0.258)	(0.255)	(0.059)	(0.205)	(0.186)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Plant FE	Yes	Yes	Yes	Yes	Yes	Yes
IVs	None	$\ln(LP_{jt-1})$	$\ln(LP_{jt-1})$ $\ln(W_{jt-1})$	None	$\ln(LP_{jt-1})$	$\ln(LP_{jt-1})$ $\ln(W_{jt-1})$
Observations	1,208	1,017	1,017	2,356	2,002	2,002
Number of Plants	485	294	294	900	546	546

Note: Standard errors are in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4 reports the estimated demand parameters. Columns (1) and (4) present the OLS estimates, which indicate inelastic demand for both products. Columns (2), (3), (5), and (6) report results from IV estimations that correct for potential price endogeneity. The IV estimates yield substantially more elastic demand. In particular, the estimated demand elasticities in Columns (3) and (6) are -2.365 for cotton fiber and -2.738 for man-made fiber, implying markups over variable cost of approximately 42 percent and 37 percent, respectively. These elasticities are used in the estimation of the output supply equations.[20] The coefficient on the demand shifter, sales in the downstream clothing industry in the city, is negative. This reflects the fact that larger markets tend to attract more competing plants, reducing the average number of customers per plant relative to smaller markets.[21]

6.2 Cost Parameters

Table 5 presents the estimated parameters of the translog cost function. Column (1) reports the results for multiproduct plants producing both cotton fiber and man-made fiber. Columns (3) and (5) report estimates for plants specializing in cotton fiber and man-made

used in Table 3.

[20] We experimented with different IVs and exogenous variables in the CES demand functions. The IV elasticity estimates, which are the only demand estimates used in estimation of the cost model in Section 4.2, are very robust.

[21] Cities with larger downstream clothing sales also tend to have more textile plants. The correlation between downstream clothing industry sales and the number of plants in the cotton fiber sector and the man-made fiber sector is 0.49 and 0.39, respectively.

fiber, respectively.[22]

Table 5: Cost function estimates

	Both Products		Cotton Fiber		Man-made Fiber	
Parameter	Coef.	(SE)	Coef.	(SE)	Coef.	(SE)
α_L	0.334^{***}	(0.087)	0.300^{***}	(0.103)	0.206^{**}	(0.100)
α_K	-0.496^{***}	(0.137)	-0.230^{***}	(0.066)	-0.072^{**}	(0.033)
β_1	0.421^{***}	(0.119)	0.846^{***}	(0.291)		
β_2	0.445^{***}	(0.115)			0.897^{***}	(0.131)
δ_{LL}	-0.154^{***}	(0.052)	-0.048	(0.032)	-0.024	(0.017)
δ_{KL}	-0.293^{***}	(0.033)	-0.124^{***}	(0.027)	0.066^{***}	(0.006)
δ_{KK}	-0.038	(0.092)	0.042	(0.053)	0.098^{***}	(0.024)
γ_{11}	0.179^{***}	(0.034)	0.130	(0.085)		
γ_{12}	-0.159^{**}	(0.076)				
γ_{22}	0.156^{***}	(0.046)			0.076^{***}	(0.029)
ϕ_{L1}	0.202^{**}	(0.082)	0.254^{***}	(0.076)		
ϕ_{L2}	0.268^{***}	(0.083)			-0.180^{***}	(0.023)
ϕ_{K1}	-0.061^{**}	(0.025)	-0.075^{***}	(0.019)		
ϕ_{K2}	0.022	(0.024)			-0.013	(0.008)
g_1	0.821^{***}	(0.028)	0.722^{***}	(0.029)		
g_2	0.779^{***}	(0.028)			0.573^{***}	(0.020)
g_L	0.308^{***}	(0.048)	0.365^{***}	(0.037)	0.431^{***}	(0.027)
g_M	0.875^{***}	(0.053)	0.846^{***}	(0.031)	0.713^{***}	(0.038)

Note: Standard errors are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Cost elasticities. In the translog specification, cost elasticities with respect to outputs and capital depend on input prices, capital stock, output levels, and plant-specific quality and efficiency terms. Table 6 reports the mean and dispersion (i.e., 10th percentile and 90th percentile) of these cost elasticities for both multiproduct and single-product plants. Among multiproduct plants, the average cost elasticities with respect to outputs of cotton fiber and man-made fiber are 0.421 and 0.445, respectively. The corresponding dispersions

[22] A regularity condition for the cost function is concavity in variable input prices. The key parameter that governs concavity in our model is the second-order parameter δ_{LL}. Specifically, concavity is implied by the inequality: $\delta_{LL} + \frac{E_{Ljt}}{C_{jt}}\left(\frac{E_{Ljt}}{C_{jt}} - 1\right) < 0$. This condition is satisfied for all observations in our sample for both multiproduct and single-product plants.

(measured as the difference between the 90th and 10th percentiles), 0.657 and 0.709, indicate substantial heterogeneity in cost elasticities across plants. Such dispersion, although smaller, is also present for the single-product plants.

Table 6: Key characteristics of estimated cost functions

	Both Products			Cotton Fiber			Man-Made Fiber		
	Mean	10 Pct	90 Pct	Mean	10 Pct	90 Pct	Mean	10 Pct	90 Pct
$\frac{\partial c_{jt}}{\partial q_{1jt}}$	0.421	0.067	0.724	0.846	0.682	1.012	-	-	-
$\frac{\partial c_{jt}}{\partial q_{2jt}}$	0.445	0.088	0.797	-	-	-	0.897	0.806	0.985
S_{jt}	1.253	0.850	1.559	1.279	0.950	1.556	1.211	0.883	1.532
$\frac{\partial c_{jt}}{\partial k_{jt}}$	-0.496	-0.717	-0.298	-0.230	-0.320	-0.140	-0.072	-0.265	0.110
σ_{LMjt}	2.194	1.696	2.947	1.343	1.196	1.592	1.201	1.107	1.333

Economies of scale. Multiproduct economies of scale are defined as the inverse of the sum of the output elasticities (Bailey and Friedlaender, 1982):

$$S_{jt} \equiv \frac{1}{\sum_n \frac{\partial c_{jt}}{\partial q^*_{njt}}}, \tag{33}$$

where $\frac{\partial c_{jt}}{\partial q^*_{njt}}$ is the cost elasticity with respect to output n, as defined in equation (8). Values of $S_{jt} > 1$ indicate increasing returns to scale, while values below 1 indicate decreasing returns.

The average value of S_{jt} is 1.25 with a standard deviation of 0.34, indicating that, on average, plants exhibit increasing short-run returns to scale. Approximately 82 percent of plant-year observations have $S_{jt} > 1$, indicating that increasing returns are common but not universal in the sample.

The estimated second-order cost parameters for outputs, γ_{11} and γ_{22}, are both positive and statistically significant. This implies that cost elasticities increase with output levels, meaning that plants producing higher output quantities experience diminishing scale economies (i.e., lower S_{jt}), holding other factors constant.

For single-product plants, the mean output elasticities are estimated at 0.846 for cotton fiber producers and 0.897 for producers of man-made fiber. The implied economies of scale, defined as the inverse of the output elasticity, are 1.279 for cotton fiber producers and 1.211 for producers of man-made fiber, on average. Approximately 86 percent of cotton fiber producers and 80 percent of producers of man-made fiber operate with increasing returns

to scale, as indicated by an economies-of-scale index above one. Consistent with the results for multiproduct plants, these elasticities increase with output levels, reflecting the positive values of γ_{11} and γ_{22}.

The mean elasticity of variable cost with respect to capital is negative across all plant types: -0.496 for multiproduct plants, -0.230 for plants producing only cotton fiber, and -0.072 for plants producing only man-made fiber. These estimates imply that higher capital stock reduces variable costs. However, there is substantial variation across plants, as demonstrated by the 10th and 90th percentiles in Table 6.

Elasticity of substitution. The elasticity of substitution between labor and materials is given by:

$$\sigma_{LMjt} = 1 + \frac{-\delta_{LL}}{\frac{E_{Ljt}}{C_{jt}} \cdot \frac{E_{Mjt}}{C_{jt}}}, \tag{34}$$

and varies across plants depending on their input expenditure shares and estimated curvature parameter δ_{LL}. As shown in Table 6, the mean elasticity of substitution is 2.19 for multiproduct plants with variation between 1.7 and 2.9 at the 10th and 90th percentiles. In contrast, single-product plants exhibit a lower mean elasticity of substitution and less heterogeneity across observations. For cotton fiber producers, the mean of σ_{LMjt} is 1.34, while for producers of man-made fiber it is 1.20. These values indicate that multiproduct plants are more responsive to changes in relative input prices, presumably reflecting opportunities to adjust the mix of outputs in production or to adjust the decision to outsource some initial stages of production.

Overall, our estimates of the elasticity of substitution exceed one for all observations, in multiproduct plants and in single-product plants. This finding is consistent with studies by Chan (2023) and Mertens and Schoefer (2025). Both studies focus on materials-labor substitution. Chan (2023) focuses on the firm's make or buy decision regarding materials and finds that the outsourcing of tasks contributes to a declining labor share. His estimates of the elasticity of substitution vary from 1.5 to 4. Mertens and Schoefer (2025) use firm and industry data for many countries and report elasticities between 1.8 and 4.2. They emphasize that the outsourcing decision for material inputs is important for explaining firm growth and the declining labor share.[23]

[23]Given that we correct for material price heterogeneity using the efficiency measure, our estimates are comparable to those in Grieco et al. (2016, 2022), Li and Zhang (2022), and Harrigan et al. (2024), which similarly control for heterogeneous material prices and estimate CES production functions using different datasets from Colombia, China, and France. These studies assume a constant elasticity of substitution across labor, materials, and capital. In addition, Berkowitz et al. (2017) report average elasticity estimates above one across Chinese industries.

Persistence of efficiency and quality. The model estimates the persistence of unobserved input efficiency and output quality as autoregressive parameters shown in equations (18) to (20). For multiproduct plants, the estimated persistence in output quality is high: $g_1 = 0.821$ and $g_2 = 0.779$ for the two outputs, indicating strong serial correlation in product-specific quality. The persistence of material input efficiency is also high, with $g_M = 0.875$. In contrast, the persistence of relative labor efficiency $\tilde{\mu}_{Ljt}$, $g_L = 0.308$, is substantially lower, suggesting greater volatility in the efficiency gap between labor and materials. For single-product plants, these persistence parameters also indicate statistically significant serial correlation, though their magnitudes are generally smaller than those estimated for multi-product plants, implying less persistence in plant-level input efficiency and output quality.

6.3 Implications of Separability and Non-joint Production

A special case of our multiproduct variable cost function is one in which inputs and outputs are separable. In our case, this implies that the plant's output vector can be aggregated into a single output $Q^*(Q^*_{1jt}, Q^*_{2jt})$ and the variable cost function can be expressed as $C_{jt} = C(Q^*(Q^*_{1jt}, Q^*_{2jt}), W^*_{Ljt}, W^*_{Mjt}, K^*_{jt})$. Input-output separability implies that the marginal rate of transformation between Q^*_{1jt} and Q^*_{2jt} is independent of factor prices and fixed factors, and that output differences, including quality differences (ν_{1jt} and ν_{2jt}), do not affect input cost shares. In our model, this implies the following testable restrictions: $\phi_{L1} = \phi_{L2} = \phi_{K1} = \phi_{K2} = 0$. The Wald test statistic for this hypothesis is 26.884, indicating that input-output separability is rejected at the 1 percent significance level. Similarly, for single-product plants, the Wald test statistics are 26.527 and 63.580 for cotton fiber plants and plants producing man-made fiber, respectively, also rejecting input-output separability in both cases at the 1 percent significance level.

The assumption of non-joint production is also testable in our framework. Non-joint production implies that the cost function can be expressed as separable cost functions for each output: $C_{jt} = C(C_1(Q^*_{1jt}, W^*_{Ljt}, W^*_{Mjt}, K^*_{1jt}), C_2(Q^*_{2jt}, W^*_{Ljt}, W^*_{Mjt}, K^*_{2jt}))$. This imposes that the marginal cost of each output is not a function of the other output level, a restrictive starting point for modeling multiproduct production. The testable restriction for our translog variable cost function is $\gamma_{12} = 0$. Our estimate of $\gamma_{12} = -0.159$ (0.076) rejects this hypothesis at the 5 percent significance level, providing evidence against the assumption of non-joint production. This complementarity, which is an important part of multiproduct production costs, cannot be estimated from single-product cost functions.

We can quantify how changes in the output of one product affect the marginal cost of

producing the other product. For example, the elasticity of the marginal cost of product 1 (Q^*_{1jt}) with respect to the output of product 2 (Q^*_{2jt}) is given by:

$$\frac{\partial \ln MC_1(Q^*_{1jt}, Q^*_{2jt})}{\partial \ln Q^*_{2jt}} = \frac{\gamma_{12}}{\frac{\partial \ln C(Q^*_{1jt}, Q^*_{2jt})}{\partial \ln Q^*_{1jt}}} + \frac{\partial \ln C(Q^*_{1jt}, Q^*_{2jt})}{\partial \ln Q^*_{2jt}}. \tag{35}$$

This elasticity is negative when $\gamma_{12} < -\frac{\partial \ln C(Q^*_{1jt}, Q^*_{2jt})}{\partial \ln Q^*_{1jt}} \frac{\partial \ln C(Q^*_{1jt}, Q^*_{2jt})}{\partial \ln Q^*_{2jt}}$, which holds for 71 percent of the observations in our data on multiproduct plants. The median of the elasticity defined in (35) is approximately -0.11, meaning that a 1 percent increase in the output of product 2 reduces the marginal cost of product 1 by 0.11 percent. Similarly, the elasticity of the marginal cost of product 2 with respect to the output of product 1 is -0.10 on average.

We use the separate multiproduct and single-product cost functions reported in Table 5 to calculate the economies of scope. A measure of economies of scope is defined as the sum of the cost of producing the two outputs in separate plants relative to the cost in the multiproduct plant:

$$\frac{C_1(Q^*_1, W^*_L, W^*_M, K^*) + C_2(Q^*_2, W^*_L, W^*_M, K^*)}{C(Q^*_1, Q^*_2, W^*_L, W^*_M, K^*)}, \tag{36}$$

where $C_1(\cdot)$ and $C_2(\cdot)$ are the estimated single-product cost functions for plants producing only product 1 or product 2, respectively, and $C(\cdot)$ is the estimated cost function for plants producing both products.

We evaluate the economies of scope (36) using the quality-adjusted vector of outputs of the multiproduct plants, (Q^*_{1jt}, Q^*_{2jt}), with efficiency-adjusted factor prices and capital stock fixed at the corresponding means of these multiproduct plants, $(W^*_L, W^*_M, K^*) = (\overline{W}^*_{Ljt}, \overline{W}^*_{Mjt}, \overline{K}^*_{jt})$. On average, the estimated economies of scope equal 1.255, implying that producing the two outputs separately would increase the short-run variable costs by 25.5 percent, compared with producing them jointly.[24] The standard deviation is 0.407, indicating that the gains from joint production vary considerably across plants.[25]

[24] This positive cost saving is consistent with that in Koh and Raval (2025), who document that a firm's revenue is higher when more shared inputs are used. Their result is based on uniquely detailed business-line-level data from the U.S. Federal Trade Commission's Line of Business Surveys which contain specific accounts of shared inputs.

[25] Alternatively, if data for single-product plants is not available, the multiproduct cost function can be used to construct a measure of economies of scope as: $\frac{C(Q^*_1, \underline{Q}^*_2, W^*_L, W^*_M, K^*) + C(\underline{Q}^*_1, Q^*_2, W^*_L, W^*_M, K^*)}{C(Q^*_1, Q^*_2, W^*_L, W^*_M, K^*)}$, where $C(\cdot)$ is the estimated cost function for multiproduct plants and $\underline{Q}^*_1$ and $\underline{Q}^*_2$ are small imputed values for each of the outputs. The translog function does not allow evaluating $\underline{Q}^*_1$ and $\underline{Q}^*_2$ at zero, so small positive levels of

7 Input Efficiencies and Output Qualities

Our model generates plant-year estimates of output quality and input efficiency. These measures are positively associated with their corresponding raw prices. For example, among multiproduct plants, the correlation coefficients between ν_1, ν_2, and $\tilde{\mu}_L$ and their respective raw prices, p_1, p_2, and $\tilde{w}_L$, are 0.28, 0.21, and 0.10. This suggests that the estimated measures capture meaningful heterogeneity in plant-product quality and input efficiency.

Heterogeneity over time vs. across plants. To summarize dynamics of these measures over time and the extent of heterogeneity across plants, we estimate a random effects regression model with a time trend as the explanatory variable. The results, presented in Table 7, indicate statistically significant and positive time trends for all four input-efficiency and output-quality measures. Input efficiency shows steady growth over the sample period. Labor efficiency increases at an average annual rate of 3.6 percent, while material efficiency grows at 2.6 percent per year. Consequently, the relative efficiency term for labor ($\tilde{\mu}_L = \mu_L - \mu_M$) exhibits an annual growth rate of 1.0 percent, which implies a decline in the relative price of labor in efficiency units. In terms of output quality, the quality of cotton fiber shows a strong upward trend, with an average annual growth rate of 6.2 percent, whereas the quality of man-made fiber grows at a slower rate of 1.5 percent per year. In addition to these time trends, all four measures exhibit substantial and persistent heterogeneity across plants. The estimated plant-level random effects account for between 62.1 percent and 78.7 percent of the total variation in the measures, highlighting the importance of plant-specific factors in shaping both input efficiencies and output qualities.

Physical units vs. quality- and efficiency-adjusted units. To explore the importance of output-quality and input-efficiency terms as sources of output and input heterogeneity, we decompose the variance of logarithmic quality-adjusted outputs and efficiency-adjusted input prices as $\text{Var}(x^*) = \text{Var}(x) + \text{Var}(e) + 2\text{Cov}(x, e)$, where x represents physical variables $(q_1, q_2, \tilde{w}_L)$, and e corresponds to $(\nu_1, \nu_2, -\tilde{\mu}_L)$. Table 8 presents the results, showing that differences in output quality and input efficiency contribute significantly to the variance of adjusted variables. For the two outputs, the variances of the quality terms in Column (4) are approximately 72 percent and 60 percent of the variances in physical units in Column (3), respectively. For the wage rate, the role of input efficiency is more pronounced, with the variance of the efficiency term equal to 262 percent of the variance in the physical wage rate.

each output can be used. As a check we set each imputed output equal to the smallest value of Q^*_{1jt} and Q^*_{2jt} within the sample of multiproduct plants. Similar to our measure using equation (36), this produces estimates of economies of scope that, on average, equal 1.305, with standard deviation of 0.221.

Table 7: Random effect regression results

Variable	time trend parameter	fraction of variation due to plant effects
μ_L	0.036^{***}	0.787
	(0.008)	
μ_M	0.026^{***}	0.749
	(0.009)	
ν_1	0.062^{***}	0.659
	(0.012)	
ν_2	0.015^{*}	0.621
	(0.008)	

Note: Standard errors are in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 8: Decomposition of efficiency-adjusted variables

x^*	$Var(x^*)$	$Var(x)$	$Var(e)$	$2Cov(x,e)$
$\tilde{w}_L^*$	0.224	0.119	0.312	-0.208
q_1^*	1.582	1.963	1.406	-1.787
q_2^*	1.062	1.012	0.602	-0.552

Note: This table decomposes the variance of efficiency-adjusted variables according to the identity: $\text{Var}(x^*) = \text{Var}(x) + \text{Var}(e) + 2\text{Cov}(x,e)$, where x denotes physical quantities $(q_1, q_2, \tilde{w}_L)$ and e corresponds to the associated efficiency terms $(\nu_1, \nu_2, -\tilde{\mu}_L)$.

As shown in the last column of the table, the quality terms for outputs are negatively correlated with physical quantities. This is consistent with large producers focusing on lower-quality, high-volume products, a pattern that has been documented in other studies.[26] The negative covariance between $\tilde{w}_L$ and $-\tilde{\mu}_L$ indicates that the physical wage rate and labor efficiency are positively correlated. Plants with a higher wage rate have higher labor efficiency, consistent with employing higher-quality workers. These correlations influence the dispersion in efficiency/quality-adjusted variables compared with their physical units. The variance of cotton fiber output declines from 1.963 to 1.582 after adjusting for quality, while for man-made fiber, quality adjustment has little effect on the output variance. In contrast, for the wage rate, the variance in the efficiency-adjusted units is substantially greater than the variance in physical units.

[26]For example, Holmes and Stevens (2014) study the U.S. furniture industry and show that large firms produce standardized outputs while small firms produce more specialized high-quality output. Grieco and McDevitt (2017) report a quality-quantity trade-off elasticity of -0.2 in U.S. dialysis centers.

Factor-biased technology and labor shares. Our model incorporates efficiency differences in labor and materials and allows for non-neutral differences in technology. Hicks neutrality, which requires $\tilde{\mu}_{Ljt} = \mu_{Ljt} - \mu_{Mjt} = 0$, cannot generally be imposed across all data points. This is because, after imposing Hicks neutrality, the system of equations in (21) would have to reconcile three composite-error equations using only the two output-quality terms, ν_{1jt} and ν_{2jt}. The estimates of $\tilde{\mu}_{Ljt}$ exhibit substantial dispersion, with a standard deviation of 0.558, indicating that Hicks neutrality does not characterize the patterns of the input efficiencies.

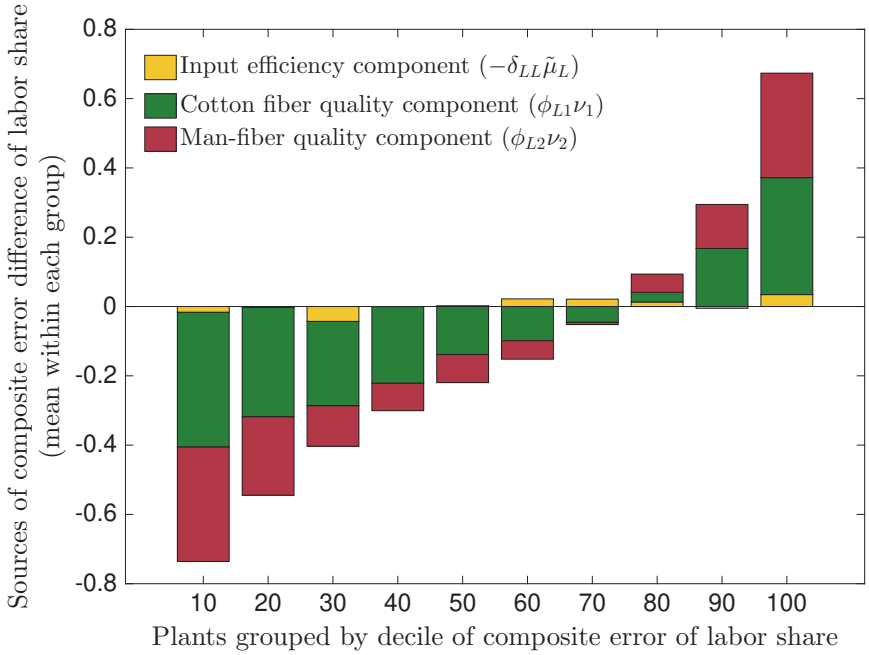
Recent studies of biased technological change have focused on its impact on labor shares of revenue (e.g., Doraszelski and Jaumandreu, 2018; Raval, 2019; Zhang, 2019; Rubens et al., 2026). A consistent finding is that the elasticity of substitution between labor and other inputs is less than one and that labor efficiency varies substantially across plants and time. Together these imply that improvements in labor efficiency due to technical change, which thus lower the effective price for labor, contribute to smaller labor shares.[27]

Our estimates imply that the elasticity of substitution between labor and materials exceeds one. As a result, an increase in labor efficiency raises the labor share of variable cost in this industry, consistent with plants bringing more stages of production "in house" and reducing their reliance on purchased inputs. Equation (15) shows this mechanism directly: holding other determinants fixed, higher labor efficiency, $\tilde{\mu}_{Ljt}$, lowers the effective wage. Since $\delta_{LL} < 0$, this reduction in the effective wage increases the labor share.

The model also allows the labor share of variable cost to depend on output quality, as reflected in the nonzero estimates of ϕ_{L1} and ϕ_{L2} in (15). This output-side heterogeneity is quantitatively more important than input-efficiency variation in accounting for differences in labor shares across plants. Figure 2 illustrates this point by decomposing the labor-share composite error, ϵ_{Ljt}, in (15) for multiproduct plants. Plants are ordered by the value of the composite error and grouped into deciles. Each bar reports the average contribution within the decile of three latent components: input efficiency, cotton fiber quality, and man-made fiber quality. The dominant pattern is that output-quality components account for most of the variation across deciles. Plants in the lower deciles have below-average composite errors primarily because their output quality components are below average. Moving toward the upper deciles, these quality components rise and contribute positively to the composite

[27] With the exception of Doraszelski and Jaumandreu (2018), the papers use CES production models where the elasticity of substitution captures substitution between labor and different combinations of capital and materials. Our model estimates the substitution elasticity between labor and materials holding capital and outputs fixed.

Figure 2: Sources of composite error differences in labor share



error, and hence to the labor share. By contrast, the input-efficiency component is small throughout the distribution and contributes only modestly relative to the two output-quality components. This output-quality channel is absent from much of the existing literature on biased technological change. If the transformation function were separable between inputs and outputs, the interaction terms ϕ_{L1} and ϕ_{L2} would be zero, and labor shares would be unaffected by changes in output quality. Our estimates reject this separability restriction.

Product quality and product mix. Our model provides estimates of within-plant heterogeneity in output quality among multiproduct plants. The estimated differences $\nu_{1jt} - \nu_{2jt}$ have a standard deviation of 0.988, indicating significant within-plant heterogeneity in output quality. The correlation between the two quality components, (ν_1, ν_2), is 0.561, implying that plants producing high-quality cotton fiber also tend to produce high-quality man-made fiber.

These findings align with recent work on within-firm heterogeneity (e.g., Orr, 2022; Caselli et al., 2026), which has examined how product-level productivity affects the relative revenue shares within the same firm. Because our translog framework does not impose input-output separability, it allows the plant's revenue shares to be affected by input efficiency as well as

output quality.

In particular, for multiproduct plants, the supply of one output in equation (16) depends not only on its own quality, but also on the quality and quantity of the other output. The relevant coefficient is γ_{12}, which we estimate as negative, implying that an increase in the quantity or quality of one output lowers the revenue share of the other output. In addition, because input prices and input efficiency affect the supply of both products, as implied by the coefficients ϕ_{L1} and ϕ_{L2}, their impacts can differ across outputs when $\phi_{L1} \neq \phi_{L2}$. The same change in factor-biased technology can shift the supply of cotton fiber and man-made fiber by different magnitudes, generating variation in product mix across plants. For example, a 10-percentage-point increase in $\tilde{\mu}_{Ljt}$ decreases the markup-adjusted revenue-to-cost ratio in equation (16) by 0.020 for cotton fiber but by 0.027 for man-made fiber.

8 A Multilateral Productivity Index

Our empirical model identifies and quantifies the various factors that influence a plant's variable costs. In this section, we develop a cost-based index that provides a natural measure of plant-level productivity and enables consistent comparisons of productivity across plants and over time.[28] We further demonstrate how this index can be decomposed into distinct components capturing the contributions of capital input, scale economies, input efficiency, and output quality.

Deriving a multilateral productivity index. We begin by totally differentiating the cost function with respect to its arguments:

$$\begin{aligned} dc_{jt} &= \sum_{x=\ell,m} \left[\frac{\partial c_{jt}}{\partial w^*_{xjt}}\right] dw^*_{xjt} + \left[\frac{\partial c_{jt}}{\partial k^*_{jt}}\right] dk_{jt} + \sum_n \left[\frac{\partial c_{jt}}{\partial q^*_{njt}}\right] dq^*_{njt} \\ &= \sum_{x=\ell,m} \left[\frac{\partial c_{jt}}{\partial w^*_{xjt}}\right] dw_{xjt} + \left[\frac{\partial c_{jt}}{\partial k^*_{jt}}\right] dk_{jt} + \sum_n \left[\frac{\partial c_{jt}}{\partial q^*_{njt}}\right] dq_{njt} + u_{jt}, \end{aligned} \tag{37}$$

where the residual term u_{jt} captures changes in input efficiency and output quality and is defined as:

[28] The measurement of total factor productivity from the dual price side was discussed by Jorgenson and Griliches (1967). They showed using a national income accounting identity, that a Divisia productivity index defined as the weighted sum of outputs minus the weighted sum of inputs was equivalent to an index of the weighted sum of input prices minus the weighted sum of output prices. The duality between productivity indexes defined from the production function and those defined using the cost function is discussed in Diewert (1980, 1981), Denny et al. (1981), Hulten (1986), and Morrison (1992).

$$u_{jt} \equiv -\sum_{x=\ell,m}\left[\frac{\partial c_{jt}}{\partial w^*_{xjt}}\right] d\mu_{xjt} + \sum_{n}\left[\frac{\partial c_{jt}}{\partial q^*_{njt}}\right] d\nu_{njt}. \tag{38}$$

The term u_{jt} isolates the contribution of unobserved changes in input efficiency and output quality to the change in variable costs.

Replace the cost elasticity with respect to the factor prices by the factor cost shares (4) and (5) and replace the cost elasticity with respect to the output by revenue-cost ratio (8):

$$\frac{\partial c_{jt}}{\partial w^*_{xjt}} = \frac{E_{xjt}}{C_{jt}} \equiv e_{xjt}, \quad x = \{l, m\} \tag{39}$$

and

$$\frac{\partial c_{jt}}{\partial q^*_{njt}} = \frac{\eta_n - 1}{\eta_n}\frac{R_{njt}}{C_{jt}} \equiv r_{njt}, \quad n = 1, \ldots, N. \tag{40}$$

Denote the cost elasticity with respect to the capital stock as:

$$\frac{\partial c_{jt}}{\partial k^*_{jt}} \equiv a_{jt}. \tag{41}$$

Substitute (39), (40), and (41) into (37) to obtain:

$$dc_{jt} \quad = \quad \sum_{x=\ell,m} e_{xjt} dw_{xjt} + a_{jt} dk_{jt} + \sum_{n} r_{njt} dq_{njt} + du_{jt}. \tag{42}$$

Because the translog variable cost function is quadratic in the logarithms of its arguments, the quadratic approximation lemma of Diewert (1976) can be used to obtain an exact discrete representation of the log cost difference between any two observations. The lemma implies that the change in log cost is equal to a weighted sum of the changes in the function's arguments, where the weights are given by the averages of the corresponding first-order derivatives evaluated at the two observations. When applying this result to panel data, Caves et al. (1982) propose a multilateral index that measures the discrete difference between each observation and a common reference point, defined as the average of the sample values. In the translog case, this reference point can be interpreted as a representative plant characterized by the sample geometric means of output levels, capital stock, and factor prices.

To construct a discrete index of the cost-function differential in (42), define the deviation of each variable from its sample mean as $\Delta x_{jt} = x_{jt} - \bar{x}$. Using the results of Diewert (1976)

and Caves et al. (1982), the difference between log cost for plant j in period t and mean log cost can be written as

$$\Delta c_{jt} = \sum_{x=\ell,m} \frac{1}{2}\left(e_{xjt} + \bar{e}_x\right)\Delta w_{xjt} + \sum_n \frac{1}{2}\left(r_{njt} + \bar{r}_n\right)\Delta q_{njt} + \frac{1}{2}\left(a_{jt} + \bar{a}\right)\Delta k_{jt} - \sum_{x=\ell,m} \frac{1}{2}\left(e_{xjt} + \bar{e}_x\right)\Delta \mu_{xjt} + \sum_n \frac{1}{2}\left(r_{njt} + \bar{r}_n\right)\Delta \nu_{njt}. \tag{43}$$

The first term on the right-hand side is a Törnqvist index of factor prices, and the second is a Törnqvist index of output quantities. The remaining terms are Törnqvist indexes of capital stock, input efficiency, and output quality.[29]

To define a multilateral productivity index, we first normalize the weights in the output quantity index so that they sum to one: $b_{njt} \equiv \frac{r_{njt}}{\sum_s r_{sjt}} = S_{jt} r_{njt}$. This normalization removes the effect of scale economies from the definition of productivity, while allowing scale economies to enter separately as a source of productivity differences. The multilateral multiproduct productivity index is then defined as

$$\Delta MPP_{jt} \equiv -\left\{\Delta c_{jt} - \sum_{x=\ell,m} \frac{1}{2}\left(e_{xjt} + \bar{e}_x\right)\Delta w_{xjt} - \sum_n \frac{1}{2}\left(b_{njt} + \bar{b}_n\right)\Delta q_{njt}\right\}. \tag{44}$$

The right-hand side measures the difference in variable cost after netting out the effects of factor prices and output quantities. It can be constructed from data on variable costs, input expenditures, factor prices, product revenues, and output quantities. The only parameters required for its construction are the demand elasticities that enter the revenue-share weights in (40).

There are two special cases in which estimates of the demand elasticities are not required to calculate the multiproduct productivity index, ΔMPP_{jt}.[30] First, if product demand elasticities are identical across products, then the output weight for each product is simply its share of total revenue, so the weights can be constructed directly from observed revenue and cost data. Second, if the plant produces a single output, the revenue-based output

[29] Using a related approach based on the translog cost function and the quadratic approximation lemma, several authors have developed decompositions of cost changes. Using a long-run multiproduct cost function with time-series data, Diewert and Fox (2008, eq. 32) decompose changes in total cost into components due to scale economies, input-price variation, and factor-biased technical change. They allow markup pricing for outputs. Using panel data for electric utilities, Gollop and Roberts (1981) decompose the growth of electricity generation costs into scale economies and biased technological change. Denny et al. (1981) study the Canadian telephone industry, attributing cost changes over time to scale economies, departures from marginal-cost pricing, and technical change.

[30] Diewert and Fox (2008) obtain a similar result in their framework.

weight is equal to one, and ΔMPP_{jt} is the change in average variable cost, $\Delta c_{jt} - \Delta q_{jt}$, minus the change in cost-share-weighted factor prices. Thus, a reduction in average variable cost, after controlling for factor-price differences, indicates an increase in productivity.

In addition to constructing an index of productivity differences, we can use the econometric estimates of the cost function to decompose those differences into components associated with capital input, scale economies, output quality, and input efficiency. Substituting the exact cost difference in (43) into the definition of productivity expresses ΔMPP_{jt} in terms of these cost-function components:

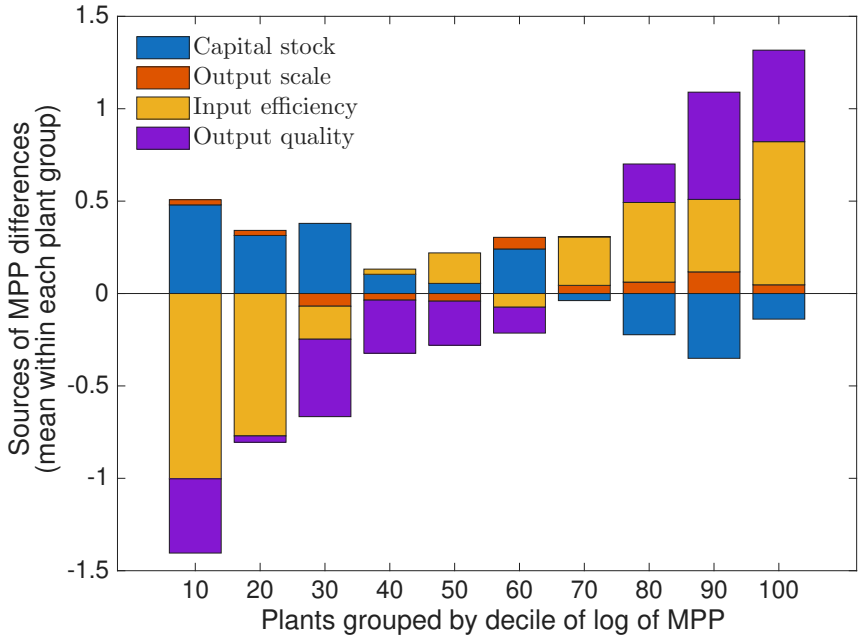
$$\begin{aligned}\Delta MPP_{jt} = &-\frac{1}{2}\left(a_{jt} + \bar{a}_K\right)\Delta k_{jt} \\ &- \sum_n \frac{1}{2}\Big[r_{njt}(1 - S_{jt}) + \bar{r}_n(1 - \bar{S})\Big]\Delta q_{njt} \\ &+ \sum_{x=\ell,m} \frac{1}{2}\left(e_{xjt} + \bar{e}_x\right)\Delta \mu_{xjt} - \sum_n \frac{1}{2}\left(r_{njt} + \bar{r}_n\right)\Delta \nu_{njt}. \end{aligned} \tag{45}$$

The first line represents the contribution of capital stock to productivity. The second line captures the contribution of outputs when there are economies or diseconomies of scale. The final line captures the contributions of input efficiency and output quality. Constructing the right-hand side of this decomposition requires econometric estimates of the cost elasticity with respect to capital, returns to scale, input efficiencies, and output qualities.

This equation illustrates how different plant characteristics contribute to the multilateral productivity index ΔMPP_{jt}. An increase in capital stock reduces variable cost because $\frac{\partial c_{jt}}{\partial k_{jt}} < 0$ and thus contributes positively to ΔMPP_{jt}. The effect of output expansion on ΔMPP_{jt} depends on the presence of economies of scale. If the plant exhibits diseconomies of scale ($S_{jt} < 1$), an increase in output raises variable cost and therefore lowers ΔMPP_{jt}. Conversely, if there are economies of scale ($S_{jt} > 1$), an output expansion reduces variable cost and raises ΔMPP_{jt}. When returns to scale are constant for all observations ($S_{jt} = \bar{S} = 1$), output differences among plants have no effect on ΔMPP_{jt}. Improvements in either labor (μ_{Ljt}) or materials (μ_{Mjt}) efficiency lower the efficiency-adjusted input prices and thus contribute positively to ΔMPP_{jt}. In contrast, increases in output qualities (ν_{1jt}, ν_{2jt}) raise the quality-adjusted output, which is costly to produce (i.e., $\frac{\partial c_{jt}}{\partial q^*_{njt}} > 0$), thereby reducing MP.

Figure 3 illustrates how the four distinct components, capital stock, output quantity, input efficiency, and output quality, contribute to differences in MPP across plants. In this figure, plants are grouped into deciles based on their MPP levels. In the first decile, the average

Figure 3: Sources of MPP differences



MPP is −0.90. These plants tend to have larger capital stock, lower output quantity, lower input efficiencies, and higher output qualities, relative to the sample means. Accordingly, capital stock contributes positively (0.48 log points), input efficiencies contribute negatively (−1.00 log points), and output quality also contributes negatively (−0.40 log points) to MPP in this group. The contribution of output quantity is small (0.03 log point) because plants' returns to scale are distributed around unity (with a mean of 1.25 and a standard deviation of 0.34). Together, these patterns indicate that plants producing high-quality output but operating with low input efficiency tend to lie at the lower end of the MPP distribution, even though they possess relatively large capital stocks.

As we move from the lowest to the highest MPP decile, the composition of plant characteristics changes substantially: capital stock declines, input efficiency improves, and output quality decreases.[31] This shift is reflected in the contribution patterns. The positive effect of capital stock on MPP diminishes and eventually turns negative in the higher MPP deciles. In contrast, the contribution of input efficiency increases steadily, becoming the dominant factor in the upper deciles. Likewise, the contribution of declining output quality becomes in-

[31] As expected, quality-adjusted output prices increase, while efficiency-adjusted wage rates decline as we move toward the highest MPP decile.

creasingly positive. In the top decile, capital stock contributes negatively (-0.14 log points), while high input efficiency and lower output quality contribute positively to MP, 0.77 and 0.50 log points, respectively. Output quantity continues to contribute little to MPP variation across deciles, consistent with our earlier finding that returns to scale are distributed around unity. Overall, plants in the highest MPP decile exhibit productivity levels that are 2.08 log points higher than those in the lowest decile. The majority of this advantage arises from producing lower-quality output with superior input efficiency. Their efficiency gains more than compensate for the disadvantage of having relatively smaller capital stocks.[32]

In summary, the MPP index offers an implementable, cost-based measure of plant performance with a clear decomposition into capital stock, scale effects, input efficiency, and output quality. This framework enables consistent productivity comparisons across plants and over time. Our empirical results show that variation in MPP is driven primarily by differences in input efficiency and output quality, while output scale plays only a minor role.

9 Conclusion

This paper develops a model of multiproduct production and demand that can be estimated with micro panel data on revenues and quantities of outputs, expenditures and prices of variable inputs, fixed input levels, and exogenous product demand variables. Our focus is on the flexibility that can be obtained when researchers can access product-level data on revenues and outputs for multiproduct plants, in addition to the more widely available data on input use, factor prices, and total revenue. The objects that are estimated include product demand curves, demand elasticities, and a flexible representation of technology that quantifies marginal rates of transformation, scale economies, elasticities of substitution, shadow values of fixed factors, input efficiencies, and output qualities that vary by data observation.

The key starting point is a flexible translog multiproduct variable cost function that does not impose input-output separability, non-joint production, constant returns to scale, or constant elasticity of substitution. The cost function incorporates input efficiencies and output qualities as unobservable variables, allowing for biased technology and product-quality heterogeneity. This is augmented with the system of input demand equations and output supply equations that are derived from plants' first-order conditions. We incorporate timing assumptions about plants' information sets that are used for constructing moment conditions

[32]Note that MPP is a short-run, cost-side productivity measure conditional on fixed capital stock, not a welfare or profit ranking. A plant producing higher-quality output may have lower MPP even if it earns higher prices and profits.

for estimation.

Compared with the seminal multiproduct-cost studies of the 1970s and 1980s, our framework treats unobserved output quality and input efficiency as determinants of the plant's endogenous input and output choices. Relative to recent "input-allocation" methods, we do not impose the assumption of non-joint production and dispense with the need for product-level input data. In contrast to the recent transformation function estimators, our approach maintains full input–output non-separability, a feature that is important for analyzing how output quality affects decisions about product mix and alters factor shares including labor share. The estimator remains compatible with any demand system that can be independently identified.

Applying the method to Taiwanese textile plants yields robust evidence of economies of scale, cost-saving scope economies, and cross-product cost complementarities, which imply that expansion of one output lowers the marginal cost of the other. Importantly, the data reject both input-output separability and non-joint production, indicating that the flexible multiproduct cost structure is not merely a theoretical refinement but is empirically important. The same change in factor-biased technology can affect the supply of cotton fiber and man-made fiber by different magnitudes, thereby generating variation in product mix across plants. At the same time, labor shares are shaped not only by factor-biased technology but also by output quality.

This model allows us to define a multilateral multiproduct-productivity index that measures the differences across plants and time in average variable cost after accounting for factor price differences and output differences under the condition of constant returns to scale. A decomposition of the index reveals that variations in input-augmenting efficiencies and output qualities, rather than capital deepening or scale economies, are the primary drivers of productivity dispersion. Plants with high input efficiencies have lower effective input prices, while plants with high-quality outputs have higher variable costs. Overall, the results show that plants with lower output quality and higher input efficiency have the highest productivity.

References

Acemoglu, D. (2002a): "Directed technical change," *The Review of Economic Studies*, 69, 781–809.

——— (2002b): "Technical change, inequality, and the labor market," *Journal of Economic Literature*, 40, 7–72.

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification properties of recent production function estimators," *Econometrica*, 83, 2411–2451.

Aw, B. Y. and Y. Lee (2025): "R&D investments, outsourcing and non-neutral productivity growth," *Journal of Productivity Analysis*, 63, 199–218.

Bailey, E. E. and A. F. Friedlaender (1982): "Market structure and multiproduct industries," *Journal of Economic Literature*, 20, 1024–1048.

Baumol, W. J., J. C. Panzar, and R. D. Willig (1982): *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich.

Berkowitz, D., H. Ma, and S. Nishioka (2017): "Recasting the iron rice bowl: the evolution of China's state owned enterprises," *The Review of Economics and Statistics*, 99, 735–747.

Binswanger, H. P. (1974): "The measurement of technical change biases with many factors of production," *American Economic Review*, 64 (6), 964–976.

Blundell, R. and S. Bond (2000): "Gmm estimation with persistent panel data: an application to production functions," *Econometric Reviews*, 19, 321–340.

Brown, R. S., D. W. Caves, and L. R. Christensen (1979): "Modelling the structure of cost and production for multiproduct firms," *Southern Economic Journal*, 46, 256–273.

Burgess, D. F. (1974): "A cost minimization approach to import demand equations," *The Review of Economics and Statistics*, 56, 225–234.

Cairncross, J., P. Morrow, S. Orr, and S. Rachapalli (2025): "Identifying firm vs. product markups using production data: micro estimates and aggregate implications," Working paper, UBC Sauder School of Business School.

Caselli, M., A. Chatterjee, and S. Li (2026): "Productivity and quality of multiproduct firms," Working paper, UNSW Sydney.

Caves, D. W., L. R. Christensen, and W. E. Diewert (1982): "Multilateral comparisons of output, input, and productivity using superlative index numbers," *The Economic Journal*, 92, 73–86.

Caves, D. W., L. R. Christensen, and J. A. Swanson (1980a): "Productivity in U.S. railroads, 1951–1974," *The Bell Journal of Economics*, 11, 166–181.

——— (1981a): "Economic performance in regulated and unregulated environments: a comparison of U.S. and Canadian railroads," *The Quarterly Journal of Economics*, 96, 559–581.

——— (1981b): "Productivity growth, scale economies, and capacity utilization in U.S. railroads, 1955–74," *American Economic Review*, 71, 994–1002.

Caves, D. W., L. R. Christensen, and M. W. Tretheway (1980b): "Flexible cost functions for multiproduct firms," *The Review of Economics and Statistics*, 62, 477–481.

——— (1984): "Economies of density versus economies of scale: why trunk and local service airline costs differ," *The RAND Journal of Economics*, 15, 471–489.

Chan, M. (2023): "How substitutable are labor and intermediates?" Working paper, University of Minnesota,.

Cowing, T. G. and A. G. Holtmann (1983): "Multiproduct short-run hospital cost functions: empirical evidence and policy implications from cross-section data," *Southern Economic Journal*, 49, 637–653.

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

De Loecker, J. and C. Syverson (2021): "An industrial organization perspective on productivity," in *Handbook of Industrial Organization*, ed. by K. Ho, A. Hortaçsu, and A. Lizzeri, North-Holland, vol. 4 of *Handbooks in Economics*, 141–223.

De Roux, N., M. Eslava, S. Franco, and E. Verhoogen (2024): "Estimating production functions in differentiated-product industries with quantity information and external instruments," NBER Working Paper 28323, National Bureau of Economic Research.

Demirer, M. (2025): "Production function estimation with factor-augmenting technology: an application to markups," Working paper, MIT Sloan School of Management.

Denny, M., M. A. Fuss, and L. Waverman (1981): "The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian telecommunications," in *Productivity Measurement in Regulated Industries*, ed. by T. G. Cowing and R. E. Stevenson, New York: Academic Press, 179–218.

Dhyne, E., A. Petrin, V. Smeets, and F. Warzynski (2024): "Theory for extending single-product production function estimation to multi-product settings," NBER Working Paper 30784, National Bureau of Economic Research.

Diewert, W. E. (1973): "Functional forms for profit and transformation functions," *Journal of Economic Theory*, 6, 284–316.

——— (1976): "Exact and superlative index numbers," *Journal of Econometrics*, 4, 115–145.

——— (1980): "Capital and the theory of productivity measurement," *American Economic Review*, 70, 260–267.

——— (1981): "The theory of total factor productivity measurement in regulated industries," in *Productivity Measurement in Regulated Industries*, ed. by T. G. Cowing and R. E. Stevenson, New York: Academic Press, 17–44.

Diewert, W. E. and K. J. Fox (2008): "On the estimation of returns to scale, technical progress and monopolistic markups," *Journal of Econometrics*, 145, 174–193.

Diewert, W. E., K. Nomura, and C. Shimizu (2025): "Estimating flexible functional forms using macroreconomic data," *Empirical Economics*, 69, 2671–2697.

Doraszelski, U. and J. Jaumandreu (2018): "Measuring the bias of technological change," *Journal of Political Economy*, 126, 1027–1084.

Evans, D. S. and J. J. Heckman (1984): "A test for subadditivity of the cost function with an application to the Bell System," *American Economic Review*, 74, 615–623.

Fuss, M. A. and L. Waverman (1981): "Regulation and the multiproduct firm: the case of telecommunications in Canada," in *Studies in Public Regulation*, The MIT Press, 277–328.

Gandhi, A., S. Navarro, and D. A. Rivers (2020): "On the identification of gross output production functions," *Journal of Political Economy*, 128, 2973–3016.

Gollop, F. M., B. Fraumeni, and D. Jorgenson (1987): *Productivity and U.S. Economic Growth*, Cambridge: Harvard University Press.

Gollop, F. M. and M. J. Roberts (1981): "The Sources of Growth in the U.S. Electric Power Industry," in *Productivity Measurement in Regulated Industries*, ed. by T. G. Cowing and R. E. Stevenson, New York: Academic Press, 107–143.

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

——— (2022): "Input prices, productivity and trade dynamics: long-run effects of liberalization on Chinese paint manufacturers," *The RAND Journal of Economics*, 53, 516–560.

Grieco, P. L. and R. C. McDevitt (2017): "Productivity and quality in health care: evidence from the dialysis industry," *The Review of Economic Studies*, 84, 1071–1105.

Griliches, Z. and J. Mairesse (1995): "Production functions: the search for identification," NBER Working Paper 5067, National Bureau of Economic Research.

Hall, R. E. (1973): "The specification of technology with several kinds of output," *Journal of Political Economy*, 81, 878–892.

Harrigan, J., A. Reshef, and F. Toubal (2024): "Techies, trade, and skill-biased productivity," CEPR Discussion Papers 15815.

Holmes, T. J. and J. J. Stevens (2014): "An alternative theory of the plant size distribution, with geography and intra- and international trade," *Journal of Political Economy*, 122, 369–421.

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

Jaumandreu, J. (2025): "Robust production function estimation when there is market power," Working paper, Boston Unniversity.

Jorgenson, D. W. and Z. Griliches (1967): "The explanation of productivity change," *The Review of Economic Studies*, 34, 249–283.

Khmelnitskaya, E., G. Marshall, and S. Orr (forthcoming): "Identifying scale and scope economies using product market data," *The RAND Journal of Economics*.

Koh, P. and D. Raval (2025): "Economies of scope from shared inputs," Working paper, Federal Trade Commission.

Lau, L. J. (1976): "A Characterization of the Normalized Restricted Profit Function," *Journal of Economic Theory*, 12, 131–163.

Levinsohn, J. and A. Petrin (2003): "Estimating production functions using inputs to control for unobservables," *The Review of Economic Studies*, 70, 317–341.

Li, S. and H. Zhang (2022): "Does external monitoring from the government improve the performance of state-owned enterprises?" *The Economic Journal*, 132, 675–708.

Maican, F. and M. Orth (2021): "Determinants of economies of scope in retail," *International Journal of Industrial Organization*, 75, 102710.

McFadden, D. L. (1978): "Cost, Revenue, and Profit Functions," in *Production Economics: A Dual Approach to Theory and Applications*, ed. by M. Fuss and D. L. McFadden, Amsterdam: North-Holland, vol. 1.

Mertens, M. and B. Schoefer (2025): "From labor to intermediates: firm growth, input substitution, and monopsony," NBER Working Paper 33172, National Bureau of Economic Research.

Morrison, C. J. (1992): "Unraveling the productivity growth slowdown in the United States, Canada and Japan: the effects of subequilibrium, scale economies and markups," *The Review of Economics and Statistics*, 74, 381–393.

Nelson, J. P., M. J. Roberts, and E. P. Tromp (1987): "An analysis of Ramsey pricing in electric utilities," in *Regulating Utilities in an Era of Deregulation*, Springer, 85–109.

Olley, G. S. and A. Pakes (1996): "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 64, 1263–1297.

Orr, S. (2022): "Within-firm productivity dispersion: estimates and implications," *Journal of Political Economy*, 130, 2771–2828.

Raval, D. R. (2019): "The micro elasticity of substitution and non-neutral technology," *The RAND Journal of Economics*, 50, 147–167.

Rubens, M., Y. Wu, and M. Xu (2026): "Exploiting or augmenting labor," *American Economic Review: Insights*, 8, 72–89.

Spady, R. H. and A. F. Friedlaender (1978): "Hedonic cost functions for the regulated trucking industry," *The Bell Journal of Economics*, 9, 159–179.

Valmari, N. (2023): "Estimating production functions of multiproduct firms," *The Review of Economic Studies*, 90, 3315–3342.

Zhang, H. (2019): "Non-neutral technology, firm heterogeneity, and labor demand," *Journal of Development Economics*, 140, 145–168.

Zhao, S., E. Malikov, and S. Kumbhakar (2025): "On the estimation of nonneutral production technologies without separability," Working paper, Oakland University.

Online Appendix

A Alternative Estimator: Dynamic Panel Approach

In the paper, our estimation approach is based on constructing moment conditions that exploit a timing assumption on the structural errors: the innovation shocks to the structural errors are orthogonal to lagged levels of the relevant variables. An alternative implementation involves using a dynamic panel estimator (i.e., Blundell and Bond, 2000), which assumes an auto-regressive process for the structural errors. This section describes the implementation of a dynamic panel estimator for our model.

We start from the system of equations (24), which can be equivalently written as:

$$\underbrace{\begin{bmatrix} q_{1jt} \\ q_{2jt} \\ -w_{Ljt} \end{bmatrix}}_{\boldsymbol{y}_{jt}} = \Omega^{-1} \left\{ \underbrace{\begin{bmatrix} \frac{\eta_1-1}{\eta_1} R_{1jt}/C_{jt} \\ \frac{\eta_2-1}{\eta_2} R_{2jt}/C_{jt} \\ E_{Ljt}/C_{jt} \end{bmatrix}}_{\boldsymbol{r}_{jt}} - \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \alpha_L \end{bmatrix}}_{\boldsymbol{d}} - \underbrace{\begin{bmatrix} \phi_{K1} \\ \phi_{K2} \\ \delta_{KL} \end{bmatrix}}_{\boldsymbol{\psi}} k_{jt} \right\} - \underbrace{\begin{bmatrix} \nu_{1jt} \\ \nu_{2jt} \\ \tilde{\mu}_{Ljt} \end{bmatrix}}_{\boldsymbol{\varepsilon}_{jt}}, \tag{A.1}$$

where Ω is defined as (22). Let $\boldsymbol{\theta}_1$ collect the relevant cost function parameters in Ω and $\boldsymbol{d}$; we can write the vector of structural errors as a function of observables and parameters:

$$\boldsymbol{\varepsilon}_{jt}(\boldsymbol{\theta}_1) \equiv \begin{bmatrix} \nu_{1jt} \\ \nu_{2jt} \\ \tilde{\mu}_{Ljt} \end{bmatrix} = \Omega^{-1} \left(\boldsymbol{r}_{jt} - \boldsymbol{d} - \boldsymbol{\psi}\, k_{jt} \right) - \boldsymbol{y}_{jt}. \tag{A.2}$$

All regressors in $\boldsymbol{r}_{jt}$ and k_{jt} are allowed to be correlated with the components of $\boldsymbol{\varepsilon}_{jt}$, as detailed below; identification will rely on internal lagged instruments.

As is typical in the dynamic panel context, we assume a three-component structure for each equation $h \in \{1, 2, L\}$ (corresponding to $\nu_1, \nu_2, \tilde{\mu}_L$):

$$\varepsilon_{jt}^{(h)} = \iota_j^{(h)} + \sigma_{jt}^{(h)} + e_{jt}^{(h)}, \qquad \sigma_{jt}^{(h)} = b_h\, \sigma_{j,t-1}^{(h)} + u_{jt}^{(h)}, \tag{A.3}$$

where we have re-defined the variables in order to make the notation more compact: $\varepsilon_{jt}^{(1)} \equiv \nu_{1jt}$, $\varepsilon_{jt}^{(2)} \equiv \nu_{2jt}$, and $\varepsilon_{jt}^{(L)} \equiv \tilde{\mu}_{Ljt}$. Note that $\iota_j^{(h)}$ is an equation-specific plant fixed effect that is unrestrictedly correlated with observables; $e_{jt}^{(h)}$ is a serially uncorrelated innovation with zero mean and finite variance; $u_{jt}^{(h)}$ is a serially uncorrelated innovation with zero mean and finite variance, independent of $e_{j\tau}^{(h)}$ for all τ; b_h is a persistence parameter. These assumptions nest the evolution processes, (18), (19), and (20) in Section 4, by setting $\iota_j^{(h)} = 0$ and $e_{jt}^{(h)} = 0$, so that $\varepsilon_{jt}^{(h)} = \sigma_{jt}^{(h)}$, for all j, t, and h.

Collect these terms in vector form:

$$\boldsymbol{\varepsilon}_{jt} = \boldsymbol{\iota}_j + \boldsymbol{\sigma}_{jt} + \boldsymbol{e}_{jt}, \qquad \boldsymbol{\sigma}_{jt} = B\,\boldsymbol{\sigma}_{j,t-1} + \boldsymbol{u}_{jt}, \qquad B = \mathrm{diag}(b_1, b_2, b_L).$$

Because $\boldsymbol{\varepsilon}_{jt}(\boldsymbol{\theta}_1)$ in (A.2) is observed up to parameters, we can form transformations that remove fixed effects and isolate innovation terms. Define the AR-filtered residuals as:

$$\boldsymbol{D}_{jt}(\boldsymbol{\theta}_1, \boldsymbol{b}) \;\equiv\; \boldsymbol{\varepsilon}_{jt}(\boldsymbol{\theta}_1) \;-\; B\,\boldsymbol{\varepsilon}_{j,t-1}(\boldsymbol{\theta}_1), \qquad \boldsymbol{b} \equiv (b_1, b_2, b_L)'. \tag{A.4}$$

Using (A.3), the fixed effects $\boldsymbol{\iota}_j$ difference out and we obtain

$$\boldsymbol{D}_{jt} = \big(\boldsymbol{u}_{jt} + \boldsymbol{e}_{jt}\big) \;-\; B\,\boldsymbol{e}_{j,t-1}.$$

Hence, $\boldsymbol{D}_{jt}$ contains only current AR innovations and current/lagged i.i.d. shocks.

To further purge serial correlation induced by $\boldsymbol{e}_{jt}$, use adjacent differences of $\boldsymbol{D}_{jt}$:

$$\begin{aligned} \Delta\boldsymbol{D}_{jt}(\boldsymbol{\theta}_1, \boldsymbol{b}) \;&\equiv\; \boldsymbol{D}_{jt}(\boldsymbol{\theta}_1, \boldsymbol{b}) - \boldsymbol{D}_{j,t-1}(\boldsymbol{\theta}_1, \boldsymbol{b}) \\ &=\; \big(\boldsymbol{u}_{jt} - \boldsymbol{u}_{j,t-1}\big) + \big(\boldsymbol{e}_{jt} - (I+B)\boldsymbol{e}_{j,t-1} + B\,\boldsymbol{e}_{j,t-2}\big). \end{aligned} \tag{A.5}$$

The difference GMM moments are:

$$\mathbb{E}\Big[\boldsymbol{z}_{j,t-3}\,\Delta\boldsymbol{D}_{jt}(\boldsymbol{\theta}_1, \boldsymbol{b})\Big] \;=\; \boldsymbol{0}, \qquad t = 4, \ldots, T, \tag{A.6}$$

where $\boldsymbol{z}_{jt} = (q_{1jt}, q_{2jt}, w_{Ljt}, k_{jt}, R_{1jt}/C_{jt}, R_{2jt}/C_{jt}, E_{Ljt}/C_{jt})$. Moment equation (A.6) is the multivariate analog of the Arellano–Bond difference moments.

The level GMM moments are:

$$\mathbb{E}\Big[\big(\boldsymbol{z}_{j,t-2} - \boldsymbol{z}_{j,t-3}\big)\,\boldsymbol{D}_{jt}(\boldsymbol{\theta}_1, \boldsymbol{b})\Big] \;=\; \boldsymbol{0}, \qquad t = 3, \ldots, T. \tag{A.7}$$

These are the system-GMM level moments, augmenting (A.6).

Let $g^{(d)}_{j,t}(\boldsymbol{\theta}_1, \boldsymbol{b})$ collect the stacked difference moments in (A.6) and $g^{(\ell)}_{j,t}(\boldsymbol{\theta}_1, \boldsymbol{b})$ the level moments in (A.7). Stack over j and admissible t to form:

$$\boldsymbol{g}(\boldsymbol{\theta}_1, \boldsymbol{b}) \;=\; \frac{1}{J}\sum_{j=1}^{J}\left(\sum_t g^{(d)}_{j,t}(\boldsymbol{\theta}_1, \boldsymbol{b}) \;\oplus\; \sum_t g^{(\ell)}_{j,t}(\boldsymbol{\theta}_1, \boldsymbol{b})\right),$$

and the GMM estimator is given by

$$(\boldsymbol{\theta}_1, \boldsymbol{b}) = \mathrm{argmin}\;\; \boldsymbol{g}'\,W\,\boldsymbol{g}, \tag{A.8}$$

where W is a consistent weighting matrix (identity in one-step; optimal two-step weighting based on first-step residuals).

Similar to the approach described in Section 4.2, this step of the estimation delivers: cost function parameters other than α_K and δ_{KK}; persistence parameters $(\hat{b}_1, \hat{b}_2, \hat{b}_L)$; estimated structural errors at the plant–time level, i.e., $\{\hat{\nu}_{1jt}, \hat{\nu}_{2jt}, \widehat{\widetilde{\mu}}_{Ljt}\}$.

The second step, following the approach in Section 4.2, is to use the cost function to estimate μ_{Mjt} and the remaining cost function parameters (i.e., α_K and δ_{KK}).

Specifically, we rewrite (29) as

$$c_{jt} - w_{Mjt} - \hat{c}_{jt} = \alpha_0 + \alpha_K k_{jt} + \frac{1}{2}\delta_{KK} k_{jt}^2 - \mu_{Mjt}, \tag{A.9}$$

where the heterogeneity in the material efficiency term is the structural error in (29) and the only unknown parameters are $\alpha_0, \alpha_K, \delta_{KK}$.

Rearranging (A.9), the structural error of interest is

$$\varepsilon_{jt}(\boldsymbol{\theta}_2) \;\equiv\; \mu_{Mjt} = \alpha_0 + \alpha_K k_{jt} + \tfrac{1}{2}\delta_{KK} k_{jt}^2 - (c_{jt} - w_{Mjt} - \hat{c}_{jt}), \tag{A.10}$$

where $\boldsymbol{\theta}_2 = (\alpha_0, \alpha_K, \delta_{KK})$.

Similar to the assumption on the other structural error terms, we assume that μ_{Mjt} has three components:

$$\mu_{Mjt} \;=\; \iota_j^{(M)} + \sigma_{jt}^{(M)} + e_{jt}^{(M)}, \qquad \sigma_{jt}^{(M)} \;=\; b_M\, \sigma_{j,t-1}^{(M)} + u_{jt}^{(M)}, \tag{A.11}$$

where $\iota_j^{(M)}$ is a plant fixed effect (materials equation), $e_{jt}^{(M)}$ and $u_{jt}^{(M)}$ are serially uncorrelated, mean-zero innovations with finite variances, and $u_{jt}^{(M)}$ is independent of $e_{j\tau}^{(M)}$ for all τ.

Define the residual

$$D_{jt}(\boldsymbol{\theta}_2, b_M) \;\equiv\; \varepsilon_{jt}(\boldsymbol{\theta}_2) - b_M\, \varepsilon_{j,t-1}(\boldsymbol{\theta}_2). \tag{A.12}$$

Take adjacent differences to obtain:

$$\Delta D_{jt}(\boldsymbol{\theta}_2, b_M) \;\equiv\; D_{jt}(\boldsymbol{\theta}_2, b_M) - D_{j,t-1}(\boldsymbol{\theta}_2, b_M). \tag{A.13}$$

The difference moments are defined as:

$$\mathbb{E}[\, z_{j,t-3}\, \Delta D_{jt}(\boldsymbol{\theta}_2, b_M)\,] \;=\; 0, \qquad t = 4, \ldots, T, \tag{A.14}$$

where $z_{j,t-3} = (c_{j,t-3}, k_{j,t-3}, k_{j,t-3}^2)$.

The level moments are defined as:

$$\mathbb{E}[\, (z_{j,t-2} - z_{j,t-3})\, D_{jt}(\boldsymbol{\theta}_2, b_M)\,] \;=\; 0, \qquad t = 3, \ldots, T. \tag{A.15}$$

Combining (A.14) and (A.15) yields a system-GMM estimator, similar to (A.8). This step

delivers the remaining cost function parameters $(\hat{\alpha}_0, \hat{\alpha}_K, \widehat{\delta}_{KK})$, the persistence of μ_{Mjt}, and the material efficiency term $\widehat{\mu}_{Mjt}$.

B Monte Carlo Experiments

This section reports Monte Carlo experiments designed to evaluate the finite-sample performance of the two estimation methods for our multiproduct cost model. The first implementation (called the main implementation) is the two-step GMM procedure described in the main text. It exploits the timing restrictions implied by the Markov evolution of the structural errors: innovation shocks to efficiency/quality are orthogonal to instruments expressed in lagged levels of observables. The second implementation (called the dynamic panel estimator) is the alternative based on (i.e., Blundell and Bond, 2000) as described in Online Appendix A. This imposes an autoregressive structure on the structural errors and uses internal lag instruments constructed from the panel structure. In this implementation, we use both the difference moments and the level moments (system GMM).

B.1 Data-generating process

We focus on the two-product case ($N = 2$) used in the paper. In each Monte Carlo replication, we take the variables that enter the cost function as *fixed design variables*: $(w_{Ljt}, w_{Mjt}, k_{jt}, q_{1jt}, q_{2jt})$, and simulate the unobserved efficiency/quality terms that shift the system. Specifically, for each plant j and each year t, we simulate $\left(\nu_{1jt}, \nu_{2jt}, \tilde{\mu}_{Ljt}, \mu_{Mjt}\right)$ according to the AR(1) processes in (18), (19), and (20). The innovation shocks are i.i.d. Gaussian with mean zero and common standard deviation σ_ξ (reported in the table notes).

Given simulated efficiency/quality terms, we construct the key observable ratios used in estimation. First, the labor cost share is generated from (15), with $\mu_{Kjt} = 0$. Second, the markup-adjusted revenue-to-cost ratios are generated from (16), with fixed demand elasticity parameter values (η_1, η_2) (reported in the table notes).

We then construct variable cost from the normalized cost identity (17) and obtain the simulated log variable cost $c_{jt} = \tilde{c}_{jt} + \left(w_{Mjt} - \mu_{Mjt}\right)$, where $\tilde{c}_{jt}$ is the normalized translog cost (defined in (10)) evaluated at the quality-adjusted quantities and efficiency-adjusted relative price. After converting to levels, $C_{jt} = \exp(c_{jt})$, we recover expenditures and revenues using the simulated ratios: $E_{Ljt} = \left(\frac{E_{Ljt}}{C_{jt}}\right) C_{jt}$, $E_{Mjt} = \left(1 - \frac{E_{Ljt}}{C_{jt}}\right) C_{jt}$ and $R_{njt} = \frac{\eta_n}{\eta_n - 1}\left(\frac{\eta_n - 1}{\eta_n}\frac{R_{njt}}{C_{jt}}\right) C_{jt}, n \in \{1, 2\}$. We compute implied output prices as revenue per unit: $P_{njt} = \frac{R_{njt}}{Q_{njt}}$, where $Q_{njt} = \exp(q_{njt})$ and hence $p_{njt} = \log R_{njt} - q_{njt}$.

Each replication therefore produces a complete simulated dataset containing

$$(w_{Ljt}, w_{Mjt}, k_{jt}, q_{1jt}, q_{2jt}, c_{jt}, E_{Ljt}, E_{Mjt}, R_{1jt}, R_{2jt}),$$

which can be used to compute the corresponding ratios used in estimation. We impose feasibility checks to ensure economically meaningful simulated outcomes. In particular, we require $0 < E_{Ljt}/C_{jt} < 1$ (equivalently $E_{Ljt} > 0$ and $E_{Mjt} > 0$) and $R_{njt} > 0$ for $n = 1, 2$.

Replications that violate these conditions are discarded.[33]

B.2 Experimental designs

We consider two experimental designs that differ in the size of the panel while holding the underlying data-generating process fixed.

Small N and T (empirically sized panel). In the first design, we use the empirical panel size and the empirical values of the fixed-design variables $(w_{Ljt}, w_{Mjt}, k_{jt}, q_{1jt}, q_{2jt})$ from the dataset reported in Section 5. This design therefore evaluates the estimators under a data environment that closely matches the empirical application, including the unbalanced-panel structure. We implement both estimators: the main implementation and the dynamic panel estimator. For the dynamic panel estimator, we use internal instruments constructed from second-order lagged observations in accordance with the setup of the AR(1) processes.

Large N and T (large panel). In the second design, we increase the panel size while keeping the dispersion of the fixed-design variables comparable to the empirical setting. Concretely, we generate $(w_{Ljt}, w_{Mjt}, k_{jt}, q_{1jt}, q_{2jt})$ so that their means and variances match those in the application sample reported in Section 5, and then apply the same simulation procedure for the efficiency/quality terms and the implied outcomes. This design isolates how estimator performance changes as the panel becomes large.

For both designs, we run 300 Monte Carlo replications and summarize performance using bias and Monte Carlo standard deviation (SD). The true parameter values used in each experiment are reported in the "True" column of Table A1. The bias is reported as the estimated mean value minus the true value.

B.3 Results and discussion

Table A1 reports the Monte Carlo bias and SD for both implementations under the two panel sizes. Three patterns stand out.

First, the main implementation performs well in both designs. Bias is negligible for essentially all cost-function parameters, and dispersion is modest even in the empirically sized panel. As expected, precision improves further in the large-panel design, reflecting the stronger concentration of the sample moments around their population values.

Second, the dynamic panel estimator performs well in the large-panel design but is markedly less stable in the empirically sized panel. In particular, several parameters exhibit substantially larger standard deviations under small N and T, while the same parameters are tightly estimated once the panel becomes large. This pattern is consistent with well-known finite-sample challenges of system-GMM estimators in short panels: internal instruments based on lagged variables can be weak when the effective time dimension is limited, and the

[33] In our baseline implementation, discarding occurs rarely for the parameterizations considered.

precision of the resulting dynamic-panel moments can deteriorate sharply in small samples (e.g., Blundell and Bond, 2000).

Third, for a given panel size, the main implementation is typically more precise than the dynamic panel estimator, as indicated by smaller standard deviations across most parameters. Intuitively, the main implementation exploits the timing assumption and applies moment conditions directly to the innovation terms and first-order lagged variables. In contrast, the dynamic panel estimator relies on transformations (differences and AR-filtering) and internal lag instruments; these operations generally discard between-plant variation and can amplify sampling noise, reducing finite-sample efficiency.

Overall, the Monte Carlo evidence shows that the main implementation is accurate and stable in panels of the size encountered in our application, and it becomes even more precise as the panel becomes larger. The dynamic panel estimator provides a useful alternative implementation, but it requires substantially larger samples to achieve the same degree of stability and precision as the main implementation in empirically sized panels.

Table A1: Monte Carlo performance of two implementations

		Small N and T				Large N and T			
Parameter	True	Main implementation		Dynamic panel estimator		Main implementation		Dynamic panel estimator	
		Bias	SD	Bias	SD	Bias	SD	Bias	SD
α_L	0.500	0.000	(0.002)	-0.067	(0.659)	-0.000	(0.000)	-0.001	(0.047)
α_K	-0.050	0.003	(0.032)	0.035	(0.493)	0.000	(0.004)	0.002	(0.049)
β_1	0.700	-0.000	(0.003)	-0.023	(1.168)	0.000	(0.000)	0.005	(0.070)
β_2	0.600	0.000	(0.004)	0.061	(0.615)	-0.000	(0.000)	-0.008	(0.069)
δ_{LL}	-0.050	-0.003	(0.045)	-0.015	(0.024)	0.000	(0.001)	0.000	(0.002)
δ_{KL}	-0.050	-0.000	(0.004)	0.003	(0.013)	0.000	(0.000)	0.000	(0.000)
δ_{KK}	-0.050	0.001	(0.036)	0.018	(0.221)	0.000	(0.003)	0.001	(0.005)
γ_{11}	0.100	0.000	(0.004)	0.003	(0.010)	0.000	(0.000)	0.000	(0.001)
γ_{12}	-0.050	-0.000	(0.005)	-0.005	(0.013)	-0.000	(0.000)	-0.000	(0.000)
γ_{22}	0.100	-0.000	(0.012)	0.014	(0.021)	0.000	(0.001)	-0.000	(0.001)
ϕ_{L1}	0.050	0.002	(0.031)	0.008	(0.017)	-0.000	(0.001)	-0.000	(0.001)
ϕ_{L2}	0.050	0.005	(0.038)	0.010	(0.019)	0.000	(0.001)	0.000	(0.001)
ϕ_{K1}	-0.050	-0.000	(0.005)	-0.002	(0.017)	-0.000	(0.000)	0.000	(0.001)
ϕ_{K2}	-0.050	-0.001	(0.007)	-0.001	(0.016)	0.000	(0.000)	0.000	(0.001)
g_1	0.500	-0.019	(0.060)	-0.001	(0.199)	-0.000	(0.009)	0.000	(0.018)
g_2	0.500	-0.018	(0.068)	-0.001	(0.163)	0.000	(0.010)	0.000	(0.019)
g_L	0.500	-0.022	(0.067)	-0.040	(0.205)	-0.001	(0.009)	-0.001	(0.019)
g_M	0.500	-0.010	(0.064)	0.047	(0.169)	-0.001	(0.009)	-0.002	(0.019)

[1] The data-generating processes are the same across different cases and implementations. The standard deviation of the innovation shocks in the AR(1) evolution processes of the quality and efficiency terms is $\sigma_\xi = 0.2$. We use fixed demand elasticity parameter values $(\eta_1, \eta_2) = (3, 4)$.
[2] The small N and T case has a data structure similar to the data (of two-product producers) reported in Section 5: $N = 38$ and $T = 13$. The high N and T case has $N = 250$ and $T = 40$. We conduct 300 simulations in each case and report Monte Carlo standard deviations of the parameter estimates in parentheses.