Productivity and Quality of Multi-product Firms*

Mauro Caselli†, University of Trento
Arpita Chatterjee‡, Federal Reserve Board
Shengyu Li§, University of New South Wales

January 4, 2026

Abstract

This paper introduces a method for estimating productivity and quality at the firm-product level using a transformation function framework. Using firms' optimization conditions, we obtain a one-to-one mapping from observed data to unobserved productivity and quality without imputing inputs. The method scales to many products, avoids specifying productivity dynamics, corrects bias from unobserved heterogeneous input prices, and accommodates productivity–quality trade-off. Applying it to Mexican manufacturing industries, we show that an exogenous, product-specific technological improvement raises firm-level productivity mainly through within-firm reallocation and delivers substantial industry-level welfare gains, further amplified by across-firm and within-firm spillovers.

Keywords: *multi-product firms, productivity, quality, spillover, within-firm reallocation*

JEL classification: *D24, L11, L15, O33.*

*The authors thank Eleni Aristodemou, Zhiyuan Chen, Jan De Loecker, Erwin Diewert, Kevin Fox, Andrea Fracasso, Maurice Kugler, Moyu Liao, Logan Lewis, Matthias Mertens, Scott Orr, Devesh Raval, Ariell Reshef, Mark Roberts, Stefano Schiavo, Petr Sedlacek, Nikos Theodoropoulos, Andreas Tryphonides, Nelli Valmari, Eric Verhoogen, Frederic Warzynski, Daniel Xu, Haiqing Xu, Hongsong Zhang, and many seminar and conference participants for very helpful comments. All errors are the authors' responsibility. All views/opinions are authors' own and do not reflect views of the Federal Reserve Board or the Federal Reserve System.

†School of International Studies & Department of Economics and Management, University of Trento. Email: mauro.caselli@unitn.it.

‡Federal Reserve Board. Email: chatterjee.econ@gmail.com.

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

1 Introduction

The production landscape of many manufacturing industries is dominated by multi-product firms, which operate across a diverse range of product lines. However, existing empirical studies that explore the determinants of firm performance have primarily focused on analyzing variations across different firms, such as heterogeneity in productivity levels and demand characteristics (e.g., Foster et al., 2008; Pozzi and Schivardi, 2016; Kumar and Zhang, 2019). Consequently, there remains a considerable gap in the understanding of the factors that drive within-firm heterogeneity and resource reallocation, as well as their subsequent impact on firm growth. This knowledge gap is due to methodological limitations and data constraints, which hinder the accurate estimation of heterogeneity at the firm-product level.

This paper introduces a method to estimate productivity and quality (product appeal) at the firm-product level, along with the transformation function and demand parameters. This method constructs a unique one-to-one mapping from observed data to unobservable variables by using firm optimization conditions. This offers several advantages over recent methods (e.g., Dhyne et al., 2022; Orr, 2022; Valmari, 2023). First, it eliminates the need for imputing within-firm input allocations. Second, it does not need to impose restrictions on productivity evolution, allowing for flexibility in exploring complex productivity dynamics after estimation. Third, it is scalable to handle a large number of products. Fourth, it addresses the estimation bias caused by heterogeneous firm-level intermediate input prices, which are usually unobservable in available data sets.[1] To demonstrate the advantages, we apply our method to three major industries in the Mexican manufacturing sector, where multi-product production is a central feature of firms. We examine the role of both across-firm and within-firm technological spillovers in the dynamic evolution of technical efficiency, as well as the role of within-firm resource reallocation in shaping firm performance.

In modeling the production side, our method is designed to address the challenges commonly faced in estimating multi-product production functions. Most production function estimation methodologies implicitly assume that each firm produces a single product (e.g., Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Gandhi et al., 2020). In this context, the input allocation is observable to researchers and each firm only has a single dimension of unobservable productivity, which can be controlled for by an observable proxy. Multi-product firms, on the contrary, may have different levels of productivity for each product. Extending the proxy-based methods to the context of multi-product firms requires at least the same number of proxies as the number of products (cf., Dhyne et al., 2022).

[1] In the cases where intermediate input prices are observed, our method can be modified to allow for non-Hicks' neutral efficiency (i.e., labor-augmenting efficiency), as shown in recent literature (e.g., Doraszelski and Jaumandreu, 2016; Zhang, 2019; Raval, 2019; Rubens et al., 2024).

Moreover, researchers do not observe the within-firm division of inputs used to produce different products because firms usually only report total inputs at the firm level.[2] Finally, intermediate input prices, which vary significantly across firms and over time due to various reasons such as bargaining power in the input market and transport costs, as documented by Atalay (2014), should be controlled for to avoid "input price bias" (Ornaghi, 2006; De Loecker et al., 2016; Grieco et al., 2016). However, these firm-level input prices are rarely observable.

To address these issues, we model the production technology using a transformation function, which is a mapping from a vector of inputs at the firm level to an aggregator of product-specific outputs. This saves us from modeling how the inputs are divided for the production of each individual product. Each product is associated with a potentially different level of physical productivity (i.e., quantity-based productivity, or TFPQ, as in Foster et al., 2008).[3] The productivity levels, together with a parameter in the transformation function that characterizes the technological substitutability of the products, govern the marginal rate of transformation between any two products. The firm observes these productivity levels before making input and output decisions to maximize profits. In the spirit of Grieco et al. (2016), we show that the optimization conditions implied from our model can be inverted to form an explicit one-to-one mapping from observed input and output decisions to unobserved productivity at the firm-product level (regardless of the number of products), while controlling for unobserved intermediate input prices. Intuitively, the variation in product prices within a firm identifies the productivity difference across products within the firm, after controlling for differences in markups and production scale. We exploit the inverted relationship to replace unobserved productivity in the transformation function, enabling estimation of the transformation function parameters. Once the parameters are estimated, we compute productivity (TFPQ) at the firm-product level from the one-to-one mapping.

Although the primary innovation of our method lies on the production side, it is flexible enough to accommodate a variety of demand systems.[4] Conditional on the availability of valid instrumental variables, the approach can be applied to widely used demand models such as Constant Elasticity of Substitution (CES) demand, discrete-choice demand (e.g., Berry, 1994), and random-coefficients logit demand (e.g., Berry et al., 1995). In our empirical

[2]This is an empirical challenge because of the potential input sharing (e.g., machinery and workers) across product lines within a firm (e.g., Cairncross et al., 2025; Koh and Raval, 2025). For example, a printing firm may use the same design software to create multiple products, such as product labels; workers with specialized skills, such as pattern makers, may be used across different product lines within the same footwear firm; in pharmaceutical industries, a firm may use the same reactors to produce different products by adjusting the process parameters.

[3]We refer to physical productivity as simply "productivity" in this paper unless explicitly stated otherwise.

[4]Specifying and estimating a demand system using product-level quantity and price data is essential if the purpose is to identify firm–product-level productivity separately from firm–product-level markups, as shown by Cairncross et al. (2025).

application, we adopt a CES demand specification, which is appropriate given the level of product aggregation in our data. To address the classic endogeneity issue in estimating the price elasticity of demand, we depart from the traditional literature by exploiting a key feature of multi-product firms: within-firm profit maximization implies a structural relationship between the revenues of products produced by the same firm. Following estimation, we recover product quality as the residual component of the demand function after controlling for price.

After demonstrating the performance of our method through Monte Carlo simulations, we apply it to establishment-level panel data from three major Mexican manufacturing industries—footwear, printing, and pharmaceuticals—that include firm-product-level prices and quantities, along with detailed firm-level input data. Multi-product firms represent approximately 56% of all firms and account for 86% of total revenues in these industries. Given the product classification used, the number of total products ranges from 4 in the footwear industry to 16 in the pharmaceutical industry, with multi-product firms producing an average of 6.9 products per year. Within each industry, the markets for different product categories (e.g., women's shoes vs. men's shoes in the footwear industry) are largely segmented. However, within each product category, firms' outputs are likely vertically differentiated, as reflected in the substantial dispersion in prices. These empirical features support our use of a CES demand model, which abstracts from competition across horizontally differentiated product markets while capturing vertical differentiation through quality differences.

After estimation, the recovered TFPQ and product quality at the firm-product level allow us to examine heterogeneity and performance both within and across firms. Following the literature (e.g., Melitz, 2000), we construct a revenue-based productivity measure (TFPR) that incorporates heterogeneity in both TFPQ and quality at the firm-product level. We find substantial variation in TFPR, with heterogeneity across firms dominating that within firms.

Interestingly, although our estimation does not impose any relationship between TFPQ and quality, we find a significant negative correlation (i.e., trade-off) between them, with a coefficient of -0.34. This implies that producing higher quality comes at the cost of lower TFPQ when inputs are held fixed.[5] We refer to the component of TFPQ that is adjusted for the cost of quality as technical efficiency. Unlike raw TFPQ, this measure is comparable across firms and products because it accounts for variation in quality.

A further advantage of our method is that it does not necessarily require any ex-ante assumptions about the dynamic evolution of technical efficiency. This feature allows us to

[5]This result is broadly consistent with the emerging literature emphasizing the negative correlation between physical productivity and quality across firms (e.g., Grieco and McDevitt, 2017; Roberts et al., 2018; Orr, 2022; Eslava et al., 2024; Forlani et al., 2023; Li et al., 2025).

investigate complex interdependencies in productivity dynamics within multi-product firms after estimating the model parameters, a task that would be considerably more difficult if the productivity process had to be estimated jointly with other parameters. To demonstrate this advantage, we study technological spillovers—both across firms and within firms—while allowing for the trade-off between TFPQ and quality. Compared to the existing literature, which typically focuses on across-firm spillovers (e.g., Malikov and Zhao, 2023), our results suggest that within-firm spillovers are also economically meaningful, although across-firm spillovers are indeed more prominent. To quantify the importance of these spillover channels, we conduct a counterfactual exercise where the technical efficiency of one product is improved exogenously. Compared with the benchmark without spillover, the across-firm spillover contributes an additional 16.6% to the total welfare gain, while the within-firm spillover contributes an extra 5.4%. More importantly, over half of the improvement in firm-level TFPR resulting from the product-specific shock is attributable to within-firm resource reallocation toward more productive products—regardless of spillover types.

Our methodology builds on recent advances in the estimation of heterogeneous productivity of multi-product firms. In addressing the common data challenge of input data being observable only at the firm level, while outputs and revenues are reported separately by product, the literature has evolved into two main approaches. The first approach, pioneered by De Loecker et al. (2016), characterizes multi-product production as a collection of single-product production functions, coupled with a rule for allocating firm inputs to each of these functions. Subsequent studies have extended this approach. In particular, Orr (2022) models product lines sharing the same technology (i.e., production parameters) but with individual productivity, and shows how demand data can be used to assist estimation under profit maximization conditions. Valmari (2023) develops a similar framework, incorporating flexible production parameters across product-specific production functions. In contrast, the second approach, led by Dhyne et al. (2022), departs from the assumption that multi-product production is a collection of single-product firms. They adopt a transformation function and show how it can be used to recover the production frontier and estimate firm-product-specific marginal costs.

We integrate the strengths of both approaches to overcome their respective limitations. First, we model multi-product production using a transformation function, similar to Dhyne et al. (2022). This avoids the need to allocate firm-level inputs, as in Orr (2022) and Valmari (2023), and allows for potential within-firm input sharing across product lines. Second, in addressing unobserved firm-product productivity, we adopt the profit maximization assumption, similar to Orr (2022) and Valmari (2023). However, instead of imputing input allocation shares, we use the profit-maximizing conditions to establish a one-to-one mapping

from observed firm decisions to unobserved productivity, extending the insights of Grieco et al. (2016, 2022), Harrigan et al. (2021) and Li and Zhang (2022) to the context of multi-product firms. Importantly, the number of profit-maximizing conditions, which naturally increase with the number of products, ensures the scalability of our method. This differs from Dhyne et al. (2022), whose method requires a separate proxy for each additional firm-product-level productivity. Rather, it is more similar to recent approaches to identify markdowns (Morlacco, 2020; Caselli et al., 2021; Kirov and Traina, 2023) or factor-augmenting productivity (Demirer, 2022; Raval, 2023) using necessary conditions for optimality with respect to multiple flexible inputs. Third, our method addresses the bias due to unobserved firm-level heterogeneity in input prices without requiring the availability of input price data. This is in contrast to the existing methods (e.g., Orr, 2022; Valmari, 2023), which typically require access to such data. Finally, our method does not rely on modeling the evolution of productivity, which offers a distinct advantage in exploring the evolution of productivity after estimation. Such an advantage is particularly beneficial in studying complex (e.g., interdependent) productivity dynamics, factors that endogenously shape the productivity trajectory (e.g., Chen et al., 2021; Malikov and Zhao, 2023), and frequent product turnovers, such as for exported products.

In terms of empirical application, our paper integrates the analysis of the productivity–quality trade-off, technological spillovers, and resource reallocation in the context of multi-product firms. Focusing on firm-level analysis, Grieco and McDevitt (2017) and Li et al. (2025) have documented a significant trade-off between productivity and quality—interpreted as the cost of quality—in the U.S. healthcare sector and the Chinese steel industry, respectively. A natural implication of their findings is that the cost of quality should be explicitly considered when modeling the evolution of productivity. Our paper identifies a similar trade-off at the firm-product level and incorporates this feature into the productivity evolution process to investigate technological spillovers. On the spillover front, our study complements the firm-level literature on productivity spillovers (e.g., Malikov and Zhao, 2023). While most existing research focuses on spillovers across firms, we demonstrate that within-firm spillovers can also be economically significant. These within-firm productivity spillovers reflect economies of scope arising from the internal sharing of knowledge (e.g., Bilir and Morales, 2020; Merlevede and Theodorakopoulos, 2023; Ding, 2025). We show that such spillovers substantially enhance both firm performance and aggregate welfare, primarily through within-firm resource reallocation, a mechanism increasingly recognized in the recent literature on multi-product firms (e.g., Mayer et al., 2021). Our evidence highlights within-firm reallocation as a novel and important channel through which firm-level productivity responds to product-specific shocks. In doing so, our study complements a large body of work emphasizing the role of across-firm resource reallocation in driving aggregate

productivity growth (e.g., Aw et al., 2001; Foster et al., 2008; Syverson, 2011; Collard-Wexler and De Loecker, 2015).

The remainder of the paper is organized as follows. Section 2 introduces the general theoretical framework of demand and production in the context of multi-product firms. Section 3 develops the estimation methodology for the general framework, while Section 4 describes the data used in the empirical analysis. Section 5 presents the empirical model and demonstrates the performance of the method using Monte Carlo simulations. Section 6 reports the estimation results. Section 7 presents our empirical application and quantitative exercise studying the dynamic evolution of technical efficiency. Section 8 concludes.

2 Theoretical Framework

This section develops a general framework of demand and production for multi-product firms, aimed at estimating firm-product-level measures of productivity and quality, along with the associated model parameters. While the general framework highlights the broad applicability of our estimation methodology, researchers may adopt specific functional forms suited to their empirical contexts. In Section 5, we demonstrate such an implementation in the context of our empirical application.

Consider an industry with J firms indexed by $j = 1, 2, \ldots, J$. There is a total of N products, indexed by $n = 1, 2, \ldots, N$, that firms can choose to produce. The timeline of the decisions is as follows. At the beginning of period t, the set of products that firm j has decided (at the end of the previous period) to produce in this period is Λ_{jt}. Each product $n \in \Lambda_{jt}$ is associated with a level of technical efficiency ω_{jnt} and a level of quality ξ_{jnt}, both of which have been determined and observed by the firm at the end of the previous period. The firm's capital stock is also determined in the previous period via an investment decision.

Given these state variables, the firm's static decisions consist of choosing material input and labor input at the firm level and the quantities of individual products to maximize total period profit, conditional on the observed material price, wage rate, and capital stock. The optimization conditions associated with these static decisions form the basis of our estimation strategy. At the end of period t, the firm also makes dynamic decisions regarding its capital stock and product portfolio for the following period, including the selection of products to produce and the associated levels of product quality and technical efficiency. Although our estimation strategy does not explicitly model these dynamic choices, Online Appendix C outlines the structure of these decisions, providing conceptual insight into how they are endogenously determined.

2.1 Demand

The demand for product n of firm j in period t is modeled as an inverse demand function:

$$P_{jnt} = P_{jnt}(\boldsymbol{Q}_{jt}, \boldsymbol{Q}_{-jt}; \boldsymbol{\xi}_t), \tag{1}$$

where P_{jnt} is the product price. Importantly, $\boldsymbol{Q}_{jt} = \{Q_{jnt}\}$, $n \in \Lambda_{jt}$ is a vector of quantities of the products produced by firm j in period t; $\boldsymbol{Q}_{-jt} = \{\boldsymbol{Q}_{kt}\}$, $k \neq j$ is a vector of quantities of the products produced by the competitors of firm j in period t; $\boldsymbol{\xi}_{jt} = \{\xi_{jnt}\}$, for all j and n, is a vector of quality levels of all products produced by firm j and its competitors. This function may also include a set of product characteristics if they are observable in the data.

Empirically, the realized (observed) price of a product is subject to an unexpected shock:

$$\tilde{P}_{jnt} = P_{jnt} e^{u_{jnt}}, \tag{2}$$

where u_{jnt} is assumed to be independent, identically distributed, across firm, product, and time, and $\mathbb{E}(e^{u_{jnt}}) = 1$. Crucially, the firm does not observe u_{jnt} (an ex-post shock) when making production decisions of inputs and outputs. In contrast, the firm observes ξ_{jnt} (an ex-ante shock) at the time of production decisions. The explicit modeling of the ex-ante and ex-post shocks is also adopted by Barrows et al. (2024). It is also worth noting that u_{jnt} is the sole source of discrepancy between the model-predicted revenue R_{jnt} and the realized revenue observed by researchers, $\tilde{R}_{jnt} = \tilde{P}_{jnt} Q_{jnt} = P_{jnt} Q_{jnt} e^{u_{jnt}} = R_{jnt} e^{u_{jnt}}$, while there is no ex-post, unexpected shock to product quantity.[6]

Depending on the empirical context, the demand system can be specified in various ways, including the widely-used CES demand, discrete-choice demand (e.g., Berry, 1994), and random-coefficients logit demand (e.g., Berry et al., 1995). These demand systems may allow for the possibility that a product's demand may be affected by cannibalization and competition, arising not only from the products of rival firms but also from other products offered by the same firm.

Note that incorporating a demand system is necessary when the goal is to identify productivity at the firm–product level from firm–product-level markups. As shown by Cairncross et al. (2025), firm–product-level productivity cannot be separately identified from firm–product-level markups using production data alone (i.e., without estimating a demand model). Accordingly, as demonstrated in Section 3, our empirical approach involves identifying firm–product-level markups using the demand system and, in turn, estimating

[6] An example of ex-post shock to prices arises if the firm commits to its product quantity before demand is realized. As a result, if the realized market demand exceeds expectations, the firm increases its price by a factor of $e^{u_{jnt}}$, and reduces it when the realized demand is weaker.

firm–product-level productivity via the production model, which is specified in the following subsection.

2.2 Production

We use a transformation function to model the production technology. Given the set of products to be produced (Λ_{jt}) and associated product quality (ξ_{jnt}, $n \in \Lambda_{jt}$), the firm uses labor (L_{jt}), material (M_{jt}), and capital (K_{jt}) to produce output quantity (Q_{jnt}, $n \in \Lambda_{jt}$) via a transformation function:

$$G(e^{-\tilde{\boldsymbol{\omega}}_{jt}}\boldsymbol{Q}_{jt}) = F(L_{jt}, M_{jt}, K_{jt}). \tag{3}$$

The transformation function (3) maps the firm-level input vector (L_{jt}, M_{jt}, K_{jt}) to a vector of outputs $\boldsymbol{Q}_{jt} \equiv \{Q_{jnt}\}$, $n \in \Lambda_{jt}$, given the vector of quantity-based productivity (i.e., physical productivity, or TFPQ) $\tilde{\boldsymbol{\omega}}_{jt} \equiv \{\tilde{\omega}_{jnt}\}$, $n \in \Lambda_{jt}$ of firm j in period t. Intuitively, a higher level of $\tilde{\omega}_{jnt}$ means that the firm is able to produce a higher quantity of output Q_{jnt}, conditional on the inputs and the quantity and productivity of other outputs. In this paper, we use TFPQ and productivity interchangeably.

The transformation function (3) represents the frontier of production possibility characterized by two aggregating functions $F(\cdot)$ and $G(\cdot)$.[7] Function $F(\cdot)$ is a general input aggregator. In empirical settings, it can take functional forms such as CES and translog. We adopt a CES function in our application in Section 5 and describe the implementation with a translog function in Online Appendix A.[8]

While we assume that the firm uses a single material input M_{jt} in production, our approach can readily accommodate cases in which firms employ multiple types of material inputs—whether horizontally differentiated (i.e., different material varieties) or vertically differentiated (i.e., different material quality levels)—when only the total firm-level expenditure on materials is observed. Details of this extension are provided in Online Appendix B.

The function $G(\cdot)$ is an output aggregator. We adopt a functional form that allows for

[7] The transformation function approach characterizes a firm's production possibility frontier following an approach pioneered by Powell and Gruen (1968) and used in recent literature (e.g. Cairncross et al., 2025; Koike-Mori and Martner, 2024).

[8] We exclude the Cobb–Douglas function for the purpose of controlling for unobservable firm heterogeneity of material prices. As will become clear in Section 3, our estimation methodology leverages the first-order conditions of profit maximization to uncover firm heterogeneity in material prices by examining the firm-level variation in the ratio of material expenditures to labor expenditures. However, the Cobb–Douglas functional form of $F(\cdot)$ implies a constant ratio of material to labor expenditures, which prevents us to do so.

potentially nonlinear technological substitution across products:

$$G(e^{-\tilde{\boldsymbol{\omega}}_{jt}}\boldsymbol{Q}_{jt}) \equiv \left\{ \sum_{n\in\Lambda_{jt}} \left[Q_{jnt} e^{-\tilde{\omega}_{jnt}} \right]^{\theta} \right\}^{\frac{1}{\theta}}, \tag{4}$$

where θ is a parameter that governs the elasticity of technological substitution across products and thereby influences the marginal cost differences among the products produced by the same firm. When $\theta = 1$, the marginal cost difference between two products depends solely on their relative productivities. In contrast, when $\theta \neq 1$, marginal cost differences are shaped by both productivity differences and relative output levels. Consequently, the value of θ characterizes the marginal rate of transformation (MRT) between any two products, defined as the ratio of their marginal costs. For any two products n and m, the MRT represents the amount of product m that must be forgone to produce an additional unit of n, holding inputs and all other outputs constant. Graphically, it corresponds to the slope of the production possibility frontier in the (Q_{jnt}, Q_{jmt}) space, conditional on everything else. In our setting, $\text{MRT}_{nm} = -e^{\theta(\tilde{\omega}_{jnt}-\tilde{\omega}_{jmt})}\left(\frac{Q_{jmt}}{Q_{jnt}}\right)^{\theta-1}$. Thus, θ influences how relative marginal costs are related to relative output levels. Such a relationship is analogous to the marginal cost implications derived from the transformation function model in Dhyne et al. (2022). In Section 3.2, we will use such dependence to identify θ. The CES functional form in (4) is also derived by Cairncross et al. (2025) from product-level production functions under a set of assumptions in the context of multi-product firms.

A few features of the transformation function are worth noticing. First, for multi-product firms, the transformation function can be interpreted as the frontier of joint production of all products, Q_{jnt}, $n \in \Lambda_{jt}$. This interpretation has three implications: (i) different products are manufactured with the same set of inputs; (ii) the inputs can be costlessly transferred across different products within the firm; (iii) producing more of one product means producing less of another product, holding inputs fixed. These implications are consistent with the modeling assumptions used by Dhyne et al. (2022), Orr (2022), and Valmari (2023). Second, our framework does not explicitly model input allocation within a firm. Instead, it accommodates the possibility of jointly utilized inputs across products, similar to the approach in Dhyne et al. (2022). This contrasts with existing methods that impute product-specific (exclusive) input allocations, thus abstracting away from the public-good nature of inputs within firms. Finally, an input-output separability assumption is embodied in our transformation function (3). Specifically, there are no interaction terms between outputs and inputs, although interaction among outputs and among inputs is allowed within the respective aggregators $G(\cdot)$ and $F(\cdot)$. This assumption implies that marginal cost differences across products produced by the same

firm do not depend on the input mix. Such separability is also assumed in the literature (e.g., Dhyne et al., 2022; Cairncross et al., 2025).

2.3 Productivity

A key element of our model is the quantity-based productivity $\tilde{\omega}_{jnt}$ in (3), which varies by firm, product, and period. We model the potential components and evolution of $\tilde{\omega}_{jnt}$ to highlight the key differences compared with the assumptions in the existing literature. Specifically, we unpack productivity into two components:

$$\tilde{\omega}_{jnt} = \omega_{jnt} - h(\xi_{jnt}), \tag{5}$$

where ω_{jnt} is technical efficiency and $h(\xi_{jnt})$ is a function of product quality ξ_{jnt}. We model $h(\xi_{jnt})$ as a part of quantity-based productivity because varieties of the same product category produced by different firms can be vertically differentiated by quality and such quality differences have potential implications for productivity. Producing one additional unit of the high-quality product may require more production procedures (e.g., longer refinements in the steel industry in Li et al., 2025), better (or more specialized, exclusive) machinery, higher-quality (or more) intermediate materials, higher standards of quality control (e.g., lower septic infections rate in the healthcare industry in Grieco and McDevitt, 2017), and extra dedicated workers (e.g., promoting quality or demand rather than production as discussed by Bond et al., 2021). In turn, this leads to a lower quantity of output, holding the inputs fixed, and thus it implies an increase in the *marginal cost* of production (or equivalently a lower productivity). Thus, we refer to $h(\xi_{jnt})$ as the cost of quality.[9]

As a result, differences in quantity-based productivity can be due to not only technical efficiency but also the cost of quality. Theoretically, explicitly modeling the cost of quality $h(\xi_{jnt})$ as a component of productivity allows for a trade-off between product quantity and quality, conditional on inputs. Empirically, this also implies that comparisons of quantity-based productivity across firms and over time require controlling for quality differences.

Thus, instead of modeling the evolution of quantity-based productivity, we model the evolution of technical efficiency, ω_{jnt}, as a Markov process:

$$\omega_{jnt} = g_n(\boldsymbol{\omega}_{t-1}, \boldsymbol{x}_{jt-1}) + \epsilon_{jnt}, \quad \forall n = 1, 2, \ldots, N, \tag{6}$$

[9]Note that the term cost of quality in this paper refers only to the impact of quality on the marginal cost of production, rather than the overall cost of quality (including research cost for new products with higher quality, which is more dynamic in nature, or the installation cost of new equipment to produce higher quality products, which are usually one-time fixed costs).

where ϵ_{jnt} is an innovation term.

The function $g_n(\cdot)$ flexibly captures the relevant determinants of the evolution of technical efficiency, depending on the focus of the application. For instance, in the context of technological spillovers (e.g., Malikov and Zhao, 2023), vector $\boldsymbol{\omega}_{t-1}$ may include the technical efficiency of other products within the same firm as well as the same product produced by other firms. Alternatively, in settings focused on the endogenous evolution of productivity, vector $\boldsymbol{x}_{jt-1}$ can include firm-level decisions made in period $t-1$—such as investment in research and development, as emphasized by Doraszelski and Jaumandreu (2013)—which affect the future trajectory of technical efficiency.

A key methodological advantage of our approach is that the estimation of production and demand functions does not necessarily rely on the evolution equation (6) or the productivity-quality trade-off (5). This allows researchers to estimate the technical efficiency process *after* obtaining efficiency measures, enabling a flexible modeling of dynamics. We demonstrate this advantage by exploring within- and across-firm technological spillovers in Section 7.

2.4 Inputs and Outputs Decisions

At the beginning of period t, the firm observes a vector of pre-determined variables, which includes the product scope Λ_{jt}, capital stock K_{jt}, intermediate input price P_{Mjt}, wage rate P_{Ljt}, technical efficiency $\boldsymbol{\omega}_{jt}$, and product quality $\boldsymbol{\xi}_{jt}$ of all the products. Note that observing technical efficiency and product quality implies that the firm also knows productivity, $\tilde{\boldsymbol{\omega}}_{jt}$, because the firm knows the trade-off (5). The intermediate input price and wage rate can differ across firms and fluctuate over time, driven by factors such as localized input markets and transportation costs. In empirical work, while the wage rate is typically observable, the intermediate input price is rarely recorded. This creates a challenge due to input price bias, as emphasized by De Loecker et al. (2016). Our empirical approach, detailed in Section 3, is able to address this issue. A key assumption is that firms' static input and output decisions do not influence input prices contemporaneously. While input prices may be endogenously determined—through negotiations or supply-chain investment decisions—and evolve over time, we treat them as predetermined with respect to static production choices.

The firm's objective is to maximize its total profit from all products in period t after observing its state, by optimally choosing the quantity of material (M_{jt}), the quantity of labor (L_{jt}), and the quantities of all the products to be produced ($\boldsymbol{Q}_{jt} = \{Q_{jnt}\}, n \in \Lambda_{jt}$):

$$\begin{aligned} &\max_{\boldsymbol{Q}_{jt}, M_{jt}, L_{jt}} \textstyle\sum_{n \in \Lambda_{jt}} \mathbb{E}(\tilde{P}_{jnt} Q_{jnt}) - P_{Mjt} M_{jt} - P_{Ljt} L_{jt} \\ &\qquad \text{subject to: (1) and (3)}, \end{aligned} \tag{7}$$

where the expectation is taken over the unexpected shock u_{jnt} embodied in the realized price $\tilde{P}_{jnt}$. However, this does not affect the firm's decisions on inputs and outputs because the firm does not observe the ex-post shock at the time of decisions and $\mathbb{E}(e^{u_{jnt}}) = 1$.

3 Estimation Methodology

The estimation method leverages a set of implications from the model that can be used to estimate productivity and quality at the firm-product-period level. The method is built upon the insights of Grieco et al. (2016, 2022), Harrigan et al. (2021) and Li and Zhang (2022), who utilize the first-order conditions of static profit maximization to control for unobservable variables in the production function estimation, but it is extended to the multi-product setting where within-firm allocation of inputs is unobserved. Specifically, while researchers do not observe key variables such as productivity and quality, the firm observes them before making optimal production decisions. Thus, the idea is to invert the implications from the profit maximization problem to establish a unique one-to-one mapping from observable production decisions to variables that are unobservable to researchers and control for them in the estimation of the transformation function. Crucially, under mild conditions, our model admits such a mapping regardless of the number of products.

Table 1: Comparison to existing estimation methods

	Production system	Firm-product productivity	Proxy free	Evolution free*	Material price unobservable	Demand system
DGKP	Product				●	
Orr	Product	●				●
Valmari	Product	●				●
DPSW	Transformation	●				
This paper	Transformation	●	●	●	●	●

Notes: DGKP refers to De Loecker et al. (2016), Orr refers to Orr (2022), Valmari refers to Valmari (2023), and DPSW refers to Dhyne et al. (2022). [*] This applies when the input aggregator has a CES form.

Compared with the existing methods in the literature, our method has several important innovations, as summarized by Table 1. First, our method models the production technology flexibly as a transformation function and not as a collection of single-product production functions (De Loecker et al., 2016; Orr, 2022; Valmari, 2023). This saves us from potentially restrictive assumptions regarding how firms allocate inputs to produce different products. This is especially important in the presence of shared inputs that serve as public goods within firms. In this regard, Dhyne et al. (2022)'s model is the most similar to ours. Second, our model offers the advantage of scalability as it does not require proxies for product-level productivity and rather relies on static optimization conditions that naturally increase with

the number of products. This advantage allows for the analysis of industries with a large number of products without relying on assumptions to aggregate products. Third, our method is designed to deal with the bias caused by unobserved material prices, like De Loecker et al. (2016). We employ the variation of labor and material expenditure ratio (conditional on the wage rate) to identify material prices. This is particularly useful when material prices are heterogeneous across firms and over time but are unobservable to researchers. Fourth, our method has the potential to explore the productivity evolution *after the estimation*, contrary to the existing methods which rely on productivity evolution *for the estimation*.[10]

This section is organized as follows. Section 3.1 establishes a one-to-one mapping between the observed data and unobservable heterogeneity using firm's static profit maximization conditions. Section 3.2 derives the estimating equations using the established mapping and develops the estimation strategy.

3.1 From Observables to Unobservables: a One-to-one Mapping

We begin the description of the estimation strategy by distinguishing the observable and unobservable variables to researchers in the estimation procedure. The researchers observe capital stock K_{jt}, labor input L_{jt}, labor expenditure E_{Ljt}, material expenditure E_{Mjt}, and the quantity Q_{jnt} and price P_{jnt} for each product $n \in \Lambda_{jt}$. The researchers do not observe the material price P_{Mjt} (or equivalently, the material input M_{jt}), as well as productivity $\tilde{\omega}_{jnt}$ and quality ξ_{jnt} for $n \in \Lambda_{jt}$. Our objective is to estimate these unobserved variables alongside the parameters of the transformation and demand functions.

We establish the relationship between the observed data and the unobservables, leveraging the firm's profit-maximization behavior. The idea is as follows. Although researchers cannot observe $\tilde{\omega}_{jnt}$, ξ_{jnt}, or P_{Mjt}, as described in Section 2.4 these variables are observed by the firm and thus influence the firm's optimal input and output decisions. By using the firm's optimization conditions, we establish a unique one-to-one mapping (up to a set of unknown production and demand parameters) from observable data K_{jt}, L_{jt}, E_{Ljt}, E_{Mjt}, Q_{jnt}, and P_{jnt} to unobservable variables $\tilde{\omega}_{jnt}$, ξ_{jnt}, and P_{Mjt}. We develop the strategy as follows.

Mapping to quality. We write quality, ξ_{jnt}, as a function of observed output price and quantity according to the inverse demand function (1). That is,

$$\xi_{jnt} = P_{jnt}^{-1}(\boldsymbol{P}_t, \boldsymbol{Q}_t), \tag{8}$$

[10]This depends on the functional form of the input aggregator, $F(\cdot)$. If $F(\cdot)$ has a CES form or a restricted version of translog, then the methodology can be implemented without relying on the productivity evolution; if $F(\cdot)$ has an unrestricted translog form, then additional conditions are required to estimate all translog parameters. Online Appendix A describes how to use the productivity evolution for such conditions.

where $\boldsymbol{P}_t$ and $\boldsymbol{Q}_t$ are the vectors of prices and qualities of all products and firms in period t.[11]

As an identification condition, the demand system must admit a unique solution for the quality levels given observable output prices and quality outcomes. This requirement is satisfied by a broad class of demand functions, including the widely adopted CES demand, discrete-choice demand, and random-coefficients logit demand models.

Standard methods for estimating these demand systems, typically relying on the use of appropriate instrumental variables, are well-established in the literature. Consequently, the identification and estimation of the demand system within our framework can be conducted as a standalone process. Once the demand system is estimated, the quality level ξ_{jnt} can be recovered using (8). Moreover, the estimated demand system allows us to compute the price elasticity of demand for any product n of firm j with respect to product m of firm j' as:

$$\frac{\partial Q_{jnt}}{\partial P_{j'mt}} \frac{P_{j'mt}}{Q_{jnt}} \equiv -\eta_{jtnm}. \tag{9}$$

Note that we have slightly abused the notation because product m can be either a product of the same firm $j' = j$ (i.e., cannibalization) or a product of another firm $j' \neq j$ (i.e., competition). That is, the elasticity can be flexible and vary by firm, product, and time.

Mapping to material price. We derive the mapping using the firm's static profit maximization problem. The Lagrange function implied by the problem (7) is:

$$\mathcal{L}_{jt} = \sum_{n \in \Lambda jt} P_{jnt}(\boldsymbol{Q}_{jt}, \boldsymbol{Q}_{-jt}; \boldsymbol{\xi}_t) Q_{jnt} - P_{Ljt} L_{jt} - P_{Mjt} M_{jt} - \lambda_{jt} \left\{ G(e^{-\tilde{\omega}_{jt}} \boldsymbol{Q}_{jt}) - F(L_{jt}, M_{jt}, K_{jt}) \right\}, \tag{10}$$

where λ_{jt} is the Lagrangian multiplier. The random shock u_{jnt} is not included in this equation because it is ex post and $\mathbb{E}(e^{u_{jnt}}) = 1$.

The first-order conditions with respect to labor and material inputs are, respectively:

$$\frac{\partial \mathcal{L}_{jt}}{\partial L_{jt}} = -P_{Ljt} + \lambda_{jt} \frac{\partial F(L_{jt}, M_{jt}, K_{jt})}{\partial L_{jt}} = 0, \tag{11}$$

$$\frac{\partial \mathcal{L}_{jt}}{\partial M_{jt}} = -P_{Mjt} + \lambda_{jt} \frac{\partial F(L_{jt}, M_{jt}, K_{jt})}{\partial M_{jt}} = 0. \tag{12}$$

Multiply them by L_{jt} and M_{jt}, respectively, and take the ratio of the two to obtain:

$$\frac{\frac{\partial F}{\partial L_{jt}} \frac{L_{jt}}{F}}{\frac{\partial F}{\partial M_{jt}} \frac{M_{jt}}{F}} = \frac{E_{Ljt}}{E_{Mjt}}. \tag{13}$$

[11] Empirically, the recovered ξ_{jnt} contains the unexpected shock u_{jnt}, which usually appears as a (log-) additive term in popular demand functions. We clarify this further in the setup of CES demand in Section 5.

This equation only involves a single unobservable variable, M_{jt}, and, for functional forms of $F(\cdot)$ such as CES and translog, admits a unique solution:[12]

$$M_{jt} = M(L_{jt}, E_{Ljt}, E_{Mjt}, K_{jt}), \tag{14}$$

and consequently,

$$P_{Mjt} = \frac{E_{Mjt}}{M(L_{jt}, E_{Ljt}, E_{Mjt}, K_{jt})}. \tag{15}$$

The identification strategy for P_{Mjt} is based on the relationship implied by the first-order conditions for labor and material inputs, and is conceptually aligned with Grieco et al. (2016). Conditional on the wage rate, changes in P_{Mjt} induce a non-Hicks-neutral effect by altering the optimal ratio of labor to material expenditures. Consequently, variations in the labor-to-material expenditure ratio observed in the data (conditional on the wage rate) provide a basis for identifying P_{Mjt}. This strategy presents an alternative to the method proposed by De Loecker et al. (2016), who use output prices as proxies for input prices to address the input price bias arising from unobserved heterogeneity in input prices.

Substituting (14) into the first order condition for labor, we obtain a unique solution for the Lagrangian multiplier:

$$\lambda_{jt} = \frac{P_{Ljt}}{\frac{\partial F(L_{jt}, M(L_{jt}, E_{Ljt}, E_{Mjt}, K_{jt}), K_{jt})}{\partial L_{jt}}}. \tag{16}$$

That is, the Lagrangian multiplier is derived from equalizing the marginal benefit and marginal cost of labor input. This equation can also be cast in terms of material input, which is equivalent to (16) due to how (unobserved) material input and its price are recovered.

Mapping to productivity. The first-order condition with respect to each product

[12] When the functional form of F satisfies the condition proposed by Proposition 1 of the Online Appendix of Grieco et al. (2016), there exists a unique solution for M_{jt}. In our empirical application, the CES functional form of $F(\cdot)$ satisfies this condition. We provide a full procedure for estimating a translog functional form of $F(\cdot)$ in Appendix A under mild conditions for the evolution of technical efficiency, and we establish the condition to ensure the unique solution for M_{jt}. For the Cobb-Douglas form of $F(\cdot)$, such a solution does not exist because the elasticity ratio on the left-hand side of (13) is always a constant. Intuitively, the material price variation in a Cobb-Douglas production function does not change the optimal ratio of labor and material expenditures. Thus, we refrain from using the Cobb-Douglas form of $F(\cdot)$.

quantity Q_{jnt}, $n \in \Lambda_{jt}$, is:

$$\begin{aligned}\frac{\partial \mathcal{L}_{jt}}{\partial Q_{jnt}} &= \left\{ \sum_{m \in \Lambda_{jt}} \frac{\partial P_{jmt}(\boldsymbol{Q}_{jt}, \boldsymbol{Q}_{-jt}; \boldsymbol{\xi}_t)}{\partial Q_{jnt}} Q_{jmt} + P_{jnt} \right\} - \lambda_{jt} \frac{\partial G(e^{-\tilde{\boldsymbol{\omega}}_{jt}} \boldsymbol{Q}_{jt})}{\partial Q_{jnt}} \\ &= \underbrace{\frac{P_{jnt}}{\mu_{jnt}}}_{\text{marginal revenue}} \underbrace{- \lambda_{jt} e^{-\theta \tilde{\omega}_{jnt}} Q_{jnt}^{\theta-1} [G(e^{-\tilde{\boldsymbol{\omega}}_{jt}} \boldsymbol{Q}_{jt})]^{1-\theta}}_{\text{marginal cost}} = 0, \end{aligned} \tag{17}$$

where

$$\mu_{jnt} \equiv \frac{1}{1 - \sum_{m \in \Lambda_{jt}} \frac{1}{\eta_{jtnm}} \frac{R_{jmt}}{R_{jnt}}} \tag{18}$$

is the markup of product n and η_{jtnm} is the price elasticity of demand defined by (9).

Notably, $\lambda_{jt} e^{-\theta \tilde{\omega}_{jnt}} Q_{jnt}^{\theta-1} [G(e^{-\tilde{\boldsymbol{\omega}}_{jt}} \boldsymbol{Q}_{jt})]^{1-\theta}$ is the marginal cost of producing Q_{jnt}. Across firms, the marginal cost of a product varies due to λ_{jt}; within a firm, the marginal cost also differs due to productivity $\tilde{\omega}_{jnt}$ and scale of production Q_{jnt}. As a result, conditional on a firm, the variation in product prices identifies the productivity difference across products within the firm, after accounting for the markup μ_{jnt} and production scale Q_{jnt}.

This idea is formally developed to derive the productivity mapping. Using (17) and substituting M_{jt} in it by (14), we obtain:

$$e^{\theta \tilde{\omega}_{jnt}} = \frac{\mu_{jnt}}{P_{jnt}} \lambda_{jt} Q_{jnt}^{\theta-1} [F(L_{jt}, M(L_{jt}, E_{Ljt}, E_{Mjt}, K_{jt}), K_{jt})]^{1-\theta}, \tag{19}$$

where we have substituted $G(\cdot)$ by using (3) and λ_{jt} is given by (16).[13]

In summary, we have established a one-to-one mapping—comprising (8), (15), and (19) — from observable data to the unobservable variables ξ_{jnt}, P_{Mjt}, and $\tilde{\omega}_{jnt}$, conditional on the demand and production parameters to be estimated. This mapping is unique for widely used demand functions, such as CES demand, discrete-choice demand, and random-coefficients demand models, as well as for common production function specifications, including CES and translog functional forms. Conceptually, our approach parallels the proxy-based methodology pioneered by Olley and Pakes (1996), and extended by a large body of methodological work, which uses observable proxies (such as investment and material inputs) to control for unobserved productivity when estimating production functions. However, in the context of multi-product firms with product-level heterogeneity, the proxy-based approach faces a scalability challenge: the number of required proxies grows with the number of products, as recognized by Dhyne et al. (2022). Our methodology leverages firms' first-order conditions to

[13] Empirically, the recovered productivity contains the unexpected shock u_{jnt} (as a log-additive term).

construct the mapping, offering a key advantage in scalability. As the number of products increases, so does the number of first-order conditions. This scalability is shared with recent approaches such as Orr (2022) and Valmari (2023), while we also adopt a transformation function approach as in Dhyne et al. (2022) to avoid assigning inputs to the production of individual outputs.

3.2 Estimating Equations and Estimation Strategy

In the previous subsection, we have constructed a one-to-one mapping from observable variables to the unobserved ξ_{jnt}, $\tilde{\omega}_{jnt}$, and P_{Mjt} (or M_{jt} equivalently) up to a set of parameters to be estimated. This mapping is the key to developing the equations to estimate these parameters, which we derive in this subsection.

Estimating a general demand system (1) is challenging due to unobservable demand factors (e.g., quality) and the endogeneity of prices. Depending on specific context and functional form of (1), strategies are well-developed (e.g. Berry, 1994; Berry et al., 1995) to address these challenges, mainly using a set of instrumental variables. Since our focus is on the production transformation function, we assume the existence of a valid set of instrumental variables, allowing researchers to estimate the demand system (1). Consequently, the firm-product-time-specific quality ξ_{jnt} and markup μ_{jnt} can be recovered via (8) and (18), respectively.

To derive our main estimating equation, we start by multiplying both sides of the equation implied by the first-order condition (17) by Q_{jnt}. Rearranging this equation gives:

$$\frac{R_{jnt}}{\mu_{jnt}} = \lambda_{jt} e^{-\theta\tilde{\omega}_{jnt}} Q_{jnt}^{\theta} [G(e^{-\tilde{\boldsymbol{\omega}}_{jt}} \boldsymbol{Q}_{jt})]^{1-\theta}, \tag{20}$$

where $R_{jnt} = P_{jnt}Q_{jnt}$.

Sum the above equation over $n \in \Lambda_{jt}$ to obtain:

$$\sum_{n\in\Lambda_{jt}} \frac{R_{jnt}}{\mu_{jnt}} = \lambda_{jt} \left[G(e^{-\tilde{\boldsymbol{\omega}}_{jt}} \boldsymbol{Q}_{jt})\right]^{1-\theta} \sum_{n\in\Lambda_{jt}} \left(e^{-\theta\tilde{\omega}_{jnt}} Q_{jnt}^{\theta}\right) = \lambda_{jt} F(L_{jt}, M_{jt}, K_{jt}), \tag{21}$$

where we have used the transformation function to replace $G(\cdot)$ to obtain the last equality.

From the first-order conditions of labor input and material input, (11) and (12), we obtain

$$\lambda_{jt} F_{jt} = \frac{E_{Ljt} + E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}} \frac{L_{jt}}{F_{jt}} + \frac{\partial F_{jt}}{\partial M_{jt}} \frac{M_{jt}}{F_{jt}}}, \tag{22}$$

where F_{jt} is a short-hand notation of $F(L_{jt}, M(L_{jt}, E_{Ljt}, E_{Mjt}, K_{jt}), K_{jt})$.

Substitute this equation into (21) to obtain:[14]

$$\sum_{n\in\Lambda_{jt}} \frac{R_{jnt}}{\mu_{jnt}} = \frac{E_{Ljt}+E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}+\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}. \tag{23}$$

Therefore, (23) describes the relationship between revenues (adjusted by the reciprocal of markups) and inputs for a general system of demand and production transformation functions in the context of multi-product firms. This equation is an analog of the popular ratio estimator of markup in De Loecker et al. (2016) in the context of single-product firms.[15] De Loecker et al. (2016) focus on uncovering markups after estimating the production function parameters (and the output elasticities), without relying on the estimation of any demand system. In the context of multi-product firms, this equation is an analog of the markup-input share relationship examined by Cairncross et al. (2025). Importantly, Cairncross et al. (2025) show that firm–product-level markups and firm–product-level productivity cannot be separately identified using production data alone. This insight implies that it is useful to specify and estimate a demand model using product-level quantity and price data to identify firm–product-level markups as in Berry et al. (1995) and, in turn, to estimate firm–product-level productivity via the production model.

Such an insight is adopted in our methodology. We first utilize the demand system (1) to uncover the markups and then proceed to estimate the production parameters using (23) as the estimating equation. In addition, this strategy offers several advantages. It is scalable for handling a large number of products, addresses bias caused by unobservable material price heterogeneity, and makes it possible to estimate production parameters without relying on productivity evolution process, because all unobserved firm heterogeneity (i.e., multi-dimensional productivity and quality as well material quantity) are substituted by observable variables in the data using the mapping developed in Section 3.1.

It is important to notice that (23) holds for the theoretically predicted revenue R_{jnt}, because it is derived from the firm's profit maximization problem. Empirically, researchers do not observe the theoretically predicted revenue R_{jnt}; instead, the observed revenue contains the unexpected shock as defined in (2): $\tilde{R}_{jnt} = \tilde{P}_{jnt}Q_{jnt} = P_{jnt}Q_{jnt}e^{u_{jnt}} = R_{jnt}e^{u_{jnt}}$. Substitute this relationship into (23) to replace the theoretically predicted revenue, and rearrange to

[14] An alternative expression is $\sum_{n\in\Lambda_{jt}} \frac{R_{jnt}}{\mu_{jnt}} = \frac{E_{Ljt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}}$ or $\sum_{n\in\Lambda_{jt}} \frac{R_{jnt}}{\mu_{jnt}} = \frac{E_{Mjt}}{\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}$. Nonetheless, because we substitute the recovered material quantity (14), which is derived from (13), into these equations, both of these equations are equivalent to (23).

[15] To see this, notice that, for single-product firms, (23) degenerates to $\frac{R_{jt}}{\mu_{jt}} = \frac{E_{Ljt}+E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}+\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}$, and consequently, the markup can be written as $\mu_{jt} = \frac{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}+\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}{(E_{Ljt}+E_{Mjt})/R_{jt}}$.

obtain an empirical estimating equation that involves the observed revenue directly:

$$\ln\left[\sum_{n\in\Lambda_{jt}}\frac{\tilde{R}_{jnt}}{\mu_{jnt}}\right]=\ln\left[\frac{E_{Ljt}+E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}+\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}\right]-u_{jt}, \tag{24}$$

where

$$u_{jt}=\ln\left\{\sum_{n\in\Lambda_{jt}}\left[\frac{\tilde{R}_{jnt}/\mu_{jnt}}{\sum_{n\in\Lambda_{jt}}\tilde{R}_{jnt}/\mu_{jnt}}e^{-u_{jnt}}\right]\right\} \tag{25}$$

is a firm-level composite error term. Intuitively, it is a geometric mean (in logarithm) of firm-product level unexpected shock u_{jnt}, using within-firm share of $\tilde{R}_{jnt}/\mu_{jnt}$ as the weights.[16]

We estimate the associated production parameters, denoted as β, using generalized method of moments (GMM):

$$\begin{aligned}\hat{\beta} &= \text{argmin}_{\beta}\left[\frac{1}{\mathbb{N}}\sum_{j,t}u_{jt}Z_{jt}\right]' W\left[\frac{1}{\mathbb{N}}\sum_{j,t}u_{jt}Z_{jt}\right], \\ \text{where}\qquad u_{jt}&=\ln\left[\frac{E_{Ljt}+E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}+\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}\right]-\ln\left[\sum_{n\in\Lambda_{jt}}\frac{\tilde{R}_{jnt}}{\mu_{jnt}}\right].\end{aligned} \tag{26}$$

W is a weight matrix, $\mathbb{N}$ is the number of firm-time observations, and Z_{jt} is a set of instrumental variables. Because u_{jt} is a composite of ex-post error terms of prices, natural candidates of instrumental variables include firm-level inputs such as L_{jt}, E_{Ljt}, E_{Mjt} and K_{jt}. In addition, in settings where firms compete in product markets, product characteristics of rival firms, if observable, may be also included in the instrumental variable set.

Although this estimation strategy offers a straightforward approach to estimating the primary production parameters, the parameter θ which characterizes the technological substitution of outputs within the firm, does not appear in the estimating equation (23). To identify and estimate θ, we leverage the influence of θ on the marginal rate of transformation (via within-firm marginal cost differences) across products within a firm.

Specifically, take the ratio of the equation implied by the first-order condition (17) of

[16] In the context of the firm-level shock (i.e., $u_{jnt}=u_{jt}, \forall n$) or in the context of single-product firms, there is only one unexpected shock per firm. Consequently, (24) simplifies to $\ln\left[\frac{\tilde{R}_{jt}}{\mu_{jt}}\right]=\ln\left[\frac{E_{Ljt}+E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}+\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}\right]+u_{jt}$. This degenerated form aligns with the estimating equation proposed by Grieco et al. (2016).

product n to that of product m. The logarithm of the ratio is:

$$\underbrace{\ln\left[\frac{P_{jnt}/\mu_{jnt}}{P_{jmt}/\mu_{jmt}}\right]}_{\text{marginal cost ratio, in log}} = \underbrace{(\theta-1)\ln\left(\frac{Q_{jnt}}{Q_{jmt}}\right)+v_{jnt}}_{\text{marginal rate of transformation, in log}}, \tag{27}$$

where $v_{jnt} = \theta(\tilde{\omega}_{jmt} - \tilde{\omega}_{jnt})$ is the relative *difference* between the productivity of the two products, adjusted by parameter θ. Intuitively, this equation aligns with the definition of the marginal rate of transformation discussed in Section 2.2. The left-hand side corresponds to the ratio of marginal costs, while the right-hand side represents the marginal rate of transformation—both expressed in logarithmic terms. Unless $\theta = 1$, the marginal cost ratio depends not only on productivity differences but also on the relative scale of production between the two products. Therefore, we identify θ by examining how marginal costs differ across products based on their relative output levels within the same firm. This identification strategy aligns with the insight of Khmelnitskaya et al. (2025).

To implement this idea, we treat v_{jnt} as an error term. Since v_{jnt} is correlated with the production scale ratio, there is an endogeneity problem. Empirically, researchers can estimate (27) using a Two-Stage Least Squares (2SLS) estimator with a set of IVs. Ideally, firm-product-level instrumental variables are preferred. For example, differences in product characteristics may shift the quantity ratio due to varying demand driven by these characteristics, while being traditionally assumed to be uncorrelated with cost-side (productivity) differences. When firm-product-level instruments are unavailable, firm-level variables, such as the wage rate, can be used as instruments. This approach benefits from examining *relative* differences between two products within the same firm. For instance, conditional on other factors, a lower wage rate decreases the firm's overall marginal cost, leading to higher production quantities for both products. However, the product with less elastic demand (e.g., product m) expands more, resulting in a lower quantity ratio (Q_{jnt}/Q_{jmt}). Thus, the wage rate and production scale ratio are correlated. Of course, the validity of firm-level instruments depends on the assumption that the *levels* of these variables are uncorrelated with the *differences* in productivity between two products, which is discussed in Online Appendix D. We examine the performance of our estimation method in a Monte Carlo setting in Section 5.3.

As a summary of the full estimating approach, the first step is to estimate the demand system (1) to obtain the estimates of demand parameters and product markups. Second, we estimate θ from the within-firm marginal rate of transformation relationship (27) using 2SLS. The third step is to estimate the production parameters using (23) via GMM. With these estimates, researchers can compute quality and productivity via (8) and (19), respectively.

Although the model and estimation strategy are presented in a general framework, re-

searchers should tailor their choice of functional forms to their specific context, balancing flexibility with the feasibility of empirical implementation. In our application, we select a CES demand system and a transformation function with a CES input aggregator, due to the consideration of the characteristics of the industries, the market structure, and the availability of instrumental variables. The next section focuses on the industry characteristics and market structure, providing the context for these choices.

4 Data

We estimate our model using firm-level Mexican manufacturing data, collected by the *Instituto Nacional de Estadística y Geografía* (National Institute of Statistics and Geography, INEGI henceforth) and covering the period 1994-2007. We use two datasets: the *Encuesta Industrial Anual* (Annual Industrial Survey, EIA henceforth), the main annual survey covering the manufacturing sector, and the *Encuesta Industrial Mensual* (Monthly Industrial Survey, EIM henceforth), a monthly survey that monitors short-term trends related to employment and output.[17] These datasets are particularly useful for our analysis because they provide quantity and sales information at the firm-product level. However, similar to most production data, information regarding inputs, viz. physical capital, intermediate input, number of workers and wage bills, are only available at the firm level.[18]

Firms are classified by INEGI into one of the classes of activity based on their principal product. A class of activity is the most disaggregated level of industrial classification and is defined at six digits according to the 1994 *Clasificación Mexicana de Actividades y Productos* (Mexican System of Classification for Activities and Products, CMAP henceforth). Firms report quantity and sales information product by product based on their industries.

We focus on three specific classes of activities: manufacturing of footwear, mainly of leather (class 324001, footwear in short); printing and binding (class 342003, printing in short); and manufacturing of pharmaceutical products (class 352100, pharmaceuticals in short).[19] These

[17]The unit of observation in both surveys is a plant rather than a firm and the sample includes all plants with more than 100 employees as well as a sample of smaller plants. For simplicity and in line with the literature, we will use the term "firm" to refer to a plant. More information on the EIA and EIM can be found in Caselli et al. (2017).

[18]All nominal variables are deflated using the consumer price index. To facilitate comparison, we normalize average industry output prices to 1. Initial capital stock and investment are deflated using industry-level price indices.

[19]For the purpose of our analysis, all products with fewer than 100 observations are aggregated together in a residual product category. The residual product category is defined as "Others" (product code 99) in Table A1 in the Online Appendix. The prices and quantities of the aggregated residual product category are estimated following Diewert et al. (2009). While this aggregation is required to estimate the demand elasticity of substitution for each product based on a large enough number of observations, it only implies that the demand elasticity of substitution is by assumption equal across all products included in the residual product category within an industry. In addition, this aggregation involves a relatively small share of products: the

three industries were chosen because each industry is made up of more than 500 firm-year observations, a number of observations large enough for implementing our estimation strategy. More importantly, multi-product firms are particularly prevalent in these industries: 56% of firms in these industries are multi-product producers and such firms account for 86% of total revenues and produce on average 6.9 products per year. They also represent a diverse set of manufacturing industries with clear concepts of product quality: for example, advanced design and assembly that provide superior comfort and durability in the footwear industry; acid-free paper and durable binding in the printing industry; potent active ingredients and degrading-preventing packaging in the pharmaceutical industry.

There are a few patterns worth noting. First, multi-product production is an essential feature of the firms in our sample. We demonstrate this point by using an index that is analogous to the traditional Herfindahl–Hirschman Index (HHI) as the sum of the squared shares of sales within a firm. A higher HHI index means a higher level of concentration of sales within a firm.[20] The index is naturally equal to one for single-product producers. For firms with a larger product scope, HHI decreases sharply becoming close to 0.3 for firm-year pairs producing 5 products and close to 0.2 for firm-year pairs producing 10 or more products.[21] These values imply that producers are genuine multi-product firms – they do not concentrate production entirely on their top products, and all products, albeit to different degrees, are important for firms' total revenues.[22] Thus, multi-product firms need to be treated and modeled as such and they cannot be simplified as single-product producers. This characteristic of the industries is also an important feature that enables us to exploit the within-firm relationship to identify model parameters, as discussed in Online Appendix D.

The importance of multiple-product production is also present in all the industries of our analysis, albeit with some degrees of variation, as shown in Table 2.[23] The percentage of multi-product firms ranges from 20% in the footwear industry to 54% in printing and 85% in pharmaceuticals and they account for an even larger share of revenues (from 39%

main (i.e., not aggregated) products account for between 81% and 93% of observations and between 82% and 90% of revenue across the three industries. Accordingly, the descriptive statistics and patterns demonstrated in this section are reported based on the aggregated categories, which is the data used in our estimation.

[20] In Figure A1 in the Online Appendix, we aggregate the firm-level index with weights equal to the firms' total revenues, by firm-year pairs' product scope.

[21] These values indeed show some degree of concentration of sales within firms. For example, if a firm produces 5 products with equal sales, the index would be 0.2. The fact that the index is close to 0.3 implies that there exists an uneven distribution of sales. We explore this heterogeneity using quality and productivity within firms in Section 7.

[22] To confirm that firms rely heavily on all products for their total sales, Online Appendix Table A2 shows the average within-firm product shares by product scope. For instance, for firms producing 5 or more products, the share of products other than the top product is 0.556 and the share of products with rank 5 and beyond is 0.147, on average.

[23] Additional descriptive statistics are available in Table A3 in the Online Appendix.

Table 2: Descriptive statistics: prevailing multi-product firms

Variable	Footwear	Printing	Pharmaceutical
Product scope, MPFs only	2.403	6.195	8.000
	(0.624)	(4.107)	(3.018)
Share of number of MPFs	0.201	0.538	0.845
Revenue share of MPFs	0.386	0.597	0.939
Total number of products	4	15	16
Total number of firms	72	79	80
Number of firm-year pairs	617	744	867

Notes: The table reports the means and standard deviations (in parenthesis) for each variable by industry. Product scope is the number of products manufactured by firm. MPFs refers to multi-product firms only.

in the footwear industry to 94% in pharmaceuticals). The average product scope is larger in printing and pharmaceuticals (respectively, 6.2 and 8.0 for multi-product firms) than in the footwear industry (2.4). These differences in average product scope are in line with the number of product categories available in each industry, which ranges from 4 in footwear to 16 in pharmaceuticals.

Second, the status of being a multi-product firm is quite persistent, and so is the product scope. In particular, using a simple autoregressive process of the number of products produced by each firm, we measure the persistence coefficients to be 0.87, 0.96, and 0.98 in the three industries, respectively.[24] Thus, multi-product firms unequivocally dominate manufacturing production in our data and their within-firm adjustment across products is more salient than the extensive margin adjustment in changing the number of products.

These patterns imply that both within-firm and across-firm heterogeneity is important. On the one hand, there exist persistent characteristics at the firm level that determine the performance across firms. On the other hand, within-firm heterogeneity and product scope play a significant role in shaping these characteristics within firms. These implications are in line with the specification for productivity (32), which contains a common component at the firm level to capture the differences across firms as well as an individual component varying at the firm-product level to explain the variation of performance within a firm.

Finally, the sample reflects patterns consistent with the choice of the empirical demand model in Section 5. On average, about 16 to 37 firms compete in the market for any given product in any given year. The majority of the firms do not command a dominant share of the market – the median (traditionally defined) HHI across firms at the product-year level

[24]The entry of new products and the exit of old products only account for 6.8 and 7.3 percent of the observations, respectively.

ranges between 0.15 in the pharmaceutical industry and 0.31 in the printing industry. More importantly, given the level of product disaggregation, the markets for different products (e.g., women's shoes vs. men's shoes in the footwear industry, and more examples in Online Appendix Table A1) are reasonably assumed as segmented. For each product, firms' outputs are vertically differentiated as evidenced by the large dispersion in prices.[25] Overall, these patterns support abstracting from demand cannibalization across products made by the same firm and assuming that firms face monopolistic competition within each product category.

5 Empirical Model and Estimation

This section presents an empirical model and applies the estimation strategy developed in Section 3 to the Mexican manufacturing industries. We choose specific functional forms for the demand and production model considering the characteristics of the industries, product classification, market structure, and availability of instrumental variables.

5.1 Empirical Specification

On the demand side, we adopt a CES demand function, assuming that while the demand for each of the N products is segmented, there is monopolistic competition among firms producing vertically differentiated products within the same product line. Formally, the realized price, $\tilde{P}_{jnt}$, from the inverse demand function (1), after accounting for the unexpected shock in (2), for product n of firm j in period t, is specified as:

$$\ln \tilde{P}_{jnt} = -\frac{1}{\eta_n} \ln Q_{jnt} + \frac{1}{\eta_n} \underbrace{(\phi_{nt} + \psi_{jn} + v_{jt})}_{\xi_{jnt}} + u_{jnt}, \tag{28}$$

where η_n denotes the constant elasticity of demand for product n. The term ξ_{jnt} represents product quality and comprises three components: ϕ_{nt} (product-time fixed effects), ψ_{jn} (firm-product fixed effects), and v_{jt} (firm-time fixed effects). The term u_{jnt} represents an idiosyncratic firm-product-time specific ex-post price shock. We follow the tradition in the literature (e.g., Melitz, 2000; Khandelwal, 2010; Hottman et al., 2016; Pozzi and Schivardi, 2016; Eslava et al., 2024) to refer to the residual recovered from the CES demand system as the perceived product quality: $\tilde{\xi}_{jnt} \equiv \ln Q_{jnt} + \eta_n \ln \tilde{P}_{jnt} = \xi_{jnt} + \eta_n u_{jnt}$.

The empirical demand model (28) excludes complementarity or substitution across different product lines. This choice is driven by empirical considerations in our context. First, given the level of product classification in our data, complementarity or substitution on the demand

[25] For example, the interquartile range of prices in logarithm is about 1.4 (i.e., a 400% difference) within a product category, on average, across the three industries.

side is unlikely. For example, the demand for women's shoes is unlikely to be influenced by competition from men's shoes. Similar functional forms have been employed by De Loecker (2011) and Valmari (2023) in modeling demand functions for multi-product contexts. Second, while estimating a richer model is conceptually appealing, a major difficulty arises from the lack of suitable instrumental variables to address the endogeneity issue associated with unobserved product quality. Traditional instruments, such as cost shifters, may fail in vertically differentiated markets where higher-quality inputs (thus higher input prices) are chosen to enhance product quality. In general, estimating flexible demand systems requires carefully constructed instruments that are uncorrelated with product quality. For example, Berry et al. (1995) utilize the characteristics of other automobiles produced by the firm itself and similar automobiles produced by its rivals. In our case, the dataset does not include such rich and strong instruments. Therefore, we adopt the CES functional form for the demand function, which enables us to leverage the within-firm variation in sales across products to estimate the constant demand elasticity parameter η_n, as described in Section 5.2.

On the production side, we use a CES input aggregator in the transformation function (3):

$$F(L_{jt}, M_{jt}, K_{jt}) \equiv \left[\alpha_L L_{jt}^{\gamma} + \alpha_M M_{jt}^{\gamma} + \alpha_K K_{jt}^{\gamma}\right]^{\frac{\rho}{\gamma}}, \tag{29}$$

where $\gamma \equiv \frac{\sigma-1}{\sigma}$ governs the elasticity of substitution across inputs. ρ is a parameter for the returns to scale in the transformation of inputs into output. α_L, α_M, and α_K are share parameters associated with labor, material, and capital, respectively. We normalize their sum to 1. We maintain the output aggregator as defined in (4).

5.2 Estimation

Applying the methodology described in Section 3.2 to the empirical model, we obtain an explicit, unique mapping from observable data to the unobservable variables:

$$\tilde{\xi}_{jnt} = \ln Q_{jnt} + \eta_n \ln \tilde{P}_{jnt}, \tag{30}$$

$$P_{Mjt} = \left[\frac{\alpha_M}{\alpha_L}\right]^{\frac{1}{\gamma}} \left[\frac{E_{Mjt}}{E_{Ljt}}\right]^{1-\frac{1}{\gamma}} P_{Ljt}, \tag{31}$$

$$e^{\theta\tilde{\omega}_{jnt}} = \frac{\eta_n}{(\eta_n - 1)\tilde{P}_{jnt}} \frac{E_{Ljt}}{\rho\alpha_L L_{jt}^{\gamma}} \left[\alpha_L L_{jt}^{\gamma}\left(1 + \frac{E_{M_{jt}}}{E_{L_{jt}}}\right) + \alpha_K K_{jt}^{\gamma}\right]^{1-\frac{\rho\theta}{\gamma}} Q_{jnt}^{\theta-1}. \tag{32}$$

Equations (30), (31), and (32) empirically represent the mappings of quality (8), material price (15), and productivity (19) in the general model. Specifically, $\tilde{\xi}_{jnt}$ is expressed as a

function of observed price and quantity, capturing how product quality can be inferred from observable market outcomes. Similarly, P_{Mjt} is determined as a function of the labor-to-material expenditure ratio, conditional on the wage rate, in the same spirit as in Grieco et al. (2016). Finally, the identification of $\tilde{\omega}_{jnt}$ integrates variations at both the firm level (i.e., L_{jt}, K_{jt}, E_{Ljt}, and E_{Mjt}) and the firm-product level (i.e., P_{jnt} and Q_{jnt}).

After substituting the mapping to the transformation function to replace the unobserved productivity and material input, we obtain an explicit expression of (24) in logarithm as:

$$\ln\left[\sum_{n\in\Lambda_{jt}}\frac{(\eta_n-1)\rho}{\eta_n}\tilde{R}_{jnt}\right]=\ln\left[E_{M_{jt}}+E_{L_{jt}}\left(1+\frac{\alpha_K}{\alpha_L}\left(\frac{K_{jt}}{L_{jt}}\right)^{\gamma}\right)\right]-u_{jt}, \tag{33}$$

where

$$u_{jt}=\ln\left\{\sum_{n\in\Lambda_{jt}}\left[\frac{\frac{(\eta_n-1)}{\eta_n}\tilde{R}_{jnt}}{\sum_{n\in\Lambda_{jt}}\frac{(\eta_n-1)}{\eta_n}\tilde{R}_{jnt}}e^{-u_{jnt}}\right]\right\}. \tag{34}$$

This equation is the multi-product version of the estimating equation proposed by Grieco et al. (2016) (see their equation 8), who assume that each firm produces a single product.[26] In the context of multi-product firms, the individual product revenues are adjusted by the reciprocal of their corresponding markups.[27]

As explained in Section 3.2, the parameter θ is not present in the estimating equation (33) and thus is not identified by (33) alone. Thus, we estimate (27) via 2SLS to identify θ. In our implementation, the IV set consists of a constant and the logarithm of the wage rate (P_{Ljt}), the capital stock (K_{jt}), and the ratio of material expenditure to labor (E_{Mjt}/L_{jt}, as a proxy for material prices, conditional on the wage rate).[28] This provides an estimate of $\hat{\theta}$.

Nonetheless, we still face two additional challenges to estimate all the parameters.[29] First, ρ is not separately identified from demand elasticities in (33). In fact, only a combination of

[26] More broadly, (33), without logarithms, is also similar to the estimating equations used by Das et al. (2007), Aw et al. (2011), and Li (2018) with data on the firm's total variable cost to estimate demand elasticities in multiple markets.

[27] If the elasticities (markups) are the same, then the estimating equation is the same as in Grieco et al. (2016). We also allow for the returns to scale parameter, ρ, to be estimated, while Grieco et al. (2016) assume it to be one.

[28] To see this, note that (31) is equivalent to $P_{Mjt}=\left[\frac{\alpha_M}{\alpha_L}\right]^{\frac{1}{\gamma}}\left[\frac{E_{Mjt}}{L_{jt}}\right]^{1-\frac{1}{\gamma}}P_{Ljt}^{\frac{1}{\gamma}}$. Taking logarithm, we obtain $\ln(P_{Mjt})=\frac{1}{\gamma}\ln\left[\frac{\alpha_M}{\alpha_L}\right]+(1-\frac{1}{\gamma})\ln\left[\frac{E_{Mjt}}{L_{jt}}\right]+\frac{1}{\gamma}\ln(P_{Ljt})$. Because we include the logarithm of the wage rate, $\ln(P_{Ljt})$, in the IV set, using $\ln\left[\frac{E_{Mjt}}{L_{jt}}\right]$ is equivalent to using $\ln(P_{Mjt})$ in this setting, although P_{Mjt} is not observable. Our result is quantitatively similar if the expenditure ratio of material and labor is used as an IV.

[29] These additional challenges arise because we are estimating the returns to scale parameter, ρ, and because the standard input data available in the Mexican dataset lacks instrumental variables to estimate the demand function (28) directly. See footnote 30 for further discussion.

η_n and ρ (i.e., $\frac{(\eta_n-1)\rho}{\eta_n}$) is identified. Second, estimating (33) via GMM requires (at least) the same number of instrumental variables as the number of products to identify $\frac{(\eta_n-1)\rho}{\eta_n}$ of each product, because product revenues are correlated with composite shock u_{jt}.

To address both challenges simultaneously, we leverage the context of multi-product firms, which provides valuable within-firm variation. We explore the relationship between the revenues of any two products implied by the firm's static maximization problem, taking into account that the markets for different products are segmented in our empirical context. Here, η_n influences the sales of *individual* products, while ρ represents the returns to scale of the production transformation function and affects the overall sales of *all* products. Thus, the firm's optimal decision on trading off the sales of different products within the firm helps identify η_n from ρ. In other words, the variation in the sales of a product relative to another product contains information on how the elasticities of the two products differ. This addresses the first challenge. Meanwhile, the identified relationship between elasticities reduces the number of parameters to be estimated in (33). Consequently, the number of instrumental variables required to estimate the rest of the parameters does not increase with the number of products. This addresses the second challenge.

To implement this idea, we take the ratio of (17) of a reference product (denoted as product 1) and that of each other product n produced by the same firm. Using $\tilde{R}_{jnt} = \tilde{P}_{jnt}Q_{jnt}$, we obtain:[30]

$$\ln(\tilde{R}_{j1t}) = c_n + \frac{1-\theta\frac{\eta_n}{\eta_n-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}\ln(\tilde{R}_{jnt}) + \zeta_{jnt}, \qquad n = 2, \ldots, N, \tag{35}$$

where

$$c_n = \frac{1}{1-\theta\frac{\eta_1}{\eta_1-1}}\ln\left[\frac{\eta_1}{\eta_1-1}\frac{\eta_n-1}{\eta_n}\right]$$

and

$$\zeta_{jnt} = \frac{\theta}{1-\theta\frac{\eta_1}{\eta_1-1}}\left[\underbrace{(\tilde{\omega}_{jnt} + \frac{1}{\eta_n-1}\tilde{\xi}_{jnt}) - (\tilde{\omega}_{j1t} + \frac{1}{\eta_1-1}\tilde{\xi}_{j1t})}_{\text{difference in TFPR}}\right].$$

The latter, ζ_{jnt}, contains the *difference* of the capability (or TFPR, $\tilde{\omega} + \frac{1}{\eta-1}\tilde{\xi}$, as will be formally defined in Section 6) of producing a product relative to that of the reference product. This equation predicts that the (logarithmic) revenues of two products are linearly related

[30]Note that this approach contrasts with much of the existing literature, which often relies on direct estimation of the demand function (28) using firm-level IVs (e.g., cost shifters). However, these IVs may be correlated with the *level* of quality, potentially biasing the results. When we estimate the demand function (28) directly using the same firm-level IVs, the resulting demand elasticities are biased downward: the mean elasticities are estimated at 2.539, −2.024, and 0.470 for the three industries, respectively. Of course, when appropriate instrumental variables exist, such as the case in Orr (2022), one can estimate demand function (28), or even a more flexible version of it, directly without resorting to this set of estimating equations.

conditional on the *difference* of production capability. In particular, firm-level inputs are not a part of the equation explicitly. This equation is similar to the estimating equation developed by Grieco et al. (2022), who explore the relationship of revenues of two markets (domestic sales and exports).[31]

Intuitively, because the demand for each product is segmented in our setting, as discussed in Section 4, the relative revenue of one product over another product in the same firm depends on their own demand elasticities (conditional on their relative levels of TFPR, ζ_{jnt}, as well as the estimated technological substitution parameter θ) rather than on complementarity or substitution between them on the demand side. As a result, the variation of one revenue *relative* to another in (35) provides the identification of the ratio, $\frac{1-\theta\frac{\eta_n}{\eta_n-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}$ for $n = 2, 3, \ldots, N$. In contrast, the variation of revenue *levels* in (33) identifies $\frac{(\eta_n-1)\rho}{\eta_n}$, $n = 1, 2, \ldots, N$. That is, the returns to scale parameter affects the sales of all products but not the relative relationship of sales between different products, while demand elasticities affect both the level and the relative relationship of sales of different products. As a result, ρ and η_n, $n = 1, 2, \ldots, N$, are separately identified as long as there are at least two products with different demand elasticities in the industry. More precisely, the elasticities and returns to scale parameter can be identified as long as there is a firm that manufactures two products with different demand elasticities for a number of periods, which is a very mild assumption. The model is over-identified when there are more than two such products produced by the firms in the industry.

To estimate (35), we choose the product produced by most firms in the industry as the reference product, in order to maximize the number of observations used in the estimation.[32] We treat ζ_{jnt} as an error term. We allow the mean of ζ_{jnt} to vary by product and year and use a set of flexible product-year dummies as controls (which also absorb c_n). ζ_{jnt} is likely correlated with $\tilde{R}_{jnt}$ – the revenue of product n is lower if the capability of producing n is lower than that of the reference product. We use a set of IVs to address the endogeneity issue. In our implementation, we use the same set of IVs used in estimating (27): the IV set consists of a constant and the logarithm of the wage rate, the capital stock, and the ratio of material expenditure to labor (as a proxy for material prices after conditional on wage rate). Grieco et al. (2022) uses a similar set of firm-level IVs to estimate an equation analogous to (35) in a two-product scenario. The same insight carries over in our context. These firm-level variables influence the *level* of revenue (i.e., $\tilde{R}_{jnt}$), but they are uncorrelated with the *difference* of capability (i.e., ζ_{jnt}) between two products. For example, conditional on everything else, a

[31] One difference is that Grieco et al. (2022) model the error term as an unexpected shock because the productivity and quality of the domestic and export products are assumed to be the same and are canceled.

[32] In our data, the percentage of firm-year pairs that produce the reference product ranges from 62% in the footwear industry to 72% in printing and 88% in the pharmaceutical industry.

higher level of capital stock potentially leads to higher revenues of a given product, but it is not necessarily associated with the production capability of one product being larger than that of another product within the same firm. We use these firm-level variables as IVs for all product pairs in (35).[33]

The validity of these IVs relies on the condition that the production of a product is not systematically more intensive in the use of a specific input (e.g., capital) than other products and that the firm-level wage rate and input price are not systematically correlated with the capability *differences* between products. We use Monte Carlo exercises to demonstrate the performance of our approach and IVs under this condition in Section 5.3. We further discuss this condition and alternative strategies in Online Appendix D.

We denote the estimated relationship between elasticities as $\hat{b}_n = \frac{1-\theta\frac{\eta_n}{\eta_n-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}$, $n = 2, \ldots, N$, and, naturally, $\hat{b}_1 = 1$ by definition. Thus, $\eta_n = 1 + \frac{\theta(\eta_1-1)}{\hat{b}_n+(1-\theta)(\eta_1-1-\hat{b}_n\eta_1)}$. Substitute it as η_n in (33) and solve for u_{jt} to construct moment conditions for the GMM estimation:[34]

$$u_{jt} = \ln\rho + \ln\left[\sum_{n\in\Lambda_{jt}} \frac{\theta(\eta_1-1)}{(\eta_1-1)+\hat{b}_n+(\theta-1)\hat{b}_n\eta_1}\tilde{R}_{jnt}\right] - \ln\left[E_{M_{jt}} + E_{L_{jt}}\left(1+\frac{\alpha_K}{\alpha_L}\left(\frac{K_{jt}}{L_{jt}}\right)^{\gamma}\right)\right]. \tag{36}$$

There are only four parameters, $\beta \equiv (\rho, \eta_1, \frac{\alpha_K}{\alpha_L}, \gamma)$, to be estimated. This means that the number of instrumental variables required does not increase with the number of products. Firm-level input choices can serve as valid IVs because they are not correlated with the unexpected shocks. In the implementation, we use $Z_{jt} = (1, E_{Mjt}, E_{Ljt}, L_{jt}, K_{jt}/L_{jt})$ as IVs.

Equation (36) can only identify $\frac{\alpha_K}{\alpha_L}$ rather than α_L, α_M, and α_K separately. The full set of $(\alpha_L, \alpha_M, \alpha_K)$ can be identified with two constraints naturally implied by the model. The first constraint is a normalization of distribution parameters in the CES production function:

[33] The model is over-identified if there is more than one IV. For example, if there are 2 IVs, then there are $2(N-1)$ moment equations that can be formed to identify $(N-1)$ coefficients (i.e., $\frac{\eta_1-1}{\eta_n-1}, n = 2, \ldots, N$).

[34] The identification of ρ relies on the condition $\mathbb{E}(e^{u_{jnt}}) = 1$, as assumed in Section 2.1. Due to the log-additivity of ρ in the main estimating equation (36), the value of ρ does not affect the estimation of the rest of the parameters. Thus, the rest of the parameters can be estimated before ρ is estimated. The following describes how ρ is estimated. Although the composite error term u_{jt}, defined in (34), does not have a zero mean (i.e., $\mathbb{E}(u_{jt}) \neq 0$), the composite error for single-product firms is the same as the product-level shock (i.e., $u_{jt} = u_{j1t}$). After the rest of the parameters are estimated, (36) can be written for these single-product firms as $\frac{e^{u_{j1t}}}{\rho} = \frac{\frac{\hat{\eta}_1-1}{\hat{\eta}_1}\tilde{R}_{j1t}}{\left[E_{M_{jt}}+E_{L_{jt}}\left(1+\frac{\hat{\alpha}_K}{\hat{\alpha}_L}\left(\frac{K_{jt}}{L_{jt}}\right)^{\hat{\gamma}}\right)\right]}$. Taking the expectation for both sides, we have $\frac{1}{\rho} = \mathbb{E}\left\{\frac{\frac{\hat{\eta}_1-1}{\hat{\eta}_1}\tilde{R}_{j1t}}{\left[E_{M_{jt}}+E_{L_{jt}}\left(1+\frac{\hat{\alpha}_K}{\hat{\alpha}_L}\left(\frac{K_{jt}}{L_{jt}}\right)^{\hat{\gamma}}\right)\right]}\right\}$. Therefore, the estimate of ρ can be recovered as $\hat{\rho} = 1/\mathbb{E}\left\{\frac{\frac{\hat{\eta}_1-1}{\hat{\eta}_1}\tilde{R}_{j1t}}{\left[E_{M_{jt}}+E_{L_{jt}}\left(1+\frac{\hat{\alpha}_K}{\hat{\alpha}_L}\left(\frac{K_{jt}}{L_{jt}}\right)^{\hat{\gamma}}\right)\right]}\right\}$.

$\alpha_L + \alpha_M + \alpha_K = 1$. The second constraint equalizes the ratio of geometric means of labor expenditure ($\overline{E}_L$) and material expenditure ($\overline{E}_M$) to the ratio of distribution parameters in the CES production function. That is, $\frac{\alpha_M}{\alpha_L} = \frac{\overline{E}_M}{\overline{E}_L}$. This constraint results from taking the geometric mean of (13), which is implied by the first-order conditions of labor and material quantities, (11) and (12), of all firms.[35]

As a result, β can be estimated as:

$$\begin{aligned} \hat{\beta} \quad &= \quad \text{argmin}_\beta \left[\frac{1}{\mathbb{N}} \sum_{j,t} u_{jt} Z_{jt}\right]' W \left[\frac{1}{\mathbb{N}} \sum_{j,t} u_{jt} Z_{jt}\right], \qquad (37) \\ \text{subject to:} \qquad & \alpha_L + \alpha_M + \alpha_K = 1 \quad \text{and} \quad \frac{\alpha_M}{\alpha_L} = \frac{\overline{E}_M}{\overline{E}_L}, \end{aligned}$$

where W is a weight matrix, $\mathbb{N}$ is the number of firm-time observations, and u_{jt} is the composite error term (36).

As a summary of the full estimating approach, the first step is to estimate $\hat{\theta}$ from (27) via 2SLS. The second step is to estimate $\hat{b}_n = \frac{1-\theta\frac{\eta_n}{\eta_n-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}$, $n = 2, \ldots, N$ via 2SLS using the relationship imposed by the within-firm relative sales in (35). The third step is to estimate $(\hat{\rho}, \hat{\eta}_1, \hat{\alpha}_L, \hat{\alpha}_M, \hat{\alpha}_K, \hat{\gamma})$ using (37) via GMM. With these estimates, the demand elasticities can be recovered as $\hat{\eta}_n = 1 + \frac{\hat{\theta}(\hat{\eta}_1 - 1)}{\hat{b}_n + (1-\hat{\theta})(\hat{\eta}_1 - 1 - \hat{b}_n \hat{\eta}_1)}$. After that, we compute $\tilde{\xi}_{jnt}$ and $\tilde{\omega}_{jnt}$ via (30) and (32), respectively.

5.3 Monte Carlo Exercise

In this section, we conduct a Monte Carlo exercise to demonstrate the performance of our estimation method. In the Monte Carlo setting, the choice of product sets is exogenous and random. The productivity and quality levels of each product are not only serially correlated over time but also negatively correlated with each other in the same period. Across products, productivity and quality exhibit different degrees of persistence and dispersion. Consequently, the levels and variability of productivity and quality differ systematically across products, generating heterogeneous revenue shares within a firm, thereby mimicking patterns observed in actual data. Wage rates, material prices, and capital stock are simulated as serially correlated and exogenous AR(1) processes. These variables are correlated with

[35] As shown by Grieco et al. (2016), this constraint holds conditional on a normalization of the CES production function. We follow the same procedure to normalize the inputs using their corresponding industry-level geometric means (e.g., Klump and de La Grandville, 2000; León-Ledesma et al., 2010). Nonetheless, to ease our notation, we directly denote the normalized input variables as (L_{jt}, M_{jt}, K_{jt}). As a result, the ratio of the geometric means of material and labor is $\frac{\overline{M}}{\overline{L}} = 1$, which implies $\frac{\alpha_M}{\alpha_L} = \frac{\overline{E}_M}{\overline{E}_L}$, by taking the geometric mean of (13) across firms.

contemporaneous input and output decisions because the firm observes their realized values before choosing inputs and outputs to maximize profit.

Table 3: Monte Carlo: Estimates of Production and Demand Function Parameters

Production			Demand		
Parameter	True	Estimate	Parameter	True	Estimate
α_L	0.400	0.400	η_1	3.000	3.022
		(0.005)			(0.257)
α_M	0.400	0.400	η_2	4.000	4.035
		(0.005)			(0.363)
α_K	0.200	0.200	η_3	5.000	5.022
		(0.010)			(0.403)
σ	2.000	2.000	η_4	6.000	6.016
		(0.023)			(0.398)
ρ	1.100	1.100	η_5	7.000	7.001
		(0.024)			(0.368)
θ	0.900	0.900			
		(0.004)			

Note: The parameter estimates are reported as the mean estimates from the Monte Carlo simulations. Standard errors in parentheses are the standard deviations of the estimates.

Our Monte Carlo exercise consists of 200 replications of simulated datasets for J firms observed over T periods, based on a model with a set of true parameters for five products: $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \alpha_L, \alpha_M, \alpha_K, \sigma, \rho, \theta)$. In each replication, we simulate productivity $(\tilde{\omega}_{jnt})$ and quality $(\tilde{\xi}_{jnt})$ for each product n, firm j, and period t, as well as wage rates $(P_{L_{jt}})$, material prices $(P_{M_{jt}})$, and capital stocks (K_{jt}) for each firm j and period t, using AR(1) processes with different persistence parameters and dispersion degrees of innovation shocks.

Given these variables, and the production and demand specifications in Section 5, we use the firm's static profit maximization problem to derive the optimal choices of labor and material inputs (L_{jt} and M_{jt}), as well as the optimal output quantity (Q_{jnt}) and price (P_{jnt}) for firm j and product n in each period t. The observed product price incorporates an idiosyncratic shock: $\tilde{P}_{jnt} = P_{jnt}e^{u_{jnt}}$, where u_{jnt} is a firm-product-time specific shock. Consequently, the observed product revenue is given by $\tilde{R}_{jnt} = \tilde{P}_{jnt}Q_{jnt}$.

The parameter values used in the data-generating process are summarized in Online Appendix Table A4. The variables used in the estimation procedure and the same set of instrumental variables as detailed in Section 5.2 are $(Q_{j1t}, \ldots, Q_{jnt}, \tilde{R}_{j1t}, \ldots, \tilde{R}_{jnt}, K_{jt}, L_{jt}, E_{L_{jt}}, E_{M_{jt}})$.

The simulated data exhibit realistic distributional patterns. First, productivity and quality are negatively correlated (coefficient: –0.2). Second, as shown in Online Appendix Table

A6, both the levels and heterogeneity of productivity and quality differ across products, reflecting technological and demand variation. Third, the mean within-firm revenue share varies across products, ranging from 6% to 57%, indicating differences in product importance within firms. Moreover, the dispersion of within-firm revenue shares differs substantially by product, suggesting heterogeneity in the relative importance of each product across firms.

Table 3 reports the mean estimates of the key parameters alongside their standard errors. In addition, the estimates for the parameters of (35) also closely match the true values, as shown in Table A5, demonstrating the effectiveness of the IVs proposed in Section 5.2. Overall, the results indicate that the estimation procedure successfully recovers the true parameters of the production and demand functions.

6 Estimation Results

This section reports the estimation results, including the production and demand function parameter estimates by industry as well as firm-product level productivity and quality. Because our empirical analysis relies on estimated variables, we employ bootstrapping with 100 samples to compute all standard errors presented in the subsequent tables.

Table 4: Production function estimates

Parameter	Footwear	Printing	Pharmaceutical
α_L	0.199	0.228	0.218
	(0.014)	(0.015)	(0.025)
α_M	0.763	0.676	0.574
	(0.039)	(0.027)	(0.068)
α_K	0.037	0.096	0.208
	(0.049)	(0.035)	(0.089)
σ	1.225	1.264	1.142
	(0.516)	(0.111)	(0.179)
ρ	1.054	1.129	1.037
	(0.146)	(0.123)	(0.196)
θ	0.950	0.779	0.720
	(0.053)	(0.066)	(0.082)

Note: Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (100 repetitions).

Table 4 presents the production parameters. α_M is significantly larger than α_L and α_K, consistent with the intensive use of intermediate material input across all industries. α_K in the pharmaceutical industry is the highest among the three industries, reflecting the importance of capital in this industry. Parameter σ, which is the elasticity of substitution across inputs,

i.e., labor, material, and capital, is greater than one across all industries. This is different from those in the classical literature which does not control for heterogeneous material prices. But it is largely consistent with the estimates in Grieco et al. (2016, 2022), Harrigan et al. (2021), and Li and Zhang (2022) based on a similar method but using different datasets from Colombia, France, and China, respectively. It is also close to the average estimate of the elasticity of substitution among Chinese industries by Berkowitz et al. (2017) using a different method. Furthermore, the returns to scale parameter ρ of the three industries is larger than one, but it is not significantly different from one, implying that production is close to constant returns to scale in these industries. Finally, the estimated values of θ range from 0.720 in the pharmaceutical industry to 0.950 in the footwear industry. Taking $\theta = 1$ as the benchmark case where products are perfectly substitutable in production, these estimates suggest that products in the footwear industry (e.g., men's vs. women's shoes) are considerably more substitutable in production than those in the pharmaceutical industry (e.g., antiparasitics vs. hormones).

Table 5 presents the estimated demand elasticity parameters for different products across the three industries. These estimates generally fall within the range reported in the existing literature (e.g., Roberts et al., 2018; Grieco et al., 2016; Dubois and Lasio, 2018). The estimated variation in demand elasticities across products implies meaningful heterogeneity in product-level markups. In the footwear industry, markups range from 1.218 to 1.400, while in the pharmaceutical industry they are significantly higher, ranging from 1.478 to 1.614. Because firms produce different sets of products in different years, we compute firm-year-level markups as weighted averages of product-level markups, using revenue shares as weights. Across the three industries, the average firm-year markup is 1.403, with a standard deviation of 0.102. This dispersion is smaller than the estimate reported by De Loecker and Warzynski (2012), which reflects a broader range of markup variation. Our measure of markup dispersion reflects only heterogeneity in product demand elasticities and composition across firms and years. Despite this narrower definition, the dispersion in firm-year markups is substantial.

The estimated quality and productivity also demonstrate substantial dispersion across firms, even conditional on a given product. However, if the objective is to compare technological production efficiency, the productivity measure (e.g., TFPQ) is not directly comparable across or within firms, as varieties within the same product category differ in quality levels. In contrast, quality-adjusted output is directly comparable across firms and products, as pointed out by Melitz (2000). Thus, we follow the literature (e.g., Orr, 2022; Li et al., 2025) to construct a revenue-based productivity (TFPR) measure that takes both quality and

Table 5: Demand function estimates

Parameter	Footwear	Printing	Pharmaceutical
η_1	4.009	4.128	2.965
	(1.620)	(1.030)	(1.166)
η_2	3.497	4.262	2.927
	(1.777)	(0.898)	(1.200)
η_3	4.263	3.699	2.998
	(1.780)	(1.192)	(1.070)
η_4	5.593	3.890	3.047
	(1.835)	(1.352)	(1.274)
η_5		4.276	2.911
		(0.881)	(1.219)
η_6		4.111	2.805
		(0.931)	(1.264)
η_7		3.787	2.923
		(1.184)	(1.158)
η_8		4.016	2.856
		(1.164)	(1.182)
η_9		4.210	2.878
		(0.909)	(1.195)
η_{10}		4.251	2.866
		(0.886)	(1.188)
η_{11}		4.004	2.926
		(1.056)	(1.282)
η_{12}		4.042	2.907
		(1.070)	(1.151)
η_{13}		4.123	3.176
		(0.968)	(1.395)
η_{14}		4.200	3.062
		(0.914)	(1.415)
η_{15}		4.147	2.628
		(0.952)	(1.334)
η_{16}			2.907
			(1.138)

Note: Elasticity parameters in each column within each industry are ordered to correspond with the column entries of products in Online Appendix Table A1, respectively. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (100 repetitions).

productivity into account:[36]

$$\text{TFPR}_{jnt} = \tilde{\omega}_{jnt} + \frac{1}{\eta_n - 1}\tilde{\xi}_{jnt}. \tag{38}$$

As expected, TFPR reflects significant dispersion across firms even within a specific product category.[37] The mean standard deviation within a product is 2.667 (calculated across all products in the three industries), which is similar in magnitude to that of revenue productivity documented by Grieco et al. (2022) in the Chinese paint industry. Regarding the components of TFPR, the standard deviation of $\tilde{\omega}_{jnt}$ within a product has a mean of 2.867, while the standard deviation of $\frac{1}{\eta_n - 1}\tilde{\xi}_{njt}$ within a product has a mean of 1.474.[38]

Interestingly, our results also reveal that within-firm heterogeneity in TFPR is substantial. Among multi-product firms, the average standard deviation of TFPR across products within a firm is 0.337, which is approximately one-eighth of the standard deviation of TFPR across firms for a given product. This indicates that while across-firm heterogeneity is more prominent, within-firm TFPR dispersion is also economically significant.

Overall, our estimation results reflect reasonable parameter estimates and productivity and quality measures at the firm-product level. In the following section, we turn to use these measures to explore the roles of productivity, quality, and within-firm resource reallocation in shaping firm performance.

7 Technological Spillovers & Within-firm Reallocation

A key advantage of our estimation method is that we do not need to impose any structure on the dynamic evolution of productivity. When the objective is to study potentially rich interdependencies in productivity dynamics—such as spillovers in the context of multi-product firms—we can estimate these processes *after* recovering the productivity measure itself rather than jointly estimating the interdependent dynamics with the model parameters. We illustrate this advantage by examining various forms of technological spillovers. In Section 7.1, we consider a conceptually novel within-firm, across-product spillover, which is particularly relevant for multi-product firms, in addition to the traditionally studied within-product, across-firm spillover (e.g., Malikov and Zhao, 2023). In Section 7.2, we show that within-firm reallocation of resources serves as an important mechanism through which multi-product firms are influenced by these technological spillovers.

[36] Note that quality enters TFPR as $\frac{1}{\eta_n - 1}\tilde{\xi}_{jnt}$. This is due to the demand specification in Section 5.1.

[37] The distributions of TFPR by product, as well as the distributions of its components, $\tilde{\omega}_{njt}$ and $\tilde{\xi}_{njt}$, are reported in Online Appendix Figures A2, A3 and A4, respectively.

[38] The standard deviation of $\tilde{\omega}_{jnt}$ is slightly larger than that of TFPR because the two components of TFPR, productivity, and quality, are negatively related, as will be clear in Section 7.

7.1 Technological Spillovers: Across-firm and Within-firm

In this section, we assess spillovers both across firms and within firms. In our context of multi-product firms, an across-firm spillover for a given product is defined as the impact of technical efficiency of the same product produced by other firms on that product's own technical efficiency. A within-firm spillover is defined as the impact of technical efficiency of other products produced by the same firm on the product in question.

Formally, we propose an evolution process for firm-product-level technical efficiency ω_{jnt}, following (6), which allows for both across-firm and within-firm spillover components:

$$\omega_{jnt} = g_1\omega_{jnt-1} + g_f\omega^f_{jnt-1} + g_p\omega^p_{jnt-1} + d_t + \epsilon_{jnt}, \tag{39}$$

where d_t is a time fixed effect and ϵ_{jnt} is an i.i.d. innovation shock. ω^f_{jnt-1} is the across-firm, within-product average technical efficiency: $\omega^f_{jnt-1} = \frac{1}{N^f_{nt-1}-1}\sum_{i\neq j}\omega_{int-1}$, where N^f_{nt-1} is the total number of firms producing product n in period $t-1$. Similarly, ω^p_{jnt-1} is the across-product, within-firm average technical efficiency: $\omega^p_{jnt-1} = \frac{1}{N^p_{jt-1}-1}\sum_{m\neq n}\omega_{jmt-1}$, where N^p_{jt-1} is the total number of products produced by firm j in period $t-1$. Both ω^f_{jnt-1} and ω^p_{jnt-1} vary at the firm-product-time level. Therefore, the term $g_f\omega^f_{jnt-1}$ captures across-firm, within-product spillovers in technical efficiency. Similarly, $g_p\omega^p_{jnt-1}$ captures across-product, within-firm spillovers—that is, the effect of changes in the average technical efficiency of a firm's other products in period $t-1$ on the technical efficiency of product n for firm j in period t.

Equation (39) can be interpreted as a multi-product extension of the single-product case in Malikov and Zhao (2023), in which we allow for the possibility of spillovers across products within firms. The lagged dependent variable captures persistence within product-firm pairs. In addition, modeling the evolution of technical efficiency rather than TFPQ makes it possible to mitigate differences across products due to different units of measurement.

While technical efficiency ω_{jnt} is not directly estimated in our procedure described in Section 5, it is shaped by two key estimated measures of heterogeneity: TFPQ ($\tilde{\omega}_{jnt}$) and quality ($\tilde{\xi}_{jnt}$), as linked through the TFPQ-quality tradeoff specified in (5). We do not impose such a trade-off in our estimation but after estimation the raw correlation between these two aspects of heterogeneity is negative, as shown in Online Appendix Figure A5. The emerging literature using firm-level data (e.g., Grieco and McDevitt, 2017; Atkin et al., 2019; Li et al., 2025) has documented a cost of quality: conditional on technical efficiency, producing higher-quality products raises marginal costs and therefore reduces measured TFPQ. This cost of quality can generate such a negative correlation between TFPQ and product appeal,

a pattern consistent with findings from the broader literature (e.g., Orr, 2022; Forlani et al., 2023; Eslava et al., 2024). We exploit this relationship to characterize technical efficiency ω_{jnt} and estimate its evolution process in (39).

Specifically, we adopt a linear version of the TFPQ-quality tradeoff (5):[39]

$$\tilde{\omega}_{jnt} = \omega_{jnt} - \gamma_\xi \tilde{\xi}_{jnt}, \tag{40}$$

where $\gamma_\xi \tilde{\xi}_{jnt}$ is interpreted as the cost (in terms of lowering productivity) of increasing quality, holding inputs fixed. γ_ξ is the elasticity of productivity with respect to the change in quality.

Replacing technical efficiency in (39) by that in (40) gives:

$$\tilde{\omega}_{jnt} = g_1 \tilde{\omega}_{jnt-1} - \gamma_\xi \tilde{\xi}_{jnt} + g_1 \gamma_\xi \tilde{\xi}_{jnt-1} + g_f \tilde{\omega}^f_{jnt-1} + g_f \gamma_\xi \tilde{\xi}^f_{jnt-1} + g_p \tilde{\omega}^p_{jnt-1} + g_p \gamma_\xi \tilde{\xi}^p_{jnt-1} + d_t + \epsilon_{jnt}, \tag{41}$$

where $\tilde{\omega}^f_{jnt-1}$ and $\tilde{\xi}^f_{jnt-1}$ are the across-firm, within-product average productivity and quality, respectively. Similarly, $\tilde{\omega}^p_{jnt-1}$ and $\tilde{\xi}^p_{jnt-1}$ are the across-product, within-firm average productivity and quality, respectively.[40]

Although all variables (except ϵ_{jnt}) are already estimated from our structural model, the innovation shock ϵ_{jnt} can be correlated with contemporaneous quality choice $\tilde{\xi}_{jnt}$. To address such an endogeneity problem, we estimate (41) via GMM using a set of instrumental variables that includes internal instruments in period $t-2$. According to the timing assumption, these variables are uncorrelated with the i.i.d innovation term ϵ_{jnt}.

The estimation results for various specifications of (41) are presented in Table 6. Column (1) reports estimates from a simplified specification that excludes both year fixed effects and spillover terms, while Column (2) adds year fixed effects. Column (3) introduces the across-firm spillover term, $g_f \omega^f_{jnt-1}$. Finally, Column (4) augments this specification by including an additional term, $g_p \omega^p_{jnt-1}$, which captures within-firm technological spillovers.

We treat Column (4) as our main specification for capturing the evolution of firm-product-level technical efficiency, allowing for a rich pattern of across- and within-firm technological spillovers and the trade-off between productivity and quality. As expected, technical efficiency is highly persistent, as indicated by the estimated value of g_1. Consistent with the literature, there is a negative trade-off between productivity and quality at the firm-product level: a 1-percent increase in quality lowers productivity by 0.340 percent, holding all else constant. The magnitude is comparable to the estimated productivity-quality trade-off elasticity of 0.2 in the U.S. healthcare industry (Grieco and McDevitt, 2017) and 0.5 in the Chinese

[39] Higher order terms of $\tilde{\xi}_{jnt}$ can be added to this relationship to allow for nonlinearity of the tradeoff.

[40] The within-product and within-firm averages are divided, respectively, by the number of firms and products minus one, since each average excludes its own observation from the sum.

Table 6: Productivity, cost of quality, and spillovers

Dep. var.: Productivity	(1)	(2)	(3)	(4)
g_1	0.905***	0.903***	0.884***	0.842***
	(0.035)	(0.035)	(0.031)	(0.030)
γ_ξ	0.129	0.131	0.337	0.340
	(0.201)	(0.202)	(0.287)	(0.242)
g_f			0.145***	0.137***
			(0.030)	(0.028)
g_p				0.056***
				(0.002)
Year FE	no	yes	yes	yes
Observations	4806	4806	4806	4806

Note: The dependent variable is quantity-based productivity at the firm-product-year level. The coefficients are estimated via GMM. The instrument set includes twice-lagged productivity and quality in all columns. In columns (3) and (4), the instrument set also includes the simple average of twice-lagged productivity and quality of the same product produced by other firms. In column (4), the instrument set also includes the simple average of twice-lagged productivity and quality of the other products produced by the same firm. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (100 repetitions). *** $p < 0.01$, ** $p < 0.05$.

steel-making industry (Li et al., 2025).

More importantly, the across-firm spillover of technical efficiency is economically significant. For a given product, a 1-percent increase in the average technical efficiency of this product produced by other firms raises the technical efficiency of this product by 0.137 percent. This magnitude is similar to that documented by Malikov and Zhao (2023), who report an across-firm spillover elasticity of 0.33 for the Chinese electric machinery manufacturing industry. In addition, our results reveal a within-firm spillover, with an elasticity approximately 40% of the magnitude of the across-firm spillover. This comparison suggests that while spillover effects are larger across firms, within-firm spillovers are also economically meaningful.

Our estimates of within-firm spillovers are closely related to economies of scope, an idea emphasized by Koike-Mori and Martner (2024), Argente et al. (2025), Ding (2025), and Khmelnitskaya et al. (2025). The importance of within-firm spillovers has also been noted for multinational firms when studying innovation and the role of intangible assets (see, e.g., Bilir and Morales, 2020; Merlevede and Theodorakopoulos, 2023). While intangible assets (ideas, knowledge) may diffuse more readily within firm boundaries than across them, they can also be rival inputs internally due to limits on managerial attention and information-processing

capacity.[41] Thus, whether within-firm or across-firm spillovers dominate is an empirical question that depends on the context. Our contribution is to establish the quantitative importance of within-firm spillovers—relative to across-firm spillovers—in a context with multi-product firms and substantial product-level heterogeneity in production efficiency. As we demonstrate in Section 7.2, product-specific shocks are amplified into firm-level and industry-level outcomes through internal resource reallocation, operating via both across-firm and within-firm spillovers.

7.2 Within-firm Reallocation in Response to Product-specific Shocks

What are the implications of an exogenous, product-specific shock for multi-product firms? Spillovers—both across and within firms—imply that the effects can be complex. Across-firm spillovers reflect the additional impact of a product-specific productivity improvement on all other firms that produce the same product, while within-firm spillovers imply that all other products within the same firm are also directly affected. These direct spillover effects on firm-level productivity are further amplified by within-firm resource reallocation: a productivity improvement in *any* product can trigger reallocation of resources across *all* products within the firm. In this section, we quantify the importance of across-firm and within-firm spillovers in terms of their impacts on social welfare and firm-level productivity via counterfactual analysis, and we highlight within-firm resource reallocation, which is not considered in the single-product firm literature, as a key mechanism through which these effects operate.

As a counterfactual exercise, we consider a 1 percentage point exogenous improvement in the technical efficiency of a representative product—denoted without loss of generality as product 1, the reference product produced by the most firms in an industry—in period t for all firms that produce this product.[42] Formally, we set $\omega'_{j1t-1} = \omega_{j1t-1} + 0.01$ for each firm that produces product 1, while holding the technical efficiency of other products unchanged: $\omega'_{jnt-1} = \omega_{jnt-1}$ for $n \neq 1$. According to the dynamics of technical efficiency in (39), this improvement in period $t-1$ directly affects the technical efficiency of product 1 in period t through the persistence term, $g_1\omega_{j1t-1}$. The across-firm direct spillover effect on all firms that produce product 1 is captured through the term $g_f\omega^f_{j1t-1}$, while the within-firm spillover effect on another product $n \neq 1$ of firm j operates through $g_p\omega^p_{jnt-1}$. Thus, this 1-percent improvement generates differential direct effects on products in period t:

[41] See Crouzet et al. (2022) for a discussion of potential rivalry in intangible capital within the firm.

[42] We focus on the short-term effects, holding all dynamic decisions (i.e., product quality, scope, and investment) described in Online Appendix C fixed. The overall long-term impacts of spillovers would likely be even larger, as firms adjust their dynamic decisions in response to spillovers. However, evaluating these long-term effects would require estimating a fully dynamic model, which we leave for future research.

$\Delta\omega_{j1t} = g_1 \times 0.01 + g_f \times 0.01$ for the reference product 1, and $\Delta\omega_{jnt} = g_p \times \frac{0.01}{N^p_{jt-1}-1}$ for other products $n \neq 1$, where N^p_{jt-1} is the number of products produced by firm j in period $t-1$.

These disproportionate impacts across products within a firm not only directly improve the productivity of individual products but also indirectly affect firm-level productivity, a traditional measure of firm performance, through within-firm resource reallocation towards more productive products in multi-product firms: firms allocate a larger share of resources to products whose productivity has improved more. Both the direct and indirect effects contribute to the improvement in firm-level TFPR. To assess the relative importance of these contributions, we aggregate firm-level TFPR from the underlying firm-product-level measures and decompose the overall effect into a direct impact and an indirect impact operating through within-firm reallocation.

In choosing an aggregation method, we follow the spirit of the standard production function estimation literature on multi-product firms, where productivity is typically allowed to vary only at the firm level, TFPR_{jt}, rather than at the firm–product level. This benchmark treats firms as multi-product producers but abstracts from within-firm productivity heterogeneity, so reallocation across products does not affect measured firm productivity. Our firm–product–level productivity measure TFPR_{jnt} is closely related to the firm-level concept, but by allowing for within-firm heterogeneity it enables us to study how resource reallocation across products shapes firm-level performance, which is a mechanism that is absent in single-product settings or in analyses that rely solely on firm-level productivity. Applying such a concept of firm-level productivity to our model setup in Section 5, Online Appendix E shows

that, this firm-level TFPR_{jt} and our measure of firm-product level TFPR_{jnt} are related as:[43]

$$e^{\text{TFPR}_{jt}} = \left\{ \sum_{\Lambda_{jt}} \left(s_{jnt} e^{-\text{TFPR}_{jnt}} \right)^{\theta} \right\}^{-\frac{1}{\theta}}, \tag{42}$$

where $s_{jnt} = \frac{\tilde{R}_{jnt}^{\frac{\eta_n}{\eta_n - 1}}}{\left\{ \sum_{\Lambda_{jt}} \left[\tilde{R}_{jnt}^{\frac{\eta_n}{\eta_n - 1}} \right]^{\theta} \right\}^{\frac{1}{\theta}}}$ is a weight based on the revenues of different products within the firm. Combining the definition of firm-product level TFPR in (38) and the TFPQ-quality trade-off in (40), TFPR_{jnt} can be expressed in terms of technical efficiency and quality: $\text{TFPR}_{jnt} = \omega_{jnt} + (\frac{1}{\eta_n - 1} - \gamma_\xi)\tilde{\xi}_{jnt}$.[44] The counterfactual exogenous shock affects TFPR_{jnt} directly through ω_{jnt}, while $\tilde{\xi}_{jnt}$ is held fixed.

Given the disproportionate changes in technical efficiency across products ($\Delta\omega_{jnt}$), each firm re-optimizes its profit by adjusting its input and product-level outputs. Let $\tilde{R}'_{jnt}$ denote the resulting product-level revenue in the counterfactual scenario. The corresponding counterfactual firm-level TFPR, computed using (42), is denoted by TFPR'_{jt}. The overall effect of a 1-percent improvement in the technical efficiency of the reference product is then measured by $(\text{TFPR}'_{jt} - \text{TFPR}_{jt})$.

To isolate the role of within-firm resource reallocation, we compute a firm-level TFPR using the updated firm-product-level TFPRs but holding the within-firm revenue shares s_{jnt} fixed at their original values. Denote this measure as TFPR^*_{jt}. The difference $(\text{TFPR}^*_{jt} - \text{TFPR}_{jt})$ reflects the direct impact of the technical efficiency improvement, while the difference $(\text{TFPR}'_{jt} - \text{TFPR}^*_{jt})$ captures the indirect impact operating through within-firm reallocation.

[43] This relationship is derived under the assumption that the firm produces multiple products, each with its own elasticity parameter in the CES demand function (28), while revenue productivity varies only at the firm level, not by product, as shown in Online Appendix E. A more conventional measure of firm-level TFPR follows the tradition of estimation methods that treat each firm as producing a single (aggregated) product. In that case, with a CES demand function for the aggregated product and elasticity parameter $\bar{\eta}$, firm-level TFPR and our firm–product-level TFPR are related by $e^{\text{TFPR}_{jt}} = \left\{ \sum_{\Lambda_{jt}} \left(s_{jnt}\, e^{-\text{TFPR}_{jnt}} \right)^{\theta} \right\}^{-\frac{1}{\theta}}$, where $s_{jnt} = \frac{\tilde{R}_{jnt}^{\frac{\eta_n}{\eta_n - 1}}}{\left[\sum_{\Lambda_{jt}} \tilde{R}_{jnt} \right]^{\frac{\bar{\eta}}{\bar{\eta} - 1}}}$ is a weight. Relative to the relationship in (42), the only difference lies in how the share s_{jnt} is constructed. We refrain from using this aggregation because the elasticity parameter $\bar{\eta}$ for the aggregated firm-level product is not a primitive object in our framework. Finally, a more straightforward measure of firm-level TFPR aggregates the firm-product-level measures using sales shares as weights: $\text{TFPR}_{jt} = \sum_{n \in \Lambda_{jt}} s_{jnt}\, \text{TFPR}_{jnt}$, where $s_{jnt} = \frac{\tilde{R}_{jnt}}{\sum_{\Lambda_{jt}} \tilde{R}_{jnt}}$ is the within-firm sales share of product n. Accordingly, the within-firm decomposition can be performed in the spirit of the within-industry, across-firm decomposition proposed by Olley and Pakes (1996). Our key result regarding the relative contributions of the direct and indirect impacts remains qualitatively similar under this alternative aggregation.

[44] Intuitively, this expression implies that while product quality promote TFPR, a sizable portion of its impact is offset by its cost, as documented in Li et al. (2025).

We conduct this decomposition for each firm and aggregate the results to the industry level using firm sales as weights.

We implement this decomposition for four distinct cases: (a) no spillover ($g^f = 0, g^p = 0$); (b) allowing for only across-firm spillover ($g^f = 0.137, g^p = 0$); (c) allowing for only within-firm spillover ($g^f = 0, g^p = 0.056$); and (d) allowing for both spillovers ($g^f = 0.137, g^p = 0.056$). In addition to the decomposition of TFPR improvement, we also evaluate the welfare implications by computing the change in total social welfare. Total welfare is defined as the sum of firm profits and consumer surplus. Given the demand functions in (28), consumer surplus is computed as $\sum_{j,n} \frac{\tilde{R}_{jnt}}{\eta_n - 1}$.

Table 7: Effects of 1-percent exogenous increase in technical efficiency of the reference product

	Spillover type			
	No	Across-firm	Within-firm	Both
Total welfare, million Pesos	1.543	1.799	1.626	1.881
	(0.327)	(0.347)	(0.327)	(0.347)
Firm-level TFPR, percentage	0.315	0.368	0.320	0.373
	(0.154)	(0.158)	(0.155)	(0.159)
– Direct impact	0.126	0.146	0.133	0.154
	(0.028)	(0.030)	(0.028)	(0.030)
– Within-firm reallocation	0.190	0.222	0.187	0.219
	(0.160)	(0.163)	(0.160)	(0.163)

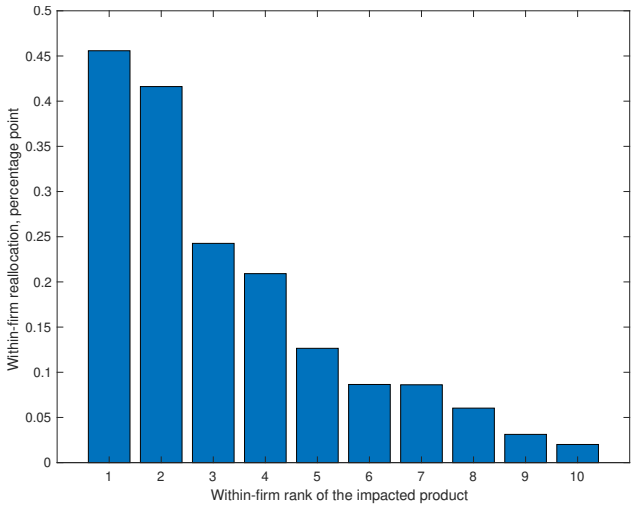
Note: The total welfare is measured as the sum of consumer surplus and producer surplus. A detailed decomposition is reported in Online Appendix Table A7. The TFPR improvement is measured in percentage point and calculated as the weighted average of the improvements in TFPR at the firm-year level with firms' total sales in the baseline scenario as weights. Bootstrapped standard errors, reported in parentheses, are computed based on 100 repetitions.

The results are presented in Table 7. A 1-percent increase in the technical efficiency of the reference product leads to an improvement in total welfare of 1.543 million Pesos in the absence of any technological spillover. When only across-firm spillover presents, the welfare gain increases to 1.799 million Pesos—approximately 16.6% higher than in the no-spillover case. In comparison, allowing only within-firm spillovers results in a 5.4% larger welfare improvement relative to the no-spillover baseline. When both spillover channels are active, the total welfare gain is approximately the sum of the net effects from the individual spillover scenarios. This comparison implies that both across-firm and within-firm technological spillovers are economically significant, though the former plays a more dominant role.

A similar pattern emerges in the impact on firm-level TFPR. Notably, within-firm resource

reallocation accounts for approximately 60% of the overall TFPR improvement across all four scenarios.[45] This highlights the importance of within-firm reallocation in shaping how multi-product firms benefit from productivity shocks. While a large literature emphasizes across-firm reallocation as a driver of aggregate productivity growth—showing that resources tend to flow toward more productive firms (e.g., Baily et al., 1992; Bartelsman and Doms, 2000; Baily et al., 2001; Aw et al., 2001; Foster et al., 2006, 2008; Syverson, 2011; Collard-Wexler and De Loecker, 2015)—our firm-product-level analysis demonstrates that within-firm resource reallocation makes a sizable contribution to the firm-level productivity growth.

Figure 1: Contribution of within-firm resource reallocation to TFPR growth



Notes: All firms producing more than 10 products are clustered in the "10" group.

Interestingly, within-firm resource reallocation plays an even greater role when a firm's top-selling products experience a productivity shock. This pattern is illustrated in Figure 1, which shows that the contribution of within-firm reallocation to firm-level TFPR improvement declines steadily with the within-firm rank of the impacted product.[46] Specifically, a 1 percentage point increase in the technical efficiency of a firm's top product (ranked 1st) results in a 0.4 percentage point improvement in firm-level TFPR attributable to within-firm reallocation. By contrast, when the same improvement occurs in the firm's least-selling product (ranked 10th), the contribution from reallocation falls to less than 0.05 percentage points. This result has important implications for the endogenous productivity dynamics of multi-product firms. In settings where firms make dynamic decisions about productivity investment, the relative sales performance of products within the firm may be a key determinant of where research efforts are directed—echoing insights from Kim (2024).

[45] This magnitude is based on the ratio of the indirect effect (Row 4) to the total effect (Row 2) across all columns of Table 7.

[46] A higher rank indicates a product further from the firm's top-selling product.

8 Conclusion

Multi-product firms account for a significant share of our economy. Yet, the traditional firm-level analysis in the literature masks the within-firm heterogeneity. In this paper, we propose a novel method to estimate firm-product-level productivity and quality along with demand and transformation function parameters. Compared with the existing methods in the literature, our method does not impose assumptions on how inputs are allocated across different products within firms, nor does it necessarily restrict how productivity evolves over time. This flexibility allows researchers to explore complex productivity dynamics after estimation. Importantly, the method can be easily scaled up to estimate production functions with a large number of products, without relying on the availability of productivity proxies. Finally, the method accounts for heterogeneous intermediate input prices that are usually unobservable to researchers and lead to biased estimation results when ignored.

We apply our method to three major industries in the Mexican manufacturing sector. Our findings reveal substantial heterogeneity in both quality and productivity—even when conditioning on a given product. Moreover, conditional on input usage, firms face a trade-off between quality and productivity. After accounting for this trade-off, we find that the underlying technical efficiency exhibits both across-firm and within-firm spillovers. This implies that an exogenous improvement in the technical efficiency of a single product not only affects the efficiency of other firms producing the same product, but also influences the efficiency of other products within the same firm. Notably, a large share of the resulting impact on firm-level TFPR is driven by within-firm resource reallocation. This highlights the quantitative importance of within-firm reallocation as a key mechanism through which multi-product firms enhance their performance. Consequently, this channel holds important implications for understanding aggregate productivity growth.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451.

Argente, D., S. Moreira, E. Oberfield, and V. Venkateswaran (2025). Scalable expertise: How standardization drives scale and scope. Technical report, National Bureau of Economic Research.

Atalay, E. (2014, June). Materials prices and productivity. *Journal of the European Economic Association 12*(3), 575–611.

Atkin, D., A. K. Khandelwal, and A. Osman (2019). Measuring Productivity: Lessons from Tailored Surveys and Productivity Benchmarking. *AEA Papers and Proceedings 109*, 444–449.

Aw, B. Y., X. Chen, and M. J. Roberts (2001). Firm-level evidence on productivity differentials

and turnover in taiwanese manufacturing. *Journal of Development Economics 66*(1), 51–86.
Aw, B. Y., M. Roberts, and D. Y. Xu (2011). R&d investment, exporting, and productivity dynamics. *American Economic Review 101*, 1312–1344.
Baily, M. N., E. J. Bartelsman, and J. Haltiwanger (2001). Labor productivity: structural change and cyclical dynamics. *Review of Economics and Statistics 83*(3), 420–433.
Baily, M. N., C. Hulten, D. Campbell, et al. (1992). Productivity dynamics in manufacturing plants. *Brookings Papers on Economic Activity 23*(1992 Microeconomics), 187–267.
Barrows, G., H. Ollivier, and A. Reshef (2024). Production function estimation with multi-destination firms. CESifo Working Papers.
Bartelsman, E. J. and M. Doms (2000). Understanding productivity: Lessons from longitudinal microdata. *Journal of Economic Literature 38*(3), 569–594.
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*(4), 735–747.
Berry, S. (1994, Summer). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics 25*(2), 242–262.
Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.
Bilir, L. K. and E. Morales (2020). Innovation in the global firm. *Journal of Political Economy 128*(4), 1566–1625.
Bond, S., A. Hashemi, G. Kaplan, and P. Zoch (2021). Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics 121*, 1–14.
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.
Caselli, M., A. Chatterjee, and A. Woodland (2017). Multi-product Exporters, Variable Markups and Exchange Rate Fluctuations. *Canadian Journal of Economics 50*(4), 1130–1160.
Caselli, M., L. Nesta, and S. Schiavo (2021). Imports and labour market imperfections: Firm-level evidence from France. *European Economic Review 131*, 103632.
Chen, Y., M. Igami, M. Sawada, and M. Xiao (2021). Privatization and productivity in china. *The RAND Journal of Economics 52*(4), 884–916.
Collard-Wexler, A. and J. De Loecker (2015). Reallocation and technology: Evidence from the us steel industry. *American Economic Review 105*(1), 131–71.
Crouzet, N., J. C. Eberly, A. L. Eisfeldt, and D. Papanikolaou (2022). The economics of intangible capital. *Journal of Economic Perspectives 36*(3), 29–52.
Das, S., M. J. Roberts, and J. R. Tybout (2007). Market entry costs, producer heterogeneity and export dynamics. *Econometrica 75*(3), 837–873.
De Loecker, J. (2011). Product differentiation, multi-product firms and estimating the impact of trade liberalization on productivity. *Econometrica 79, No. 5*, 1407–1451.
De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.
De Loecker, J. and F. Warzynski (2012). Markups and firm-level export status. *American Economic Review 102*(6), 2437–71.
Demirer, M. (2022). Production Function Estimation with Factor-Augmenting Technology:

An Application to Markups. mimeo.

Dhyne, E., A. Petrin, V. Smeets, and F. Warzynski (2022). Theory for extending single-product production function estimation to multi-product settings. Nber working paper, National Bureau of Economic Research.

Diewert, E., K. J. Fox, and L. Ivancic (2009). Scanner Data, Time Aggregation and the Construction of Price Indexes. UBC Discussion Paper 09-09, Department of Economics, University of British Columbia.

Ding, X. (2025). Industry linkages from joint production. *Working Paper*.

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

Doraszelski, U. and J. Jaumandreu (2016). Measuring the bias of technological change. *Journal of Political Economy (forthcoming)*.

Dubois, P. and L. Lasio (2018). Identifying industry margins with price constraints: Structural estimation on pharmaceuticals. *American Economic Review 108*(12), 3685–3724.

Eslava, M., J. Haltiwanger, and N. Urdaneta (2024). The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand, and Wedges. *The Review of Economic Studies 91*(1), 259–300.

Forlani, E., R. Martin, G. Mion, and M. Muûls (2023). Unraveling Firms: Demand, Productivity and Markups Heterogeneity. *The Economic Journal 133*(654), 2251–2302.

Foster, L., J. Haltiwanger, and C. J. Krizan (2006). Market selection, reallocation, and restructuring in the us retail trade sector in the 1990s. *The Review of Economics and Statistics 88*(4), 748–758.

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

Gandhi, A., S. Navarro, and D. A. Rivers (2020). On the identification of gross output production functions. *Journal of Political Economy 128*(8), 2973–3016.

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

Grieco, P., S. Li, and H. Zhang (2022). Input Prices, Productivity and Trade Dynamics: Long-run Effects of Liberalization on Chinese Paint Manufacturers. *The RAND Journal of Economics 53*(3), 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*(3), 1071–1105.

Harrigan, J., A. Reshef, and F. Toubal (2021). Techies, Trade, and Skill-Biased Productivity. CEPR Discussion Papers 15815, C.E.P.R. Discussion Papers.

Hottman, C. J., S. J. Redding, and D. E. Weinstein (2016). Quantifying the sources of firm heterogeneity. *The Quarterly Journal of Economics 131*(3), 1291–1364.

Khandelwal, A. K. (2010). The long and short (of) quality ladders. *Review of Economic Studies 77*, 1450–1476.

Khmelnitskaya, E., G. Marshall, and S. Orr (2025). Identifying Scale and Scope Economies using Product Market Data. *RAND Journal of Economics forthcoming*.

Kim, C. (2024). From Research to Development: How Globalization Shapes Corporate Innovation. mimeo.

Kirov, I. and J. Traina (2023). Labor Market Power and Technological Change in US Manufacturing. mimeo.

Klump, R. and O. de La Grandville (2000). Economic growth and the elasticity of subsitituion: Two theorems and some suggestions. *American Economic Review 90*(1), 282–291.

Koh, P. and D. Raval (2025). Economies of scope from shared inputs. *Working Paper*.

Koike-Mori, Y. and A. Martner (2024). Aggregating distortions in networks with multi-product firms. *Available at SSRN 5020563*.

Kumar, P. and H. Zhang (2019). Productivity or unexpected demand shocks: What determines firms'investment and exit decisions? *International Economic Review 60*(1), 303–327.

León-Ledesma, M. A., P. McAdam, and A. Willman (2010). Identifying the elasticity of substitition with biased technical change. *American Economic Review 100*(4), 1330–1357.

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies 70*(2), 317–341.

Li, J., S. Li, and H. Zhang (2025). Output Quality, Productivity, and Demand Advantage: Evidence from the Chinese Steel Industry. Working paper, University of New South Wales.

Li, S. (2018). A structural model of productivity, uncertain demand, and export dynamics. *Journal of International Economics 115*, 1–15.

Li, S. and H. Zhang (2022). Does External Monitoring from the Government Improve the Performance of State-Owned Enterprises? *The Economic Journal 132*(642), 675–708.

Malikov, E. and S. Zhao (2023). On the estimation of cross-firm productivity spillovers with an application to FDI. *Review of Economics and Statistics 105*(5), 1207–1223.

Mayer, T., M. J. Melitz, and G. I. Ottaviano (2021). Product mix and firm productivity responses to trade competition. *Review of Economics and Statistics 103*(5), 874–891.

Melitz, M. J. (2000). Estimating firm-level productivity in differentiated product industries. unpublished paper.

Merlevede, B. and A. Theodorakopoulos (2023). Intangibles within firm boundaries. *Working Paper*.

Morlacco, M. (2020). Market Power in Input Markets: Theory and Evidence from French Manufacturing. mimeo.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

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

Orr, S. (2022). Within-firm productivity dispersion: Estimates and implications. *Journal of Political Economy 130*(11), 2771–2828.

Powell, A. A. and F. Gruen (1968). The constant elasticity of transformation production frontier and linear supply system. *International economic review 9*(3), 315–328.

Pozzi, A. and F. Schivardi (2016). Demand or productivity: What determines firm growth? *RAND Journal of Economics 47*(3), 608–630.

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

Raval, D. R. (2019). The micro elasticity of substitution and non-neutral technology. *The RAND Journal of Economics 50*(1), 147–167.

Roberts, M., D. Y. Xu, X. Fan, and S. Zhang (2018). The Role of Firm Factors in Demand, Cost, and Export Market Selection for Chinese Footwear Producers. *Review of Economic Studies 85*(4), 2429–2461.

Rubens, M., Y. Wu, and M. Xu (2024). Exploiting or augmenting labor. Technical report,

Working Paper.

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

Valmari, N. (2023). Estimating production functions of multiproduct firms. *Review of Economic Studies 130*(11), 3315–3342.

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

Online Appendix

A Input Aggregator: Translog Functional Form

While Section 5 adopts a CES input aggregator for the empirical implementation, the method is not restricted to that functional form. In this appendix, we outline an estimation strategy when the input aggregator instead takes a translog functional form.

We maintain the setup of the demand system and output aggregator as (1) and (4), respectively. As a result, the parameters associated with the demand model and output aggregator are estimated in the same way as described in the paper. In particular, denote the estimated markup from the demand model as $\hat{\mu}_{njt}$ and denote the estimated parameter in the output aggregator as $\hat{\theta}$. In what follows, we focus on estimating the parameters specific to the translog input aggregator.

The input aggregator takes a full translog functional form:

$$\begin{aligned} F(L_{jt}, M_{jt}, K_{jt}) = \exp\Big\{ & \alpha_l \ln L_{jt} + \alpha_m \ln M_{jt} + \alpha_k \ln K_{jt} \\ & + \alpha_{kl} (\ln K_{jt})(\ln L_{jt}) + \alpha_{km} (\ln K_{jt})(\ln M_{jt}) + \alpha_{lm} (\ln L_{jt})(\ln M_{jt}) \Big\} \\ & + \frac{1}{2}\alpha_{ll} (\ln L_{jt})^2 + \frac{1}{2}\alpha_{mm} (\ln M_{jt})^2 + \frac{1}{2}\alpha_{kk} (\ln K_{jt})^2, \end{aligned} \tag{A.1}$$

where the input variables, (L_{jt}, M_{jt}, K_{jt}), are normalized by their geometric means (e.g., $\frac{1}{\mathbb{N}}\sum_{j,t} \ln L_{jt} = 0$) respectively. Such a normalization is analogous to the normalization conducted for the specification of the CES input aggregator, as described in Footnote 35.

Applying the methodology described in Section 3.2, we obtain a mapping from observable data to the unobservable variables:[47]

$$\xi_{jnt} = P_{jnt}^{-1}(\boldsymbol{P}_t, \boldsymbol{Q}_t), \tag{A.2}$$

$$\ln M_{jt} = \frac{\frac{E_{Ljt}}{E_{Mjt}}(\alpha_m + \alpha_{km} \ln K_{jt} + \alpha_{ml} \ln L_{jt}) - (\alpha_l + \alpha_{ll} \ln L_{jt} + \alpha_{kl} \ln K_{jt})}{(\alpha_{ml} - \alpha_{mm}\frac{E_{Ljt}}{E_{Mjt}})}, \tag{A.3}$$

and

$$e^{\theta \tilde{\omega}_{jnt}} = \frac{\mu_{jnt}}{P_{jnt}} \underbrace{\frac{P_{Ljt}}{F_L(L_{jt}, M_{jt}, K_{jt})}}_{\lambda_{jt}} Q_{jnt}^{\theta-1} [F(L_{jt}, M_{jt}, K_{jt})]^{1-\theta}, \tag{A.4}$$

[47] We assume that $\frac{\alpha_{ml}}{\alpha_{mm}} \neq \frac{E_{Ljt}}{E_{Mjt}}$, so the unique solution of M_{jt} exists.

where μ_{jnt} in (A.4) is the markup specified in (18), M_{jt} in (A.4) is the function presented as (A.3), and $F_L(L_{jt}, M_{jt}, K_{jt}) = \frac{\partial F(L_{jt}, M_{jt}, K_{jt})}{\partial L_{jt}}$. The above equations explicitly represent the mappings of quality (8), material quantity (14), and productivity (19) in the general model.

Consequently, the expression of the main estimating equation, (24), in this setup, is:

$$\begin{aligned} \ln\left[\sum_{n\in\Lambda_{jt}} \frac{\tilde{R}_{jnt}}{\mu_{jnt}}\right] &= \ln\left[\frac{E_{Ljt}+E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}+\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}}\right] - u_{jt} \\ &= -\ln\beta_0 + \ln[E_{M_{jt}} - \beta_1 E_{L_{jt}}] - \ln\left[1+\beta_2 \ln K_{jt} + \beta_3 \ln L_{jt}\right] - u_{jt}, \quad \text{(A.5)} \end{aligned}$$

where $\beta_0 = \alpha_m - \alpha_l\beta_1$, $\beta_1 = \frac{\alpha_{mm}}{\alpha_{ml}}$, $\beta_2 = (\alpha_{km} - \alpha_{kl}\beta_1)/\beta_0$, $\beta_3 = (\alpha_{lm} - \alpha_{ll}\beta_1)/\beta_0$, and $u_{jt} = \ln\left\{\sum_{n\in\Lambda_{jt}}\left[\frac{\tilde{R}_{jnt}/\mu_{jnt}}{\sum_{n\in\Lambda_{jt}}\tilde{R}_{jnt}/\mu_{jnt}} e^{-u_{jnt}}\right]\right\}$ is a firm-level composite error term.

As with the CES input aggregator in our empirical implementation, (A.5) alone does not identify all parameters of the translog function as the input aggregator. This is because some of the parameters are cancelled out when substituting the unobserved productivity and material input into the translog function using the mapping. Hence, additional conditions are required to identify all the translog parameters. Because the translog function is more flexible than the CES function, we explore both cross-sectional and time-series assumptions.

In particular, the first time-series assumption regards material prices. For demonstration purposes, we assume that material prices evolve exogenously according to an AR(1) process:

$$\ln P_{Mjt} = h_0 + h_1 \ln P_{Mjt-1} + \epsilon_{Mjt}, \quad \text{(A.6)}$$

where ϵ_{Mjt} is an i.i.d. shock.

The second time-series assumption regards technical efficiency. We only need to utilize the evolution of one product (e.g., reference product 1) within each firm. For demonstration purposes, we assume that the technical efficiency evolution of this reference product in (6) is independent of the evolution of other products in the same firm (i.e., abstracts away from spillovers):

$$\omega_{j1t} = g_0 + \omega_{j1t-1} + \epsilon_{j1t}. \quad \text{(A.7)}$$

The full estimation procedure is described by the following three steps.

Step 1: Estimate the revenue relationship (A.5).

We first estimate (A.5) via GMM with instrumental variables specified when estimating its general version (26) in Section 3.2. This provides an estimate of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$.[48] Another important output of the estimation is the fitted value of the right-hand side of (A.5), which

[48] We do not interpret the constant estimated from (A.5) as β_0 because the composite error term u_{jt} does not have a zero mean (i.e., $\mathbb{E}(u_{jt}) \neq 0$).

is denoted as $\hat{\Psi}_{jt}$:

$$\hat{\Psi}_{jt} \equiv \frac{E_{Ljt} + E_{Mjt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}} + \frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}} = \frac{E_{Mjt}}{\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}}} = \frac{E_{Ljt}}{\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}}}, \tag{A.8}$$

where the last two equalities come from the ratio of first-order conditions (13).

Thus, the elasticity of function F with respect to labor can be computed as $\hat{v}_{Ljt}$:

$$\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}} = \frac{E_{Ljt}}{\hat{\Psi}_{jt}} \equiv \hat{v}_{Ljt}. \tag{A.9}$$

Given the translog functional form (A.1), this elasticity can be written as:

$$\frac{\partial F_{jt}}{\partial L_{jt}}\frac{L_{jt}}{F_{jt}} = \alpha_l + \alpha_{ll} \ln L_{jt} + \alpha_{lm} \ln M_{jt} + \alpha_{kl} \ln K_{jt} = \hat{v}_{Ljt}. \tag{A.10}$$

Similarly, the elasticity of function F with respect to material can be computed as $\hat{v}_{Mjt}$:

$$\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}} = \frac{E_{Mjt}}{\hat{\Psi}_{jt}} \equiv \hat{v}_{Mjt}, \tag{A.11}$$

and this elasticity can be written as:

$$\frac{\partial F_{jt}}{\partial M_{jt}}\frac{M_{jt}}{F_{jt}} = \alpha_m + \alpha_{lm} \ln L_{jt} + \alpha_{mm} \ln M_{jt} + \alpha_{km} \ln K_{jt} = \hat{v}_{Mjt}. \tag{A.12}$$

Because the input variables are normalized by their geometric means (e.g., $\frac{1}{\mathbb{N}}\sum_{j,t} \ln L_{jt} = 0$) respectively, taking the average of (A.10) and (A.12) we obtain:

$$\hat{\alpha}_l = \frac{1}{\mathbb{N}}\sum_{j,t} \hat{v}_{Ljt}, \qquad \hat{\alpha}_m = \frac{1}{\mathbb{N}}\sum_{j,t} \hat{v}_{Mjt}, \tag{A.13}$$

where $\mathbb{N}$ is the total number of observations. These are the across-sectional restrictions that are analogous to the restriction imposed for the specification of the CES input aggregator in Section 5.2.

With these estimates, we can compute $\hat{\beta}_0$ as $\hat{\beta}_0 = \hat{\alpha}_m - \hat{\alpha}_l\hat{\beta}_1$.

Step 2: Estimate the evolution process of material prices (A.6).

Using (A.12), we can express $\ln M_{jt}$ as:

$$\ln M_{jt} = \frac{1}{\alpha_{mm}}(\hat{v}_{Mjt} - \hat{\alpha}_m - \alpha_{lm} \ln L_{jt} - \alpha_{km} \ln K_{jt}). \tag{A.14}$$

This is an equivalent expression of (A.3).

As a result, $\ln P_{Mjt}$ can be written as:

$$\ln P_{Mjt} = \left(\ln E_{jt} + \frac{1}{\hat{\beta}_1}\ln L_{jt}\right) + \frac{\hat{\alpha}_m}{\alpha_{mm}} - \frac{1}{\alpha_{mm}}\hat{\upsilon}_{Mjt} + \frac{\alpha_{km}}{\alpha_{mm}}\ln K_{jt}. \tag{A.15}$$

Therefore, we can estimate (A.6) via GMM using moment conditions

$$\mathbb{E}(\epsilon_{Mjt} Z_{Mjt}) = 0, \tag{A.16}$$

where $\epsilon_{Mjt} = \ln P_{Mjt} - h_0 - h_1 \ln P_{Mjt-1}$ and Z_{Mjt} is the instrument variables $Z_{Mjt} = (1, \ln E_{jt-1}, \ln L_{jt-1}, \hat{\upsilon}_{Mjt-1}, \ln K_{jt-1})$. Z_{Mjt} is uncorrelated with ϵ_{Mjt} because ϵ_{Mjt} is not in the information set of period t. The estimated parameters are $(\hat{h}_0, \hat{h}_1, \hat{\alpha}_{km}, \hat{\alpha}_{mm})$.

With these estimates, we can recover the following parameters using the estimate in step 1 as: $\hat{\alpha}_{ml} = \frac{\hat{\alpha}_{mm}}{\hat{\beta}_1}$, $\hat{\alpha}_{kl} = \frac{\hat{\alpha}_{km} - \hat{\beta}_0\hat{\beta}_2}{\hat{\beta}_1}$, and $\hat{\alpha}_{ll} = \frac{\hat{\alpha}_{ml} - \hat{\beta}_0\hat{\beta}_3}{\hat{\beta}_1}$.

Step 3: Estimate the technical efficiency evolution process of the reference product (A.7)**.**

We re-write the mapping (A.4) for the reference product as

$$\hat{\Phi}_{1jt} = e^{\tilde{\omega}_{1jt}} F(L_{jt}, M_{jt}, K_{jt}), \tag{A.17}$$

where $\hat{\Phi}_{1jt} \equiv \left[\frac{\hat{\mu}_{1jt}}{R_{1jt}}\frac{E_{Ljt}}{\hat{\upsilon}_{Ljt}}\right]^{\frac{1}{\theta}} Q_{1jt}$, which can be directly computed using the estimates from the previous steps.

Substitute the unobserved M_{jt} in (A.17) by (A.14), and reorganize the terms to obtain:

$$\ln F(L_{jt}, M_{jt}, K_{jt}) = \ln \hat{F}_{jt} + \gamma_k \ln K_{jt} + \tfrac{1}{2}\gamma_{kk}(\ln K_{jt})^2, \tag{A.18}$$

where $\gamma_k = \alpha_k - \frac{\hat{\alpha}_m \hat{\alpha}_{km}}{\hat{\alpha}_{mm}}$, $\gamma_{kk} = \alpha_{kk} - \frac{\hat{\alpha}_{km}^2}{\hat{\alpha}_{mm}}$, and $\ln \hat{F}_{jt} = \hat{\gamma}_0 + \hat{\gamma}_l \ln L_{jt} + \frac{1}{2}\hat{\gamma}_{ll}(\ln L_{jt})^2 + \hat{\gamma}_{lk}(\ln L_{jt})(\ln K_{jt}) + \frac{1}{2}\hat{\gamma}_{vv}(\ln \hat{\upsilon}_{Mjt})^2$ with coefficients $\hat{\gamma}_0 = -\frac{\hat{\alpha}_m^2}{2\hat{\alpha}_{mm}}$, $\hat{\gamma}_l = \hat{\alpha}_l - \frac{\hat{\alpha}_m \hat{\alpha}_{lm}}{\hat{\alpha}_{mm}}$, $\hat{\gamma}_{ll} = \hat{\alpha}_{ll} - \frac{\hat{\alpha}_{lm}^2}{\hat{\alpha}_{mm}}$, $\hat{\gamma}_{lk} = \hat{\alpha}_{kl} - \frac{\hat{\alpha}_{lm}\hat{\alpha}_{km}}{\hat{\alpha}_{mm}}$, and $\hat{\gamma}_{vv} = \frac{1}{\hat{\alpha}_{mm}}$.

Note that the only parameters unknown are γ_k and γ_{kk} in the right-hand side of (A.18) and $\ln \hat{F}_{jt}$ is directly computed from the data and the estimated parameters in the previous steps.

As a result, (A.17) can be used to solve $\tilde{\omega}_{1jt}$ as:

$$\tilde{\omega}_{1jt} = \ln \hat{\Phi}_{1jt} - \ln \hat{F}_{jt} - \gamma_k \ln K_{jt} - \tfrac{1}{2}\gamma_{kk}(\ln K_{jt})^2. \tag{A.19}$$

According to the evolution of technical efficiency (A.7) of the reference product and the cost of quality specification (40), we derive the explicit evolution process of technical efficiency

of the reference product as:

$$\tilde{\omega}_{j1t} = g_0 + g_1(\tilde{\omega}_{j1t-1} + \gamma_\xi \hat{\xi}_{jt-1}) - \gamma_\xi \hat{\xi}_{j1t} + \epsilon_{j1t}. \tag{A.20}$$

This equation can be estimated via GMM using moment conditions

$$\mathbb{E}(\epsilon_{1jt} Z_{1jt}) = 0. \tag{A.21}$$

Note that $\epsilon_{1jt} = \tilde{\omega}_{j1t} - g_0 - g_1(\tilde{\omega}_{j1t-1} + \gamma_\xi \hat{\xi}_{jt-1}) + \gamma_\xi \hat{\xi}_{j1t}$, where $\tilde{\omega}_{j1t-1}$ and $\tilde{\omega}_{j1t}$ are replaced by (A.19) for period $t-1$ and t, respectively. Z_{1jt} is the instrument variables $Z_{1jt} = (1, \ln \Phi_{1jt-1}, \ln F_{jt-1}, \ln K_{jt-1}, (\ln K_{jt-1})^2, \hat{\xi}_{j1t-1})$. Z_{Mjt} is uncorrelated with ϵ_{Mjt} because ϵ_{Mjt} is not in the information set of period t. The estimated parameters are $(\hat{g}_0, \hat{g}_1, \hat{\gamma}_k, \hat{\gamma}_{kk})$. With these estimates, the final parameters of the translog function parameters can be recovered as $\alpha_k = \hat{\gamma}_k + \frac{\hat{\alpha}_m \hat{\alpha}_{km}}{\hat{\alpha}_{mm}}$ and $\alpha_{kk} = \gamma_{kk} + \frac{\hat{\alpha}_{km}^2}{\hat{\alpha}_{mm}}$.

The choice of input aggregator functional form (CES vs. translog) depends on the empirical context, because each functional form has advantages and disadvantages. Although a CES aggregator is a more restrictive functional form, its estimation does not rely on the technical efficiency evolution process. This allows researchers to specify a more flexible or complex technical efficiency evolution *after* estimating the rest of the model as shown in the paper. In contrast, the translog aggregator offers greater flexibility in modeling inputs. However, estimating the translog function requires jointly estimating the technical efficiency evolution process and the translog parameters, which can limit the flexibility of the technical efficiency evolution specification in empirical applications.

B Multiple Materials Inputs

In the paper, we follow standard practice in the literature by assuming that each firm uses a single intermediate input in the production process. In reality, however, especially for multi-product firms, total intermediate input expenditures encompass a variety of goods that may differ both horizontally (e.g., rubber versus foam) and vertically (e.g., genuine leather versus synthetic leather). Unfortunately, most datasets report only total material expenditures, with no breakdown of types, prices, or quantities. This data limitation constrains researchers' ability to isolate the effects of individual inputs in the production process.

In this context, a key question is: under what conditions can our method accommodate the fact that firms use multiple (i.e., horizontally and vertically differentiated) material inputs, without requiring additional data? More specifically, can we still estimate the model parameters and recover productivity and quality at the firm-product level when only firm-level data on intermediate input expenditure (rather than input-type specific expenditures or more

disaggregated data) are available? The Online Appendix of Grieco et al. (2016) provides a positive answer in their context of single-product firms. The key assumption they need is that the effect of different intermediate inputs on production can be summarized through a homogeneous material index function. With this assumption, the production function parameters and thus productivity can be recovered even if the full vector of intermediate input expenditures is not directly observed. Such an idea can be directly applied to our context of multi-product firms. We present the details as follows.

Suppose a firm utilizes a vector of material inputs, $\mathbf{M}_{jt} = (M_{1jt}, M_{2jt}, \ldots, M_{Djt})$, in production. These inputs may include different input types and variations of the same input at different quality levels.[49] However, the researcher observes only the total expenditure on all materials, $E_{M_{jt}} = \sum_{d=1}^{D} P_{M_{djt}} M_{djt}$, rather than the quantity of each specific input, M_{djt}, or its corresponding price, $P_{M_{djt}}$.

We assume that these material inputs enter the transformation function as follows:

$$G(e^{-\tilde{\omega}_{jt}} \boldsymbol{Q}_{jt}) = F(L_{jt}, \tau(\mathbf{M}_{jt}), K_{jt}), \tag{B.1}$$

where $\tau : \mathbb{R}_+^D \to \mathbb{R}_+$ is an index function that aggregates the contribution of all material inputs to production.[50] We assume that $\tau(\cdot)$ is homogeneous of degree κ. As part of the production technology, the firm is assumed to know τ. Of course, without observing individual material inputs, we are not able to estimate the parameters associated within $\tau(\cdot)$, although the value of $\tau(\cdot)$ can be recovered. Our goal is to show how our method can be extended to such a context of multiple material inputs without estimating the parameters associated within $\tau(\cdot)$.

Note that this setup allows firms to use different material inputs to produce different products within the same firm, without explicitly modeling the allocation of each input to specific products. Such an assumption aligns with our broader modeling approach in the paper: rather than specifying separate production functions for individual products, we treat production as a transformation process. This approach enables us to account for input differentiation—whether across input types (horizontal differentiation) or quality levels

[49] An illustrative example in the Online Appendix of Grieco et al. (2016) is as follows. The material vector consists of three components: (M_1, M_2, M_3). M_1 and M_2 are vertically differentiated versions of the same type of input. While the quality for M_1 is normalized to be 1, the quality for M_2 is modeled as a scale parameter $\delta > 1$. M_3 is a component that is horizontally differentiated to the other two. For example, consider M_1, M_2, and M_3 as foam sole (lower quality), rubber sole (higher quality), and leather upper, respectively, for footwear industry. The material index function is modeled as: $\tau(M_{jt}) = \max\left(\left[M_{1jt}^{\gamma_1} + M_{3jt}^{\gamma_1}\right]^{1/\gamma_1}, \left[(\delta M_{2jt})^{\gamma_2} + M_{3jt}^{\gamma_2}\right]^{1/\gamma_2}\right)$.

[50] Some firms may not use certain inputs. The Online Appendix of Grieco et al. (2016) demonstrate how a discrete choice model of input selection can accommodate such cases.

(vertical differentiation)—without requiring an allocation rule for how each material input is assigned to each product. In the context that individual material inputs remain unobserved by researchers, this approach makes it possible to estimate the transformation function parameters and recover firm-product-level productivity. In the following, we demonstrate how this estimation is carried out.

As described in the paper, the firm's static optimization problem is now to choose L_{jt} and the vector $\mathbf{M}_{jt}$ to maximize the profit. In the setup with multiple material inputs, the Lagrange function is:

$$\begin{aligned}\mathcal{L}_{jt} &= \sum_{n\in\Lambda jt} P_{jnt}(\boldsymbol{Q}_{jt}, \boldsymbol{Q}_{-jt}; \boldsymbol{\xi}_t) Q_{jnt} - P_{Ljt}L_{jt} - \sum_{d=1}^{D} P_{Mdjt}M_{djt} \\ &\quad -\lambda_{jt}\left\{ G(e^{-\tilde{\boldsymbol{\omega}}_{jt}}\boldsymbol{Q}_{jt}) - F(L_{jt}, \tau(\mathbf{M}_{jt}), K_{jt}) \right\},\end{aligned} \tag{B.2}$$

where λ_{jt} is the Lagrangian multiplier.

The first-order conditions with respect to labor and each individual material inputs imply:

$$\lambda_{jt}\frac{\partial F(L_{jt}, \tau(\mathbf{M}_{jt}), K_{jt})}{\partial L_{jt}} = P_{Ljt}, \tag{B.3}$$

$$\lambda_{jt}\frac{\partial F(L_{jt}, \tau(\mathbf{M}_{jt}), K_{jt})}{\partial \tau_{jt}}\tau_d(\mathbf{M}_{jt}) = P_{Mdjt}, \qquad \forall d = 1, 2, \ldots, D. \tag{B.4}$$

where $\tau_d(\mathbf{M}_{jt}) = \frac{\partial \tau(\mathbf{M}_{jt})}{\partial M_{djt}}$.

Define a material price index as $P_{\tau jt} = \frac{E_{M_{jt}}}{\psi(\mathbf{M}_{jt})}$, where $\psi(\mathbf{M}_{jt}) = \sum_{d=1}^{D} M_{djt}\tau_d(M_{jt})$. Using this price index, the information in (B.4) can be summarized into a single equation by multiplying by M_{djt}, summing across d, and dividing it by $\psi(\mathbf{M}_{jt})$,

$$\lambda_{jt}\frac{\partial F(L_{jt}, \tau(\mathbf{M}_{jt}), K_{jt})}{\partial \tau_{jt}} = P_{\tau jt}. \tag{B.5}$$

This equation, together with (B.3), can be interpreted as the firm's first-order conditions, as if it were optimizing while facing a wage rate $P_{L_{jt}}$ and a material price index $P_{\tau_{jt}}$ for a single aggregated material input, represented by the quantity index $\tau_{jt} = \tau(\mathbf{M}_{jt})$. This is analog to the setup described in Section 3 for the baseline case where only a single material input is considered. Importantly, neither $P_{\tau_{jt}}$ nor τ_{jt} needs to be observable.

In this setting, total observed material expenditure is related to the material price index and the material quantity index via: $P_{\tau jt}\tau_{jt} = \frac{E_{M_{jt}}}{\kappa}$, where $E_{M_{jt}}$ is the total expenditure on materials, and κ is the degree of homogeneity of the function $\tau(\cdot)$. This relationship follows

directly from Euler's Theorem for homogeneous functions: $\sum_{d=1}^{D} M_{djt}\tau_d(\mathbf{M}_{jt}) = \kappa\tau(\mathbf{M}_{jt})$. In the special case where the firm uses a single material input, $\tau(\cdot)$ reduces to the identity function, which is homogeneous of degree 1, implying $\kappa = 1$.

Thus, we can treat τ_{jt} as analogous to M_{jt} in the single-input case. Thus, the estimation strategy described in Section 3 remains applicable, with one key modification: material expenditure E_{Mjt} is replaced by $\frac{E_{Mjt}}{\kappa}$, where κ serves as an additional scaling parameter.

The identification of κ depends on the specification of the input aggregator function. In some cases (e.g., a translog input aggregator), κ may not be separately identified from the production function parameters. In such cases, it can be normalized to 1 without loss of generality, as it is absorbed into the primary parameters of the production function. In other cases (e.g., a CES input aggregator), κ is identifiable through the revenue function, where it captures the returns to scale of the material aggregator index $\tau(\cdot)$. For example, in setup with the CES functional form of input aggregator as specified in Section 5, the main estimating equation (33) becomes:

$$\ln\left[\sum_{n\in\Lambda_{jt}} \frac{(\eta_n-1)\rho}{\eta_n}\tilde{R}_{jnt}\right] = \ln\left[\frac{E_{M_{jt}}}{\kappa} + E_{L_{jt}}\left(1+\frac{\alpha_K}{\alpha_L}\left(\frac{K_{jt}}{L_{jt}}\right)^{\gamma}\right)\right] - u_{jt}, \tag{B.6}$$

where

$$u_{jt} = \ln\left\{\sum_{n\in\Lambda_{jt}}\left[\frac{\frac{(\eta_n-1)}{\eta_n}\tilde{R}_{jnt}}{\sum_{n\in\Lambda_{jt}}\frac{(\eta_n-1)}{\eta_n}\tilde{R}_{jnt}}e^{-u_{jnt}}\right]\right\}. \tag{B.7}$$

After the model parameters are estimated, the firm-product level productivity and quality can be computed in the same way as specified in Section 5.

In summary, our method extends naturally to the multiple-material input setting (where the production involves different types or quality levels of material inputs) without requiring additional data, provided that the contribution of material inputs to production can be captured by a homogeneous materials index function in the transformation function rather than being modeled as inputs of individual products. While the specific functional form of this index function is not identified without additional information, this lack of identification does not hinder the recovery of the model's parameters or firm-product level productivity. In fact, neither the precise functional form nor the dimensionality of the index function needs to be explicitly specified for our estimation approach to remain valid.

C Firm's Dynamic Decisions

This appendix describes the dynamic decisions made by the firm as a completion of the full model. At the end of each period t, the firm chooses the set of products to produce, their

associated quality levels, and investment in technical efficiency improvement (e.g., research and development), for the next period $t+1$. These decisions are made conditional on the current state and after observing the adjustment costs of product scope and quality levels. Although the evolution of some state variables such as capital stock can be endogenous, we remain agnostic on modeling their exact evolution processes because our estimation method focuses on the static decisions and does not rely on how these variables evolve over time. The adjustment costs of product scope capture the costs incurred by the firm to install and arrange new production lines. The adjustment costs of product quality contain the costs of modifying the production procedure and sourcing new suppliers of the material input to meet the new quality levels.

In making decisions regarding product scope, quality levels, and investment, the firm is forward-looking and takes into account the impact of the current decisions on the future paths of the state variables. In particular, the firm knows that the choice of improving the quality of a product for the next period will reduce the associated (quantity-based) productivity in the next period (i.e., due to the cost of quality).

Although we do not estimate the full dynamic model in this paper—due to the high dimensionality of the state space—the model plays a crucial role in clarifying the firm's dynamic decision-making and serves as the conceptual foundation for the static model.[51] Specifically, while the firm's choices regarding product scope, quality levels, and technical efficiency are inherently endogenous in a dynamic setting, we treat them as predetermined and observed at the time the firm chooses inputs and outputs to maximize current-period profit. Our estimation method, presented in Section 3, relies on this assumption to establish the mapping from observed variables to unobserved productivity and quality.

D Discussion of the Instrumental Variables

Our strategy for estimating the relationship of demand elasticity parameters η_n in Section 5.2 exploits the within-firm relationship between the revenues of two products conditional on relative production capability (TFPR), implemented via an instrumental variable (IV) approach. This section discusses the validity of our IVs and outlines alternative methods for estimating η_n.

A key requirement for IV validity is a non-trivial degree of heterogeneity in production capability (i.e., TFPR) across firms for each product. If the dispersion in TFPR for a given product is extremely small, our identification strategy—relying on within-firm variation

[51] For instance, even in the footwear industry with only four products, the dynamic state includes at least 10 continuous variables: four for technical efficiency, four for product quality, and two for input prices (material and labor).

in revenues to estimate the elasticity relationship (35)—will fail. As an extreme example, consider an industry where firms produce a primary product (1) and a secondary product (2). If TFPR varies substantially across firms for product 1 but is constant for product 2, then in estimating (35), the firm-level inputs used as IVs would be correlated with the error term because ζ_{j2t} would be mechanically determined by the TFPR of product 1. Thus, a necessary empirical condition for our IV strategy is that all products exhibit sufficiently large TFPR heterogeneity. In our application to Mexican manufacturing industries, this condition is satisfied.

First, products at our level of aggregation display substantial across-firm variation in TFPR, as reflected in the dispersion of prices and sales across firms. In particular, the lowest standard deviation of log output prices for any product is 0.27 and the lowest standard deviation of log sales is 1.25. The TFPR dispersion estimates reported in Online Appendix Figure A2 further confirm that heterogeneity is sufficiently large for each product in our sample.

Moreover, within a firm, sales are not concentrated entirely in a single product. Online Appendix Table A2 and Online Appendix Figure A1 (discussed in Section 4) show that all products contribute non-trivially to firm-level revenues. For example, Online Appendix Table A2 reports average within-firm product shares by product scope: among firms producing five or more products, the average share of all products other than the top-selling one is 0.556, and the average share of products ranked fifth or lower is 0.147. Online Appendix Figure A1 shows the within-firm Herfindahl–Hirschman Index (HHI), where a lower value indicates greater diversification within a firm. The HHI falls to around 0.3 for firm-year pairs with five products and to about 0.2 for those with ten or more products, indicating that revenues are not dominated by their top product in these multi-products firms.

The insights from the above discussion also apply to the validity of the same IVs used in estimating θ from (27), as this estimation likewise exploits the within-firm relationship.

If the above heterogeneity condition does not hold, alternative approaches are available to estimate the demand elasticities. First, demand elasticities can, in principle, be identified directly from the demand function (28) using variation in prices and quantities, provided that suitable IVs uncorrelated with product quality are available. For instance, Orr (2022) estimates a demand system by constructing IVs that exploit variation in product sets and input price growth across firms operating in similar input markets but serving different output markets. In this case, one could directly estimate the demand function as discussed in Section 3 without relying on (35). Second, in the context where the assumption of constant returns to scale (i.e., $\rho = 1$) can be plausibly imposed, the demand elasticities can be identified from (33) alone, bypassing the need for the strategy of estimating (35). In this case, (33) simplifies

to the estimating equation used in Das et al. (2007), Aw et al. (2011), and Li (2018), which relates total variable cost (the counterpart to the right-hand side of (33)) to export revenues (the counterpart to the left-hand side of (33)) across multiple export markets for the same firm.

E Aggregating to Firm-level TFPR

The literature has a tradition of using revenue-based productivity (TFPR) as a measure of firm performance. While our framework yields a measure of TFPR at the firm-product level, these product-level measures can be aggregated when the interest is in evaluating overall firm performance. This appendix derives the aggregation.

We begin with our framework and impose the standard assumption implicitly assumed in the literature using firm-level data: the productivity of producing different products within the same firm is identical (i.e., a common firm-level productivity). Specifically, from the demand function (28), the revenue for product n can be written as

$$\tilde{R}_{jnt} = \tilde{P}_{jnt}Q_{jnt} = Q_{jnt}^{\frac{\eta_n-1}{\eta_n}} e^{\frac{1}{\eta_n}\tilde{\xi}_{jnt}} = \left[Q_{jnt}e^{-\tilde{\omega}_{jnt}}e^{\text{TFPR}_{jnt}}\right]^{\frac{\eta_n-1}{\eta_n}}, \tag{E.1}$$

where TFPR_{jnt} is defined in (38).

Rearranging this expression, we obtain:

$$\tilde{R}_{jnt}^{\frac{\eta_n}{\eta_n-1}} e^{-\text{TFPR}_{jnt}} = Q_{jnt}e^{-\tilde{\omega}_{jnt}}.$$

Raising both sides to the power θ, summing across all products $n \in \Lambda_{jt}$, and then taking the $1/\theta$ root yields:

$$\left\{\sum_{n\in\Lambda_{jt}}\left[\tilde{R}_{jnt}^{\frac{\eta_n}{\eta_n-1}} e^{-\text{TFPR}_{jnt}}\right]^\theta\right\}^{1/\theta} = \left\{\sum_{n\in\Lambda_{jt}}\left[Q_{jnt}e^{-\tilde{\omega}_{jnt}}\right]^\theta\right\}^{1/\theta} = F(L_{jt}, M_{jt}, K_{jt}), \tag{E.2}$$

where the second equality follows from the transformation function (3).

Given the common revenue-based productivity at the firm level, denoted TFPR_{jt}, we replace TFPR_{jnt} with TFPR_{jt} in the left-hand side to obtain:

$$\left\{\sum_{n\in\Lambda_{jt}}\left[\tilde{R}_{jnt}^{\frac{\eta_n}{\eta_n-1}}\right]^\theta\right\}^{1/\theta} e^{-\text{TFPR}_{jt}} = F(L_{jt}, M_{jt}, K_{jt}). \tag{E.3}$$

Therefore, comparing the two equations above, the firm-level TFPR can be related to

firm-product-level TFPR as:

$$e^{\text{TFPR}_{jt}} = \left\{ \sum_{n \in \Lambda_{jt}} \left(s_{jnt} e^{-\text{TFPR}_{jnt}} \right)^{\theta} \right\}^{-1/\theta}, \tag{E.4}$$

where

$$s_{jnt} = \frac{\tilde{R}_{jnt}^{\frac{\eta_n}{\eta_n - 1}}}{\left\{ \sum_{m \in \Lambda_{jt}} \left[\tilde{R}_{jt}^{\frac{\eta_m}{\eta_m - 1}} \right]^{\theta} \right\}^{1/\theta}} \tag{E.5}$$

is a weight that depends on the relative contribution of product n to firm j's aggregate revenue. For a single-product firm, this relationship degenerates to an identity equation.

F Additional Tables and Figures

Table A1: Product list by industry

Product name (product code)		
Footwear, leather (324001)	Printing and binding (342003)	Pharmaceutical products (352100)
Cow leather, for men (1)	Printing of calendars and almanacs (5)	Bactericides (11)
Cow leather, for women (2)	Folding boxes (6)	Antiparasitics (13)
Cow leather, for kids (3)	Notebooks and pads (7)	Dermatological (15)
Others (99)	Labels and prints (13)	Products with specific actions (19)
	Brochures and catalogs (14)	Circulatory system (21)
	Continuous forms (15)	Digestive system and metabolism (22)
	Accounting/admin/tax forms (16)	Musculoskeletal system (23)
	Telephone directories (17)	Respiratory system (24)
	Books (18)	Sensory organs (25)
	Journals (19)	Genitourinary system (26)
	Checks (21)	Blood and hematopoietic organs (27)
	Commemorative/business cards (23)	Central nervous system (28)
	Commercial flyers (24)	Hormones (32)
	Posters (25)	Vitamins and compounds (43)
	Others (99)	Non-therapeutic products (59)
		Others (99)

Table A2: Within-firm product shares by product scope

	Product rank (by sales level)				
Product scope	1	2	3	4	5+
1	1.000				
2	0.770	0.230			
3	0.670	0.240	0.090		
4	0.568	0.283	0.117	0.032	
5+	0.444	0.204	0.123	0.082	0.147

Note: All firm-year pairs producing 5 products or more are clustered in the "5+" group. All products ranked 5 or lower are clustered in the "5+" group.

Table A3: Descriptive statistics

Variable	Footwear	Printing	Pharmaceutical
Revenue per product (R)	70.890	30.713	104.376
	(106.830)	(75.803)	(211.836)
Number of workers (L)	248.722	160.826	465.352
	(383.053)	(157.036)	(492.768)
Labor expenditure (E_L)	14.532	18.238	92.608
	(30.514)	(22.983)	(112.902)
Material expenditure (E_M)	53.287	65.952	269.550
	(81.526)	(91.617)	(384.567)
Capital stock (K)	3.413	22.839	23.196
	(7.428)	(49.486)	(32.074)

Notes: The table reports the means and standard deviations (in parenthesis) for each variable by industry. R is revenues by product (1 million 2007 Mexican Peso, 1M MXN); L is the number of workers by firm, K is the capital stock by firm (1000 physical units); E_L is the expenditure on labor (wage bill) by firm (1M MXN); E_M is the expenditure on intermediates by firm (1M MXN).

Table A4: Monte Carlo parameter values

Parameter	Description	Value
N	Number of products	5
T	Number of periods	15
J	Number of firms	500
η_n, $n = 1, \ldots, 5$	Demand elasticity parameters	3, 4, 5, 6, 7
α_L	CES parameter of labor	0.4
α_M	CES parameter of material	0.4
α_K	CES parameter of capital	0.2
σ	Elasticity of substitution of inputs	2
ρ	Returns to scale parameter	1.1
θ	Substitution parameter of output	0.9
g_n^{ω}, , $n = 1, \ldots, 5$	Persistence parameters in productivity evolution	0.81 0.82 0.83 0.84 0.85
g_n^{ξ}, , $n = 1, \ldots, 5$	Persistence parameter in quality evolution	0.79 0.78 0.77 0.76 0.75
g^l	Persistence parameter in wage rate evolution	0.85
g^m	Persistence parameter in material price evolution	0.8
g^k	Persistence parameter in capital evolution	0.8
r	Productivity and quality shock correlation	-0.2
$sd(\varepsilon_n^{\omega})$, $n = 1, \ldots, 5$	S.D. of productivity shock	0.025 0.020 0.015 0.010 0.005
$sd(\varepsilon_n^{\xi})$, $n = 1, \ldots, 5$	S.D. of quality shock	0.025 0.020 0.015 0.010 0.005
$sd(\varepsilon^{\ell})$	S.D. of wage rate shock	0.1
$sd(\varepsilon^{m})$	S.D. of material price shock	0.1
$sd(\varepsilon^{k})$	S.D. of capital stock shock	0.1
$sd(u)$	S.D. of unexpected firm-product price shock (u_{jnt})	0.05

Table A5: Monte Carlo: Estimates of within-firm revenue relationship

	$\frac{1-\theta\frac{\eta_2}{\eta_2-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}$	$\frac{1-\theta\frac{\eta_3}{\eta_3-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}$	$\frac{1-\theta\frac{\eta_4}{\eta_4-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}$	$\frac{1-\theta\frac{\eta_5}{\eta_5-1}}{1-\theta\frac{\eta_1}{\eta_1-1}}$
True	0.571	0.357	0.229	0.143
Estimate	0.569	0.357	0.228	0.143
Standard error	(0.033)	(0.020)	(0.012)	(0.007)

Note: The estimates, for the parameters of (35), are reported as the mean estimates from the Monte Carlo simulations. Standard errors in parentheses are computed as the standard deviation of the estimates.

Table A6: Monte Carlo: distributional characteristics of key simulated variables

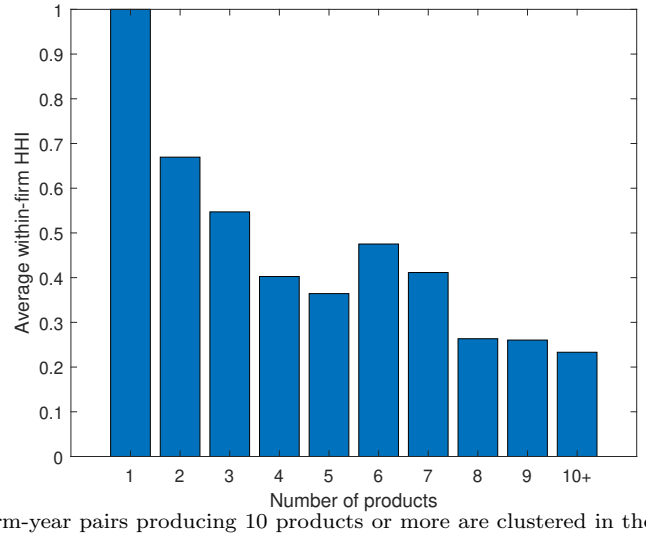
	Productivity				
	$\tilde{\omega}_1$	$\tilde{\omega}_2$	$\tilde{\omega}_3$	$\tilde{\omega}_4$	$\tilde{\omega}_5$
Mean	0.921	0.945	0.971	1.000	1.033
Std. deviation	0.040	0.033	0.025	0.017	0.009
	Quality				
	$\tilde{\xi}_1$	$\tilde{\xi}_2$	$\tilde{\xi}_3$	$\tilde{\xi}_4$	$\tilde{\xi}_5$
Mean	0.607	0.591	0.576	0.563	0.550
Std. deviation	0.039	0.030	0.022	0.015	0.007
	Within-firm revenue share				
	share_1	share_2	share_3	share_4	share_5
Mean	0.569	0.371	0.275	0.159	0.057
Std. deviation	0.179	0.091	0.105	0.100	0.070

Note: The reported means and standard deviations are calculated as the average and standard deviation of the key variables across Monte Carlo simulations.

Table A7: Welfare improvement of 1-percent increase of technical efficiency, million Pesos

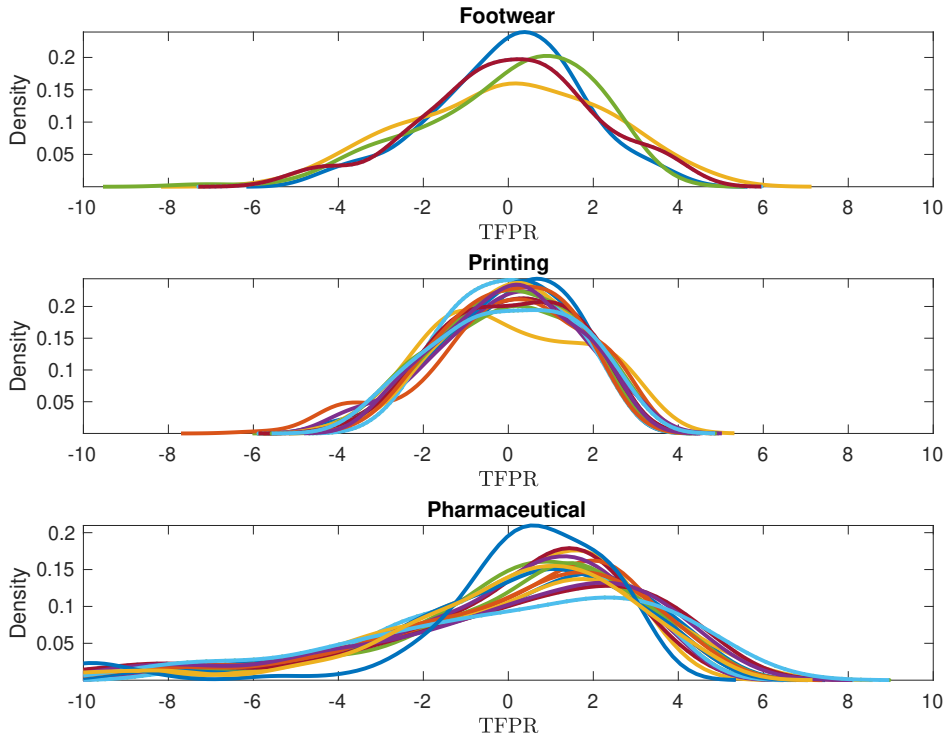
	No	Across-firm	With-firm	Both
Total welfare	1.543	1.799	1.626	1.881
Consumer surplus	0.845	0.984	0.889	1.029
Producer surplus	0.699	0.814	0.736	0.852

Figure A1: Weighted average within-firm HHI, by number of products



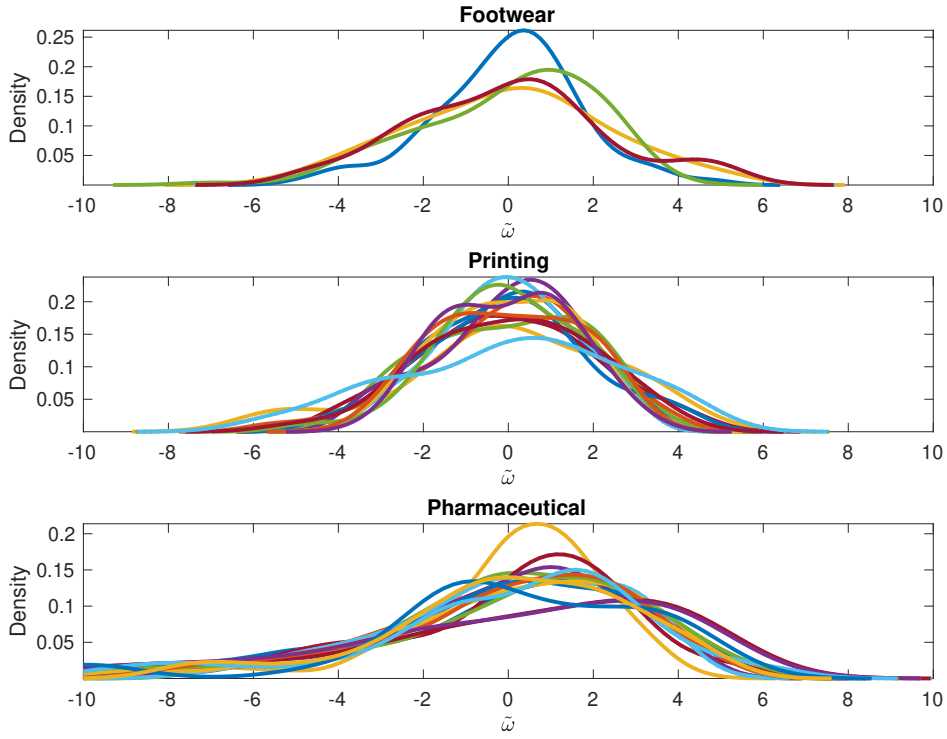
Notes: All firm-year pairs producing 10 products or more are clustered in the "10+" group. The weighted average is calculated using revenues as weights.

Figure A2: Distribution of TFPR



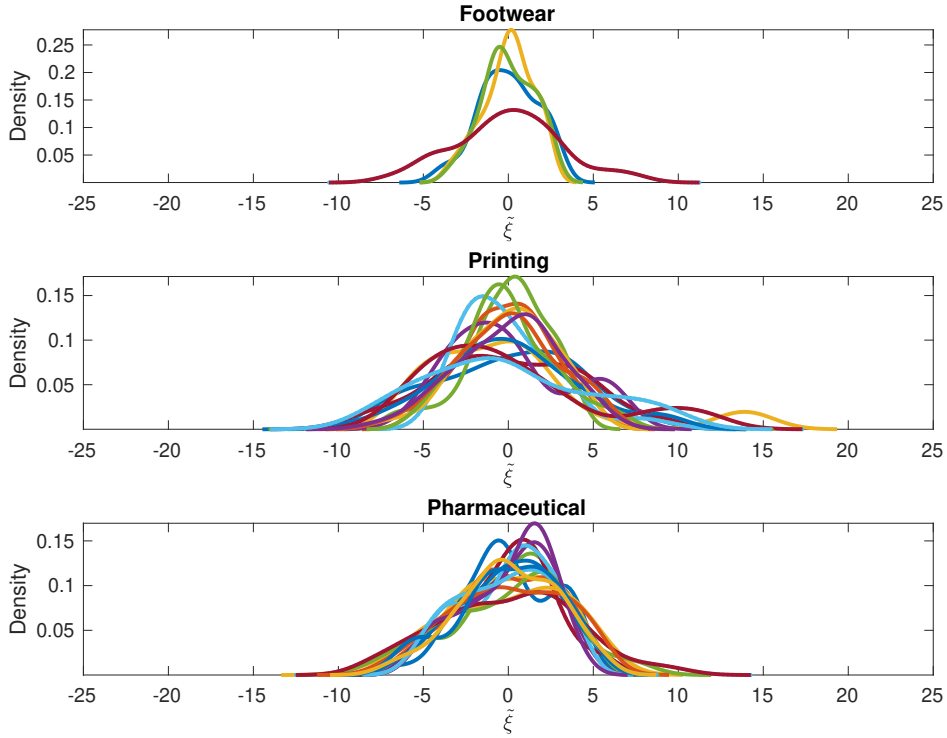
Notes: TFPR is demeaned, and only products with at least 100 observations are included.

Figure A3: Distribution of productivity, $\tilde{\omega}$



Notes: $\tilde{\omega}$ is demeaned, and only products with at least 100 observations are included.

Figure A4: Distribution of quality, $\tilde{\xi}$



Notes: $\tilde{\xi}$ is demeaned, and only products with at least 100 observations are included.

Figure A5: The relationship between productivity and quality