Productivity and Quality of Multi-product Firms[*]

Mauro Caselli[†], University of Trento
Arpita Chatterjee[‡], University of New South Wales
Shengyu Li[§], University of New South Wales

March 18, 2024

Abstract

This paper proposes a novel method to estimate productivity and quality of multi-product firms at the firm-product level, together with transformation function and demand parameters. The method utilizes firm optimization conditions to establish a one-to-one mapping between observed data and unobserved productivity and quality, offering distinct advantages: it eliminates the need for imputing firm-product input shares or imposing productivity evolution processes, while also exhibiting scalability to accommodate numerous products and the capability to address the bias caused by unobserved heterogeneous intermediate input prices. We apply this method to a set of Mexican manufacturing industries. We find that multi-product firms' better-performing products have both higher productivity and higher quality, with the former emerging as a stronger predictor of within-firm performance. However, firms face a trade-off between quality and productivity, which we refer to as the cost of quality. The cost of quality is higher for more differentiated products and declines with product age. In a counterfactual exercise, we show that a reduction in the cost of quality can lead to substantial firm-level productivity gains and that, on average, about 30.3 percent of these gains are due to the within-firm reallocation of production. Notably, a larger product scope allows more room for intra-firm resource reallocation, leading to higher productivity gains. This reveals a new mechanism for enhancing the performance of multi-product firms.

Keywords: *multi-product firms, production function, productivity, output quality, intra-firm reallocation*

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

[*]The authors thank Eleni Aristodemou, Zhiyuan Chen, Erwin Diewert, Kevin Fox, Andrea Fracasso, Moyu Liao, Matthias Mertens, Scott Orr, Ariell Reshef, Mark Roberts, Stefano Schiavo, Petr Sedlacek, Nikos Theodoropoulos, Andreas Tryphonides, Nelli Valmari, Eric Verhoogen, Frederic Warzynski, Daniel Xu, Hongsong Zhang, and seminar participants at the 2022 Australasian Trade Workshop, 2023 Firm and Industry Dynamics Workshop, 2023 Econometric Society Australasian Meeting, University of Melbourne, Monash University, University of New South Wales, University of Auckland, National University of Singapore, Singapore Management University, University of Neuchâtel, University of Cyprus, Indian Statistical Institute Delhi, Renmin University of China, CAFRAL, Australian Microdata reading group, Shanghai University of Finance and Economics, Zhejiang University, and Wuhan University for very helpful comments. All errors are the authors' responsibility.

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

[‡]School of Economics, Business School, the University of New South Wales, Australia. 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 performance and growth. This knowledge gap is mainly due to methodological limitations and data constraints, which hinder the accurate estimation of different aspects of heterogeneity at the firm-product level.

Addressing this critical gap and building upon recent studies (e.g., Dhyne et al., 2022; Orr, 2022; Valmari, 2022), this paper introduces an innovative method to estimate physical productivity and quality (product appeal) at the firm-product level, along with the transformation function and demand parameters. This method constructs a one-to-one mapping from observed data to unobservable variables by leveraging firm optimization conditions. This provides distinct advantages compared to the existing methods. First, it eliminates the need for imputing intra-firm input allocations or relying on productivity evolution processes. Second, it is scalable to handle a large number of products in monopolistically competitive markets. Third, it addresses the estimation bias caused by heterogeneous firm-level intermediate input prices, which are usually unobservable in commonly available data sets. Drawing on comprehensive firm-product-level data from three major industries in the Mexican manufacturing sector, we employ this method to study the trade-off between productivity and quality within firms, their relative importance in determining intra-firm revenue heterogeneity, and the role of product scope in shaping firm growth through intra-firm resource reallocation.

In modelling the production side, our method is designed to address the challenges commonly faced in estimating multi-product production functions. The recent strand of production function estimation methodologies implicitly assumes 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 produce different products and thus have different levels of productivity in these products, even within the same firm. 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). Moreover, researchers do not observe the within-firm division of inputs used to produce different products because firms usually only report total inputs. The potentially heterogeneous ability of input sharing (e.g., machinery and workers) across product lines within firms (e.g., Cairncross et al., 2023), which is observable to the firms but is unobservable to researchers, further complicates the problem.[1] Finally, intermediate input prices, which significantly vary across firms and over time due to various reasons such as quality differentiation, bargaining power in the input market, transport costs, and suppliers' marginal cost as documented by Atalay (2014) using US Census Bureau data, should be controlled for in the estimation to avoid biased estimates of production function and input elasticities (i.e., input price bias as emphasized in Ornaghi, 2006; De Loecker et al., 2016; Grieco et al., 2016). However, only input expenditure (rather than input price and input quantity) is observable to researchers at the firm level in commonly available data sets.

To address these issues, we model the production technology using a constant elasticity of substitution (CES) transformation function, which transforms inputs into different products. The inputs can be shared in production across products within the same firm. Each product is associated with a potentially different level of physical productivity (i.e., quantity-based productivity, as in Foster et al., 2008).[2] 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 always 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. We use the inverted relationship to substitute unobserved productivity to estimate the parameters of the transformation function. Once the parameters are estimated, we compute productivity at the firm-product level from the one-to-one mapping.

In modelling the demand side, we adopt a commonly used CES demand function. The firm's products are chosen from a set of horizontally differentiated categories. Within each product category, each firm's product variety is vertically differentiated according to its quality level. Because the optimal product prices are chosen after the firm's decisions on the product quality levels, we face a classic endogeneity problem in estimating the price elasticity of demand. The

[1] For example, a printing firm may use the same design software to create multiple products, such as logos and product labels; workers with specialized skills, such as pattern makers and shoe designers, may be used across different product lines within the same footwear firm; in pharmaceutical industries, a firm may use the same reactors and mixing tanks to produce different products, by adjusting the formulation and process parameters.

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

traditional solution is to use cost shifters (such as capital stock) as instrumental variables (IVs) for the price to estimate the demand function directly. However, if the cost of producing high-quality products is higher as suggested by the recent literature (e.g., Grieco and McDevitt, 2017; Forlani et al., 2023; Li et al., 2023), then cost shifters may still be correlated with quality. Thus, we depart from the existing literature to examine the advantage offered by intra-firm decisions in multi-product firms – profit maximization of the firm implies a relationship between the revenues of products within the same firm. This relationship depends on the demand elasticities and intra-firm relative quality *differences* (as opposed to quality *levels*), which can be instrumented by the commonly used firm-level cost shifters. Therefore, we use this relationship to help identify the demand elasticities. We use Monte Carlo exercises to demonstrate that this approach is able to recover the true parameters well. After the estimation, we compute quality as the residual of demand after controlling for price in the spirit of Khandelwal (2010).[3]

We apply our method to establishment-level panel production data from three major Mexican manufacturing industries (i.e., footwear, printing, and pharmaceuticals) that record prices and quantities at the firm-product level along with rich input data at the firm level. Multi-product production is an essential feature of the firms in our sample. Multi-product firms account for around 65 percent of the total number of firms and 86% of total revenues, and their average number of products is 6.7 per year, albeit with differences between industries. Within each industry, the markets for different products (e.g., women's shoes vs. men's shoes in the footwear industry) are largely segmented. Nevertheless, for each product, firms' output is vertically differentiated, as evidenced by the large dispersion in prices. These features are consistent with the model's assumption of a monopolistically competitive market structure with vertically differentiated products.

After estimation, we first follow the literature (e.g., Melitz, 2000) to construct a (firm-product) quality-adjusted productivity (ATFP) measure that accounts for heterogeneity in both productivity and quality. We find significant dispersion of ATFP across firms, even conditional on the product. More importantly, both components of ATFP (i.e., productivity and quality), are important for the within-firm performance of multi-product firms. Products closer to the core competence of the firm (defined by the highest revenue within firms) have both higher productivity and higher quality. However, differences in productivity across products within a firm turn out to be a

[3]Of course, the residual of demand is essentially demand heterogeneity which embodies a set of demand shifters. We leverage the rich fixed effects offered by the firm-product level data to refine the demand residual as a measure of quality in the empirical exercises. Nonetheless, we acknowledge that the refined measure of quality may still have different components, such as product appeal perceived by consumers, if they vary at the firm-product-time level.

stronger determinant of within-firm performance, as measured by product rank in terms of either level or growth of sales.

These different dimensions of within-firm heterogeneity are not, however, unrelated to each other. Within a firm, improving quality at the product level comes at the cost of reducing productivity. This result is broadly consistent with the emerging literature highlighting the trade-off across firms between these two unique dimensions of firm heterogeneity (e.g., Jaumandreu and Yin, 2014; Grieco and McDevitt, 2017; Roberts et al., 2018; Atkin et al., 2019; Orr, 2022; Eslava et al., 2023; Forlani et al., 2023; Li et al., 2023). Intuitively, producing one unit of a high-quality product may require more (or longer) production processes, better (or more specialized, exclusive) machinery, higher quality (or more) intermediate materials, and higher standards of quality control, all of which lead to a lower quantity of output holding inputs fixed and consequently an increase in *marginal cost* of production (or lower productivity, equivalently). After controlling for a rich set of fixed effects, offered by the advantage of using data at the firm-product level, as well as using different sets of IVs, we find that, on average, a 1 percent increase in quality reduces productivity by 0.198 percent, holding all other variables constant. Moreover, this trade-off is heterogeneous – it is stronger for more differentiated or younger products. This result suggests that, while it is more costly to produce a high-quality level of more differentiated products, long experience in producing a particular product allows the firm to improve quality with less sacrifice in efficiency.

Quantitatively, the cost of quality bears significant implications for firm productivity growth and intra-firm resource allocation. A reduction in the cost of quality not only directly increases the firm's ATFP but also indirectly influences it through the firm's endogenous reallocation of resources towards the production of higher-quality products. This is due to the positive relationship observed between ATFP and product quality within the firm (correlation coefficient: 0.440). In a counterfactual analysis, we find that a 1 percent reduction in the cost of quality corresponds to an average 2.562 percent improvement in firm-level ATFP. Notably, a substantial 30.3 percent of this improvement can be attributed to the within-firm reallocation of production towards high-quality, high-ATFP products.

Importantly, our findings reveal that the impact of the quality cost reduction on firm performance is particularly pronounced for multi-product firms with larger product scope. This is because a broader product scope provides these firms with greater flexibility for intra-firm resource reallocation, resulting in a higher gain in ATFP when the cost of quality decreases. This result uncovers a novel mechanism for productivity growth for multi-product firms, which dominate manufacturing production. Their ability to leverage a larger range of products through reallocation allows them to capitalize on the opportunities arising from reduced quality cost, thus boosting their overall productivity.

In terms of methodology, our paper builds on recent advances in the estimation of heterogeneity within 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 efficiency shocks, and shows how demand data can be used to assist estimation under profit maximization conditions. Valmari (2022) develops a similar framework, incorporating flexible production parameters across different product production functions. Chen and Liao (2022) generalize the previous papers by allowing single-product firms and multi-product firms to have different production functions and by estimating both non-parametric and parametric production functions for multi-product firms. 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 introduce a transformation production function and show how it can be used to recover the production frontier and estimate firm-product-specific marginal costs, taking into account complementarities and spillovers in multi-product production.

Our methodology integrates the strengths of both approaches to overcome their respective limitations. First, we model multi-product production using a transformation production function, similar to Dhyne et al. (2022). This avoids the need to allocate firm-level inputs, as in Orr (2022) and Valmari (2022), and allows for intra-firm input sharing across product lines, which may contribute to economies of scope in multi-product production. Second, in addressing unobserved firm-product productivity, we adopt the profit maximization assumption, similar to Orr (2022) and Valmari (2022). 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, ensure the scalability of our method. This differs from Dhyne et al. (2022), which requires a separate proxy for each additional firm-product-level productivity. 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, 2022), which typically require access to such data. Finally, our method does not rely on modelling 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 factors that endogenously shape the productivity trajectory (e.g., Chen et al., 2021) or frequent product turnover decisions, such as for exported products, where the observation of products is truncated by latent variables.

Regarding the application, our paper contributes to the literature that analyzes the role of productivity and quality in driving firm performance. Based on cross-firm analysis using firm-level data, a growing literature shows that demand is at least as important as productivity in explaining firm turnover and growth (e.g., Foster et al., 2008; Pozzi and Schivardi, 2016). By using data at the firm-product level, our paper is closely related to the literature on multi-product firms, which has long focused on cost (productivity) and quality as determinants of within-firm relative performance, measured as the sales rank of products (e.g., Berman et al., 2012; Chatterjee et al., 2013; Mayer et al., 2014, 2021; Eckel et al., 2015; Arkolakis et al., 2021). To our knowledge, an empirical investigation of the relative merits of these different determinants of within-firm performance is lacking in the literature due to methodological limitations in estimating productivity and quality at the firm-product level. We complement the literature by uncovering rich dimensions of within-firm heterogeneity and showing that productivity plays a more important role in shaping heterogeneous revenue and growth within firms.

Our paper also contributes to the recent literature analyzing the trade-off between productivity and quality (i.e., the cost of quality). Focusing on the healthcare industry, Grieco and McDevitt (2017) show that reducing the quality standards of a healthcare center can increase its patient load. Atkin et al. (2019) reveal a reverse correlation between quantity productivity and quality productivity among rug-makers in Egypt, drawing insights from data that include direct quality assessments. Forlani et al. (2023) document a strong negative correlation between demand and quantity-based productivity in various Belgian manufacturing industries. Using an objective measure of output quality, Li et al. (2023) find that about half of the benefits of quality are offset by the cost of producing the quality in the Chinese steel industry. These papers document such a trade-off across firms. Our paper advances the finding by showing a similar trade-off at the firm-product level. To this end, our analysis is consistent with the negative relationship between productivity and "product appeal" documented at the same level of disaggregation by Orr (2022).

We further show that the cost of quality is heterogeneous across degrees of product differentiation and varies with product age. Nonetheless, after taking both the cost and the benefit of quality into account, ATFP is documented to be positively correlated with quality. This result is consistent with endogenous quality choice models (Kugler and Verhoogen, 2009, 2012) and empirical analysis using an objective quality measure by Li et al. (2023).

Finally, our paper is related to a large literature on resource reallocation, which focuses on across-firm analyses and shows that much of the aggregate productivity growth is attributable to the resource reallocation towards 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). Complementing the literature, our counterfactual analysis shows that there can be a substantial contribution to productivity growth due to within-firm resource reallocation – a mechanism that is emphasized in the recent literature studying multi-product firms (e.g., Mayer et al., 2021). Importantly, we focus on the channel of the cost of quality and document a positive relationship between product scope and the contribution of intra-firm resource reallocation. This result illustrates the quantitative importance of intra-firm resource reallocation within multi-product firms due to quality differences as a novel channel affecting overall productivity at the firm level.

In the rest of the paper, Section 2 develops a framework to describe the production and demand functions and the firm's endogenous decisions on output, quality, product scope, and investment. Section 3 describes the methodology and steps for estimating the production and demand functions. Section 4 describes the data used in the estimation. Section 5 presents the estimation results. Section 6 documents the trade-off between productivity and quality and examines their role in shaping intra-firm heterogeneity. Section 7 quantitatively illustrates the significance of the cost of quality and the role of product scope in intra-firm resource reallocation using a counterfactual exercise. We conclude in Section 8.

2 Model

This section develops a framework to describe the firm's static production decisions. Our empirical estimation utilizes the static optimization conditions implied by this framework. We also sketch out firms' dynamic choices with respect to output quality, product scope, and investment, which provide conceptual insights regarding how these dynamic choices are endogenously determined.

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_{jtn} and a level of quality ξ_{jtn}, 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 a capital investment decision.

The firm's static decisions in the current period consist of the material input, labor input, and quantities of individual products to maximize its total period profit subject to demand and production functions, after observing the material price, wage rate, and capital stock. At the end of period t, the firm makes dynamic decisions on the capital stock and the set of products to be produced with their levels of product quality and technical efficiency for the following period, after observing the associated adjustment or investment costs.

2.1 Demand

The entire set of products that the firm can choose to produce is divided into N horizontal categories, such as women's and men's shoes. For each product category $n \in \{1, 2, \ldots, N\}$, the output of each firm is vertically differentiated according to its choice of quality level Ξ_{jtn}. This means that, although the demand for each of the N product categories is segmented, there is monopolistic competition across firms that produce vertically differentiated products in the same category. This assumption is also adopted by De Loecker (2011) and Valmari (2022) in modelling the demand functions in the multi-product context.[4]

Specifically, for each product category n, a representative consumer has constant elasticity of substitution (CES) preferences in terms of both the quality and the quantity of the products offered by firms:[5]

$$U_{tn} = \left[\sum_j \left(\Xi_{jtn}^{\frac{1}{\eta_n - 1}} Q_{jtn} \right)^{\frac{\eta_n - 1}{\eta_n}} \right]^{\frac{\eta_n}{\eta_n - 1}}, \tag{1}$$

where $\eta_n > 1$ is the elasticity of substitution across the varieties offered by the firms. Q_{jtn} is the physical quantity and Ξ_{jtn} is the product quality produced by firm j in period t, respectively.

[4] Orr (2022) allows for a more flexible demand structure with cannibalization across products. Nonetheless, at the level of product classification of our data, markets of different products are largely segmented, as will be discussed in Section 4. Thus, we abstract away from the across-product competition. This also implies that our setup is suitable to model firms manufacturing products for different (or segmented) markets, such as exporters selling to different destination markets.

[5] The power of Ξ_{jtn}, $\frac{1}{\eta_n - 1}$, is used to simplify the notation to reach a commonly used demand function (2). A large literature that treats demand residual as output quality implicitly shares the same setup (e.g., Melitz, 2000; Khandelwal, 2010; Pozzi and Schivardi, 2016; Valmari, 2022).

That is, the consumer values the quality-adjusted quantity of the product, $\Xi_{jtn}^{\frac{1}{\eta_n - 1}} Q_{jtn}$, which forms the basis for constructing the quality-adjusted productivity in Section 5.

Given the consumer's total expenditure B_{tn} and the product price P_{jtn}, the consumer's utility maximization problem implies the following demand function for product n from firm j:

$$\ln Q_{jtn} = -\eta_n \ln P_{jtn} + \xi_{jtn} + \phi_{tn} + \psi_{jn} + v_{jt}, \tag{2}$$

where $\xi_{jtn} = \ln \Xi_{jtn}$. Intuitively, a higher quality level shifts the demand curve upwards. Beyond quality, three other components also influence demand. First, $\phi_{tn} = \ln \left(\frac{B_{tn}}{\sum_j \Xi_{jtn} P_{jtn}^{1-\eta_n}} \right)$ is a product-specific expenditure index that depends on macroeconomic conditions captured in B_{tn} such as consumer income and market size in period t. Second, ψ_{jn} represents factors that affect demand at the firm-product level but do not vary over time such as consumers' subjective tastes, brand image related to specific products, number (or variety) of subcategories contained in each product category under our classification and product measurement units (e.g., grams vs. liters).[6] Finally, v_{jt} captures the demand heterogeneity such as firm effort in marketing that varies by firm and year. For the purpose of demonstration, we summarize the structural terms that shift the demand function as $\tilde{\xi}_{jtn} = \xi_{jtn} + \phi_{tn} + \psi_{jn} + v_{jt}$. The firm observes $\tilde{\xi}_{jtn}$ for all products before making production decisions.

Remark: Essentially, $\tilde{\xi}_{jtn}$ is a demand shifter, which captures all sorts of demand heterogeneity that influences product demand but is not accounted for by product prices. Empirically, $\tilde{\xi}_{jtn}$ is usually referred to as "perceived product appeal/demand" (e.g., Pozzi and Schivardi, 2016; Orr, 2022; Valmari, 2022; Eslava et al., 2023) or "quality" (e.g., Melitz, 2000; Khandelwal, 2010; Hottman et al., 2016). In our paper, we follow this tradition of notation and acknowledge that it embodies quality (ξ_{jtn}) as well as non-quality components, such as consumer tastes, brand/firm image, marketing efforts and market size. Yet, our setting with multiple-product firms provides us with a rich set of fixed effects at the product-year (ϕ_{tn}), firm-product (ψ_{jn}), and firm-year (v_{jt}) levels to control for the non-quality component that varies at these levels. For this reason, we define $\chi_{jnt} = \phi_{tn} + \psi_{jn} + v_{jt}$ and refer to χ_{jnt} as a demand shock in this paper. Notably, this advantage is not available in the traditional across-firm analysis (i.e., using firm-level data), and thus it helps to tease out a finer measure of quality (i.e., ξ_{jtn}) from residual demand (i.e., product appeal, $\tilde{\xi}_{jtn} = \xi_{jtn} + \chi_{jnt}$) that is traditionally used as quality.

[6] Units of measurement can be different across product categories. Consequently, the quantities and prices of different product categories are not readily comparable. In the demand function (2), ψ_{jn} absorb such differences. Similarly, in our empirical analysis in Section 6, we use firm-product dummies to tease out ξ_{jtn} from such differences.

2.2 Production Technology

We use a transformation function to model the production technology. Specifically, given the set of products to be produced (Λ_{jt}) and associated product appeal ($\tilde{\xi}_{jtn}$, $n \in \Lambda_{jt}$), the firm uses labor (L_{jt}), material (M_{jt}), and capital (K_{jt}) to produce output quantity (Q_{jtn}, $n \in \Lambda_{jt}$) following a constant elasticity of substitution (CES) transformation function:[7]

$$\sum_{n \in \Lambda_{jt}} e^{-\tilde{\omega}_{jtn}} Q_{jtn} = F(L_{jt}, M_{jt}, K_{jt}) \equiv \left[\alpha_L L_{jt}^{\gamma} + \alpha_M M_{jt}^{\gamma} + \alpha_K K_{jt}^{\gamma}\right]^{\frac{\rho}{\gamma}}, \tag{3}$$

where $\tilde{\omega}_{jtn}$ is the Hicks neutral, quantity-based productivity (i.e., so-called physical productivity, or TFPQ) of firm j in producing product n in period t. In this paper, we use quantity-based productivity, TFPQ, and productivity interchangeably. $\gamma \equiv \frac{\sigma-1}{\sigma}$ governs the elasticity of substitution across inputs, i.e., labor, material, and capital. ρ is a parameter for the returns to scale in the transformation of inputs into output. α_L, α_M, and α_K are distribution parameters associated with labor, material, and capital, respectively. We normalize their sum to 1.

Remark: A few features of the transformation function are worth noticing. First, the transformation function and the modelling of productivity are compatible with the single-product CES production functions traditionally used in the literature. To see this, consider a firm producing only a single product with quantity Q_{jt}. In this case, the transformation function degenerates into $Q_{jt} = e^{\tilde{\omega}_{jt}} \left[\alpha_L L_{jt}^{\gamma} + \alpha_M M_{jt}^{\gamma} + \alpha_K K_{jt}^{\gamma}\right]^{\frac{\rho}{\gamma}}$.

Second, for multi-product firms, the transformation function can be interpreted as the frontier of joint production of all products, Q_{jtn}, $n \in \Lambda_{jt}$. This interpretation has three implications: (i) different products are manufactured with the same set of inputs (i.e., labor, material, and capital); (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 modelling assumptions used by Dhyne et al. (2022), Orr (2022), and Valmari (2022). In our context, although the quantity substitution of one product and another is linear, the rate of substitution between them is determined by their

[7] A similar transformation function is adopted by Cairncross et al. (2023), who derive the transformation function (as a general output distance function) from individual product production function with shared inputs across products. In fact, the transformation function (3) in our setup is a restricted version of the output distance function proposed by Cairncross et al. (2023), which includes a CES aggregator on the output side with a parameter θ:

$$\sum_{n \in \Lambda_{jt}} [e^{-\tilde{\omega}_{jtn}} Q_{jtn}^{\theta}]^{\frac{1}{\theta}} = F(L_{jt}, M_{jt}, K_{jt}) \equiv \left[\alpha_L L_{jt}^{\gamma} + \alpha_M M_{jt}^{\gamma} + \alpha_K K_{jt}^{\gamma}\right]^{\frac{\rho}{\gamma}}.$$

Our restriction is $\theta = 1$. In this restricted setup, $\rho > 1$ implies that there is input sharing in production, as shown Cairncross et al. (2023).

relative productivity, which flexibly varies at the firm-product-period level.

Third, conditional on inputs, the output of a product does not only depend on its own productivity but also the productivity levels of other products. This can be observed from a two-product example of (3): $Q_{jt1} = e^{\tilde{\omega}_{jt1}} \left[\alpha_L L_{jt}^{\gamma} + \alpha_M M_{jt}^{\gamma} + \alpha_K K_{jt}^{\gamma}\right]^{\frac{\rho}{\gamma}} - e^{\tilde{\omega}_{jt1} - \tilde{\omega}_{jt2}} Q_{jt2}$. This is analogous to the transformation function proposed by Dhyne et al. (2022), with one key difference in terms of how productivity is defined. We explicitly model how the output of a product depends on its own productivity (i.e., $\tilde{\omega}_{jt1}$) as well as the productivity levels of other products (via the difference, $\tilde{\omega}_{jt2} - \tilde{\omega}_{jt1}$). As a result, our definition and identification of product 1's productivity is the transformation efficiency from inputs to output of product 1, conditional on not only the (observable) output of product 2 but also the (unobservable) productivity difference between the two products.[8] The (unobservable) productivity difference, if any, will be reflected by price differences across products within the same firm. We use such an insight to identify productivity differences and ultimately productivity levels in Section 3.

Finally and more importantly, our transformation function allows for shared inputs or the joint utilization of inputs across different products, which may contribute to economies of scope in the spirit of Panzar and Willig (1977, 1981). In fact, input allocation within a firm is not explicitly modelled in our framework. This methodology is in contrast to the existing methods that rely on imputing the intra-firm (exclusive) allocation of inputs and thus abstract away from imperfectly divisible inputs with properties of a public good within a firm.

The quantity-based productivity $\tilde{\omega}_{jtn}$ in (3) varies by firm, product, and period. While we do not impose restrictions on $\tilde{\omega}_{jtn}$ to estimate the parameters in (3) as will be shown in Section 3, in the rest of this subsection we discuss the potential components and evolution of $\tilde{\omega}_{jtn}$ to highlight the key differences compared with the assumptions in the existing literature.

Departing from the literature, we unpack productivity into two components:

$$\tilde{\omega}_{jtn} = \omega_{jtn} - h(\xi_{jtn}), \tag{4}$$

where ω_{jtn} is technical efficiency and $h(\xi_{jtn})$ is a function of product quality ξ_{jtn}. It is crucial to model $h(\xi_{jtn})$ as a part of quantity-based productivity because varieties of the same product category produced by different firms can be vertically differentiated by their quality levels. Intuitively, producing one unit of the high-quality product may require more production procedures

[8] Dhyne et al. (2022) model such a difference as a constant, product-specific parameter. We relax it as a productivity difference that can vary by firm, product, and time.

(e.g., longer refinements in the steel industry in Li et al., 2023), 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 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). For this reason, we refer to $h(\xi_{jtn})$ 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 modelling the cost of quality $h(\xi_{jtn})$ as a component of productivity allows for a trade-off between product quantity and quality, conditional on inputs. From an empirical perspective, this also implies that comparisons of quantity-based productivity across firms and over time require control for quality differences. Accordingly, we deal with $\tilde{\omega}_{jtn}$ as a whole rather than its components (ω_{jtn} and $h(\xi_{jtn})$) when estimating (3) in Section 3, and we explore the trade-off between (quantity-based) productivity and output quality in Section 6 after they are estimated.

While our empirical model does not rely on the evolution of productivity, we include it to facilitate the modelling of dynamic decisions in Section 2.4. Specifically, we model the evolution of ω_{jtn} as a flexible, endogenous Markov process:

$$\omega_{jt+1n} = g_n(\boldsymbol{\omega}_{jt}, I_{jt}) + v_{jt+1n}, \forall n = 1, 2, \ldots, N, \tag{5}$$

where $g_n(\cdot)$ is a function specific to product category n, v_{jt+1n} is an innovation term, and I_{jt} is firm-level investment (such as investment in research and development as emphasized by Doraszelski and Jaumandreu, 2013) that influences the future path of technical efficiency. Importantly, $\boldsymbol{\omega}_{jt} = (\omega_{jt1}, \omega_{jt2}, \ldots, \omega_{jtn})$ is a vector of firm-product level technical efficiency of *all* products of firm j in period t.[10] That is, the evolution process of the technical efficiency of one product can be influenced by the previous levels of technical efficiency of other products due to, for instance, intra-firm technology spillovers. The firm observes the realization of $\boldsymbol{\omega}_{jt}$ before making the production decisions specified below in Section 2.3.

[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).

[10] The firm observes the evolution processes of all products, even if the firm only selectively produces a subset of products.

Remark: Our modelling of the evolution processes is different from that of the literature in three aspects. First, we model the evolution of the underlying technical efficiency rather than quantity-based productivity as in the literature. When quality is an endogenous choice made by the firm and has an impact on quantity-based productivity, quantity-based productivity may no longer evolve in an auto-regressive way, even if the underlying technical efficiency is auto-regressive.

Second, we allow the evolution processes to be interdependent across products. From a computational perspective, adopting and estimating such flexible evolution processes would add a significant computational burden to the existing proxy-based approach in dealing with firms producing many products. Fortunately, our estimation methodology utilizes the first-order conditions of profit maximization to map observable firm input, output, and price choices to unobservable productivity, without relying on the evolution processes, as will become clear in Section 3. This feature is in contrast to the existing estimation methods (e.g., Orr, 2022; Valmari, 2022), which rely on the evolution assumption of productivity and thus exclude a flexible interdependency of productivity among different products.

Third, the literature usually only models the evolution processes of manufactured products due to data and computational limitations. However, this approach potentially suffers from an endogenous selection problem because firms only manufacture products when they are profitable. This problem could be severe if the product turnover (i.e., adding and dropping products) is frequent. An appropriate approach is to model the evolution processes of *all* products. But this imposes a challenge in dealing with the latent variables that determine product selection, which is usually dynamic. Our estimation methodology saves us from the data and computational challenges, because it does not rely on the productivity evolution processes.

2.3 Static Decisions: Inputs and Outputs

At the beginning of period t, the firm observes the vector of state variables, which includes the product scope, capital stock, intermediate input price, wage rate, technical efficiency, and product quality of all the products. We summarize the state variables in $s_{jt} = (\Lambda_{jt}, \boldsymbol{\omega}_{jt}, \boldsymbol{\xi}_{jt}, K_{jt}, P_{Mjt}, P_{Ljt}, \boldsymbol{\chi}_{jt})$, where $\boldsymbol{\omega}_{jt}$, $\boldsymbol{\xi}_{jt}$ and $\boldsymbol{\chi}_{jt}$ are the vectors of technical efficiency, product quality and demand shocks of *all* the products of firm j in period t, respectively. Note that the observation of technical efficiency and product quality implies that the firm also knows productivity, $\tilde{\boldsymbol{\omega}}_{jt}$, because the firm knows the trade-off (4). P_{Mjt} and P_{Ljt} are the firm-level material price and the wage rate, respectively. Importantly, both of them can be different across firms and vary over time.

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 ($\boldsymbol{Q}_{jt} = \{Q_{jtn}\}, n \in \Lambda_{jt}$). Specifically, the period (static) profit is:

$$\begin{aligned} \pi(s_{jt}) \quad &= \quad \max_{\boldsymbol{Q}_{jt}, M_{jt}, L_{jt}} \sum_{n \in \Lambda_{jt}} P_{jtn} Q_{jtn} - P_{Mjt} M_{jt} - P_{Ljt} L_{jt} \\ \text{subject to:} \quad & \quad (2) \text{ and } (3). \end{aligned} \tag{6}$$

Remark: In commonly available data, while P_{Ljt} is usually observable to researchers as the wage rate, P_{Mjt} is rarely recorded at the firm level. As documented by Atalay (2014) using US Census Bureau data, P_{Mjt} can be significantly heterogeneous across firms due to geography, bargaining power and access to the input market, suppliers' marginal costs, etc. It is well understood that such input price heterogeneity should be controlled for in the production function estimation to avoid bias (i.e., input price bias as emphasized in Ornaghi, 2006; De Loecker et al., 2016; Grieco et al., 2016). Recent developments in the estimation of multi-product production functions usually assume the availability of P_{Mjt} (or a firm-level index of it, e.g., Orr, 2022; Valmari, 2022). In contrast, our method is tailored to accommodate common situations where input prices vary at the firm level but are unobservable to researchers. In particular, we maintain the assumption of the literature that P_{Mjt} varies at the firm level (as opposed to the firm-product level) because we model the production as a transformation function (rather than an individual production plant for each product).[11] We control for P_{Mjt} following the insights of Grieco et al. (2016, 2022), Harrigan et al. (2021), and Li and Zhang (2022), as will be shown in Section 3. Consequently, our empirical method for estimating multi-product production functions offers broader applicability in commonly available datasets compared to existing methods.

2.4 Dynamic Decisions

This subsection briefly 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 $s_{jt} = (\Lambda_{jt}, \boldsymbol{\omega}_{jt}, \boldsymbol{\xi}_{jt}, K_{jt}, P_{Mjt}, P_{Ljt}, \boldsymbol{\chi}_{jt})$ and after observing the adjustment costs of product scope and quality levels. Although the evolution of K_{jt}, P_{Mjt}, P_{Ljt} and $\boldsymbol{\chi}_{jt}$ can be endogenous,

[11] This assumption holds if the input can be costlessly transferred across product lines within the firm, as assumed by Orr (2022) and Valmari (2022).

we remain agnostic on modelling 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). As a result, these decisions are dynamic.

Remark: Although we do not estimate the complex dynamic model in this paper (due to the considerably high dimension of the state variables),[12] the model serves the crucial purpose of clarifying the (dynamic) choices made by the firm and their implications when we estimate the static model. In particular, the dynamic model implies that even if the underlying technical efficiency follows a simple AR(1) process, the resulting productivity (i.e., TFPQ) is not an AR(1) process as assumed in the literature. To see this, note that quality ξ_{jt+1n} is endogenously determined by the firm based on the state variable vector s_{jt}, including the technical efficiency of all products (i.e., $\boldsymbol{\omega}_{jt}$). Considering the impact of quality on productivity shown in (4), productivity of any product n in period $t+1$ depends on the entire state vector s_{jt} in a highly nonlinear way. Ignoring such interdependent relationships may potentially result in biased estimation. Fortunately, our empirical method does not use any assumptions regarding how technical efficiency and productivity evolve, as will be discussed in Section 3.

3 Estimation

In the estimation method, we focus on the static component of the model, which produces a set of implications 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

[12] For example, even in the footwear industry with only four products, the dynamic state includes at least 10 continuous variables – 4 variables for technical efficiency, 4 variables for product quality, and 2 for the material and labor prices.

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, our model always admits such a mapping regardless of the number of products.

Compared with the existing methods in the literature, our method has several important innovations and advantages. The first advantage is scalability. Our method is readily applicable to industries with many products because the method does not require proxies for product-level productivity, and rather relies on static optimization conditions that naturally increase with the number of products. Second, our method does not rely on productivity evolution processes. This enables researchers to explore the productivity evolution *after the estimation*, contrary to the existing methods which rely on productivity evolution *for the estimation*. More broadly, this advantage is useful in applications when product turnover is frequent and endogenously depends on latent variables (e.g., in the context of exported products). Third, our method models the production technology flexibly as a transformation function. 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. Fourth, our method is designed to deal with the challenge of unobserved material prices. This is particularly useful when material prices are heterogeneous across firms and over time but are unobservable to researchers. Fifth, our estimation of demand functions leverages the within-firm revenue relationship implied by profit maximization to estimate demand elasticities with commonly available firm-level IVs. This alleviates the need for firm-product level IVs in the demand estimation that are rarely available.

This section is organized as follows. In Section 3.1, we first describe how the static profit maximization conditions lead to one-to-one mapping between the observed data and variables that are unobservable to researchers. In Section 3.2, we derive the estimating equations using the mapping established in Section 3.1 and describe our estimation strategy in detail.

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

We start the description of the estimation strategy by clarifying the observable and unobserved variables in the estimation procedure. We observe capital stock K_{jt}, labor input L_{jt}, labor expenditure E_{Ljt}, material expenditure E_{Mjt}, and quantity Q_{jtn} and price P_{jtn} of each product $n \in \Lambda_{jt}$. We do not observe P_{Mjt} (or M_{jt}), and $\tilde{\xi}_{jtn}$ and $\tilde{\omega}_{jtn}$ for $n \in \Lambda_{jt}$. Our goal is to estimate these unobserved variables together with the production and demand function parameters. Next, we describe the mapping between observed data and unobservables on the basis of the firm's profit maximization.

Note that the firm observes the state s_{jt} (in particular, $\tilde{\omega}_{jtn}$ and $\tilde{\xi}_{jtn}$ for all $n \in \Lambda_{jt}$ and P_{Mjt}) as described in Section 2.3 and optimally chooses quantities of inputs and outputs subject to the demand and production functions. The Lagrange function implied by the static profit maximization problem (6) is:

$$\begin{aligned}\mathcal{L}_{jt} \quad = \quad & \sum_{n \in \Lambda jt} (Q_{jtn})^{1-\frac{1}{\eta_n}} e^{\frac{\tilde{\xi}_{jtn}}{\eta_n}} - P_{Ljt}L_{jt} - P_{Mjt}M_{jt} \\ & -\lambda_{jt}\Bigg\{ \Bigg[\sum_{n \in \Lambda_{jt}} e^{-\tilde{\omega}_{jtn}} Q_{jtn} \Bigg] - F(L_{jt}, M_{jt}, K_{jt}) \Bigg\}. \end{aligned} \tag{7}$$

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

$$\frac{\partial \mathcal{L}_{jt}}{\partial L_{jt}} = -P_{Ljt} + \lambda_{jt} \frac{\partial F(L_{jt}, M_{jt}, K_{jt})}{\partial L_{jt}} = 0, \tag{8}$$

$$\frac{\partial \mathcal{L}_{jt}}{\partial M_{jt}} = -P_{Mjt} + \lambda_{jt} \frac{\partial F(L_{jt}, M_{jt}, K_{jt})}{\partial M_{jt}} = 0. \tag{9}$$

The first-order condition with respect to each product quantity Q_{jtn}, $n \in \Lambda_{jt}$, is:

$$\frac{\partial \mathcal{L}}{\partial Q_{jtn}} = \frac{\eta_n - 1}{\eta_n} P_{jtn} - \lambda_{jt} e^{-\tilde{\omega}_{jtn}} = 0, \tag{10}$$

where we have used $P_{jtn} = (Q_{jtn})^{-\frac{1}{\eta_n}} e^{\frac{\tilde{\xi}_{jtn}}{\eta_n}}$ according to the demand function (2). The implication of (10), $P_{jtn} = \frac{\eta_n}{\eta_n - 1} \lambda_{jt} e^{-\tilde{\omega}_{jtn}}$, is intuitive: the price is the product of the markup $(\frac{\eta_n}{\eta_n-1})$ and the marginal cost $(\lambda_{jt} e^{-\tilde{\omega}_{jtn}})$.[13] Within a firm, the marginal cost of a given product differs only due to productivity $\tilde{\omega}_{jtn}$, although the marginal cost also varies across firms due to λ_{jt}. This is a direct result of the costless input transferability assumption of the production transformation function and profit maximization. Therefore, conditional on a firm, the variation in product prices identifies the productivity difference across products within the firm (after accounting for the markup).

[13] It should be noted that because the demand elasticities vary by product and firms produce different sets of products, markups also vary by firm-year pairs. We report the estimated variation of markups at the firm-year in Section 5.

From the perspective of researchers, we do not observe $\tilde{\xi}_{jtn}$, $\tilde{\omega}_{jtn}$ and P_{Mjt}. Nonetheless, we observe the optimal choices which are made based on them by the firm. Thus, utilizing the optimization conditions allows us to recover the unobserved state variables as functions of the observable variables. Specifically, our strategy is to recover $\tilde{\xi}_{jtn}$, $\tilde{\omega}_{jtn}$ and P_{Mjt} as functions of parameters and observable variables including capital stock K_{jt}, labor input L_{jt}, labor expenditure E_{Ljt}, material expenditure E_{Mjt}, quantity Q_{jtn} and price P_{jtn} of each product $n \in \Lambda_{jt}$.

First, we write $\tilde{\xi}_{jtn}$ as a function of observed price and quantity according to the demand function (2):

$$\tilde{\xi}_{jtn} = \ln Q_{jtn} + \eta_n \ln P_{jtn}. \tag{11}$$

Once η_n is estimated, we can recover $\tilde{\xi}_{jtn}$ as above.

Second, we write P_{Mjt} as a function of observable variables. Taking the ratio of equations (8) and (9) and utilizing the expenditure identities (i.e., $E_{Ljt} = L_{jt}P_{Ljt}$ and $E_{Mjt} = M_{jt}P_{Mjt}$), we have:

$$M_{jt} = \left[\frac{\alpha_L E_{Mjt}}{\alpha_M E_{Ljt}}\right]^{\frac{1}{\gamma}} L_{jt}. \tag{12}$$

This implies that material quantity can be recovered from observable variables up to unknown parameters $(\alpha_L, \alpha_M, \gamma)$. Thus, P_{Mjt} is naturally derived by substituting (12) in the expenditure identity (i.e., $E_{Mjt} = M_{jt}P_{Mjt}$):

$$P_{Mjt} = \left[\frac{\alpha_M}{\alpha_L}\right]^{\frac{1}{\gamma}} \left[\frac{E_{Mjt}}{E_{Ljt}}\right]^{1-\frac{1}{\gamma}} P_{Ljt}. \tag{13}$$

In the same spirit of Grieco et al. (2016), the identification of P_{Mjt} comes from the variation of labor and material expenditure ratio (conditional on wage rate), which is implied by the optimality condition under non-Hicks neutrality of the material price in the transformation function.

The third step is to recover $\tilde{\omega}_{jtn}$ for $n \in \Lambda_{jt}$. Specifically, by substituting (12) into (8), we can solve for λ_{jt} as:

$$\lambda_{jt} = \frac{E_{Ljt}}{\rho \alpha_L L_{jt}^{\gamma}} \left[\alpha_L L_{jt}^{\gamma} \left(1 + \frac{E_{M_{jt}}}{E_{L_{jt}}}\right) + \alpha_K K_{jt}^{\gamma}\right]^{1-\frac{\rho}{\gamma}}. \tag{14}$$

Then, we substitute (14) into (10) to get:

$$e^{\tilde{\omega}_{jtn}} = \frac{\eta_n}{(\eta_n - 1)P_{jtn}} \underbrace{\frac{E_{Ljt}}{\rho \alpha_L L_{jt}^{\gamma}} \left[\alpha_L L_{jt}^{\gamma} \left(1 + \frac{E_{M_{jt}}}{E_{L_{jt}}}\right) + \alpha_K K_{jt}^{\gamma}\right]^{1-\frac{\rho}{\gamma}}}_{\lambda_{jt}}. \tag{15}$$

Noticeably, there are two major components in (15) that identify firm-product-period specific productivity, $\tilde{\omega}_{jtn}$. The first is a firm-level component, λ_{jt} as in (14). This component is the analog of single-product-firm productivity modelled by Grieco et al. (2016) (see their equation (7)). This (unobserved) productivity component is identified from the (unobserved) material price because productivity is Hicks-neutral while the material price is not in our framework.[14] That is, a change in the material price causes a change in the (observable) labor-material expenditure ratio, but a productivity change does not. The second major component, which varies by firm and by product, consists of P_{jtn} and η_n. Intuitively, the variation in product prices helps identify the differences in productivity across products within the same firm, conditional on the elasticity of demand. That is, firms with higher (quantity-based) productivity pass the cost-saving to the product prices (as in Foster et al., 2008). Consequently, the product with a lower price has higher productivity compared with another product manufactured by the same firm, after controlling for the markup (implied by the elasticity of demand). In sum, our identification of $\tilde{\omega}_{jtn}$ uses the variations both at the firm level and at the firm-product level.

Remark: The proxy-based methodology, originated from Olley and Pakes (1996) along with a long list of methodological papers, uses observable variables (such as capital investment and material input) to control for productivity when estimating production functions. Extending the proxy-based approach to the multiple-product context requires valid proxies, which have to admit a one-to-one mapping between the proxies and firm-product level productivity. This is a challenging assumption in the context of a large number of products due to the high dimension of the problem. More importantly, the number of proxies has to increase with the number of products (as recogenized by Dhyne et al., 2022), making the extension even more challenging without additional assumptions. The recent development in methods (i.e., Chen and Liao, 2022; Orr, 2022; Valmari, 2022) circumvents this challenge by using production functions of individual products as proxy functions directly, after imputing intra-firm input allocation from firm optimization conditions. This approach assumes that there is no transitory error in production (which is explicitly modelled and dealt with by Olley and Pakes, 1996) and that the persistent error (as in the traditional notion of productivity) evolves independently according to a Markov process.

In contrast, our methodology uses first-order conditions to construct an explicit one-to-one mapping for productivity (up to the parameters to be estimated). This does not only guarantee the existence and uniqueness of the mapping from observable data to unobservable heterogeneity,

[14] If both productivity and the material price are Hicks-neutral in the production function, as in Cobb-Douglas production functions, then this identification strategy fails. However, in this case, the labor-material expenditure ratio would be a constant under the optimality condition, which is not supported by the data.

but also lends us a significant advantage in dealing with scenarios where firms produce a large number of products, because the number of first-order conditions naturally increases with the number of products. In addition, this methodology of recovering unobservable heterogeneity (instead of imputing input allocation) saves us from estimating productivity evolution processes as a part of the production estimation, which can dramatically complicate the existing methods in the literature, especially when there are endogenous, frequent entry and exit of products or the evolution processes of productivity are interdependent. More broadly, this feature allows our method to be widely applied to analyzing the impact of policy shocks on productivity, which would have to be otherwise considered as a part of the evolution processes (as emphasized by Chen et al., 2021) and further complicate the estimation process using the existing methods.

3.2 Estimating Equations and Strategy

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

In contrast to the existing methods (i.e., Orr, 2022; Valmari, 2022) that abstract away from unexpected shocks of production, we relax this assumption by following the spirit of Olley and Pakes (1996) to allow for a transitory shock u_{jt} (in addition to productivity) to the transformation function (3) in the estimation:

$$\sum_{n\in\Lambda_{jt}} e^{-\tilde{\omega}_{jtn}} Q_{jtn} = \left[\alpha_L L_{jt}^{\gamma} + \alpha_M M_{jt}^{\gamma} + \alpha_K K_{jt}^{\gamma}\right]^{\frac{\rho}{\gamma}} e^{u_{jt}}. \tag{16}$$

Specifically, u_{jt} is a non-structural firm-year level unexpected shock (or measurement error), which has a mean of zero: $E(u_{jt}) = 0$. The distinction between u_{jt} and productivity ($\tilde{\boldsymbol{\omega}}_{\boldsymbol{jt}}$) is that the firm observes productivity when making decisions and thus productivity is correlated with input choices, while u_{jt} is not observed by the firm and thus is uncorrelated with input choices.

Remark: Essentially, u_{jt} is an unforecastable shock (beyond productivity) that influences the quantity of output at the firm level. Orr (2022) and Valmari (2022) assume away such a shock (i.e., productivity is the only unobserved variable in their production models after input allocation is recovered) in order to directly use the production function itself as a control function for productivity in the productivity evolution process. This assumption is the key to their scalability to handle a high dimension of productivity at the firm-product level. In contrast, our approach does not treat the production (transformation) function as a control function and thus we can allow for such an unforecastable shock.[15]

Because u_{jt} is unobserved by the firm, it does not affect the production decisions (i.e., the first-order conditions) derived above. Substitute (12) and (15) into (16), and after some algebra, we obtain the following estimating equation:

$$\ln\left[\sum_{n\in\Lambda_{jt}} \frac{(\eta_n-1)\rho}{\eta_n} R_{jtn}\right] = \ln\left[E_{M_{jt}} + E_{L_{jt}}\left(1+\frac{\alpha_K}{\alpha_L}\left(\frac{K_{jt}}{L_{jt}}\right)^{\gamma}\right)\right] + u_{jt}, \tag{17}$$

where R_{jtn} is the revenue of product n of firm j at period t, and u_{jt} is the firm-year level unexpected shock in (16). In fact, this equation is the multi-product equivalent of the estimating equation proposed by Grieco et al. (2016) (see their equation 8), who assume that each firm produces a single product.[16] In the context of multi-product firms, the individual product revenues are adjusted by their corresponding markups (reciprocal).[17] Although u_{jt} does not affect production decisions, it does appear as a part of observed product revenues. A higher shock implies a higher realized revenue R_{jtn}. Thus, u_{jt} and R_{jtn} are correlated. This also implies that we need to estimate the model via the generalized method of moments (GMM).[18]

Nonetheless, estimating all the parameters using (17) alone faces two challenges. First, ρ is not separately identified from demand elasticities in (17). In fact, only a combination of η_n and ρ (i.e., $\frac{(\eta_n-1)\rho}{\eta_n}$) is identified by (17). Second, (17) 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 all product revenues are correlated with u_{jt}.

To address the two challenges at the same time, 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. Notably, η_n influences the sales of *individual*

[15] After estimating our model, we find that the dispersion of such shock u_{jt}, as reflected as (19), is significant. For example, the inter-quartile range of u_{jt} in the pharmaceutical industry is 0.44, implying that it should not be ignored or treated as a part of productivity.

[16] More broadly, (17), 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.

[17] 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.

[18] If one is willing to assume that u_{jt} is uncorrelated with the revenues or the demand elasticities are estimated (e.g., by estimating the demand functions directly) before the production estimation, it is possible to estimate (17) using a Nonlinear Least Squares (NLLS) estimator. The key results from the NLLS estimation are quantitatively and qualitatively similar to our main results.

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 (17). 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 define a reference product. In principle, the reference product can be any product. From an empirical point of view, we use the product that is manufactured by most firms in the industry, in order to maximize the number of observations one could use in the estimation. Without loss of generality, we denote the reference product as product 1. For any firm j, taking the ratio of (10) of the reference product and that of another product n and using $R_{jtn} = P_{jtn}Q_{jtn}$, we obtain:

$$\ln(R_{jt1}) = c_n + \frac{\eta_1 - 1}{\eta_n - 1}\ln(R_{jtn}) + \mu_{jtn}, \qquad n = 2, \dots, N, \tag{18}$$

where

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

and

$$\mu_{jtn} = (\eta_1 - 1)\left[\underbrace{(\tilde{\omega}_{jt1} + \frac{1}{\eta_1 - 1}\tilde{\xi}_{jt1}) - (\tilde{\omega}_{jtn} + \frac{1}{\eta_n - 1}\tilde{\xi}_{jtn})}_{\text{difference in quality-adjusted productivity}} + \underbrace{\frac{\eta_1 - \eta_n}{(\eta_1 - 1)(\eta_n - 1)}u_{jt}}_{\text{measurement error component}}\right].$$

The latter, μ_{jtn}, contains the *difference* of the capability (or quality-adjusted productivity, $\tilde{\omega} + \frac{1}{\eta - 1}\tilde{\xi}$, as will be formally defined in Section 5) of producing a product relative to that of the reference product and composition of the unexpected shock. This equation predicts that the (logarithmic) revenues of two products are linearly related 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).[19]

[19] One crucial 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 thus they cancel each other out.

Intuitively, because the demand for each product is segmented in our setting, as discussed in Sections 2.1 and 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 productivity and quality, measured as μ_{jtn}) rather than on complementarity or substitution between them. As a result, the variation of one revenue *relative* to another in (18) provides the identification of the ratio, $\frac{\eta_1-1}{\eta_n-1}$ for $n = 2, 3, \ldots, N$. In contrast, the variation of revenue *levels* in (17) 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. The model is over-identified when there are more than two products produced by the firms 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.

To estimate (18), we treat μ_{jtn} as an error term. We allow the mean of μ_{jtn} to vary by product and year and use a set of flexible product-year dummies as controls (which also absorb c_n). Still, μ_{jtn} is likely correlated with R_{jtn} – 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, 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).[20] Grieco et al. (2022) uses a similar set of firm-level IVs to estimate an equation analogous to (18) in a two-product scenario. The same insight carries over in our context. These firm-level variables influence the *level* of revenue (i.e., R_{jtn}), but they are uncorrelated with the *difference* of capability (i.e., μ_{jtn}) between two products. For example, conditional on everything else, a 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. Thus, we use these firm-level variables as IVs for all product pairs in (18).[21]

[20] To see this, note that (13) 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.

[21] 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., $\frac{\eta_1-1}{\eta_n-1}, n = 2, \ldots, N$).

These firm-level IVs are valid under 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, because otherwise that input (as an IV) will be endogenously affected by the capability of producing that product relative to other products. In our structural framework, given the transformation function and costless transferability of inputs across products, this condition is always satisfied.[22] We use Monte Carlo exercises to demonstrate the performance of our approach and IVs. As shown in the first panel of Table A7 in Appendix B, the approach estimates the true value of $\frac{\eta_1 - 1}{\eta_n - 1}$ very well.

Remark: It is worth noting three features of our strategy to estimate the relationship of the elasticities of demand between products. First, although the demand elasticities can be, in principle, identified by the variation of prices and quantities in (2), our strategy does not use such variation. In fact, our strategy has an advantage over the traditional ways of using such variation in estimating η_n via the demand function (2) directly. In general, one needs IVs (such as firm-product-time varying cost shifters) that are uncorrelated with product quality to estimate (2) directly. But commonly used cost shifters can hardly serve this purpose if firms producing high-quality products use high-quality inputs that come with higher costs. Nonetheless, if one has appropriate IVs to estimate the demand function (2) directly, then it is not necessary to estimate (18), and, consequently, our main equation (17) can be simply estimated using a Nonlinear Least Square estimator directly (instead of using GMM). For example, Orr (2022) designs sophisticated IVs that leverage the variation of product sets and material input price growth experienced by firms in other output markets that use similar inputs. However, his strategy does not apply to broader data settings with little variation of product sets or unobserved input prices and inputs (like in our context). Fortunately, simple cost shifters such as firm-level capital stock are appropriate IVs to estimate (18) for all product pairs, which explores the relative difference (as opposed to the level) of revenue across products.

Second, our strategy is also different from the practice of estimating the first-order differences of demand function (2), which implicitly assume that the unobserved quality is constant from one period to another (i.e., Valmari, 2022). In contrast to this assumption, our strategy explores the advantage offered by the intra-firm decisions in multi-product firms to estimate the time-varying

[22] To examine this condition empirically, we check whether the IVs are correlated with either the within-firm product shares or the ratio of log sales of a product over that of the baseline product as alternative measures of relative production capability. Specifically, we regress each IV on either the interactions between product fixed effects and within-firm revenue shares (including firm and year fixed effects) or the interactions between product fixed effects and the ratio of log sales of a given product over that of the baseline product (including firm and year fixed effects). We find that at least 85% of coefficients (i.e., products) are not significant at the 1% level in these tests for our IVs.

unobserved quality at the firm-product level.

Finally, if one is willing to assume constant return to scale (i.e., $\rho = 1$), then the demand elasticities can be identified using (17) alone, without relying on the strategy involving (18). In fact, with the constant return to scale assumption, (17) degenerates to the estimating equations used by Das et al. (2007), Aw et al. (2011), and Li (2018). These papers utilize the relationship between the total variable cost (as our counterpart of the right-hand side of (17)) and export revenues (as our counterpart of the left-hand side of (17)) of the same firm to estimate demand elasticities in multiple export markets.

We denote the estimated relationship between elasticities as $\hat{b}_n = \frac{\eta_1 - 1}{\eta_n - 1}$, $n = 2, \ldots, N$, and, naturally, $\hat{b}_1 = 1$ by definition. Thus, $\eta_n = \frac{1}{\hat{b}_n}(\eta_1 - 1) + 1$. Substitute it as η_n in (17) and solve for u_{jt} to construct moment conditions for the GMM estimation:

$$u_{jt} = \ln \rho + \ln \left[\sum_{n \in \Lambda_{jt}} \frac{\eta_1 - 1}{\eta_1 - 1 + \hat{b}_n} R_{jtn} \right] - \ln \left[E_{M_{jt}} + E_{L_{jt}} \left(1 + \frac{\alpha_K}{\alpha_L} \left(\frac{K_{jt}}{L_{jt}} \right)^{\gamma} \right) \right]. \tag{19}$$

Importantly, 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. In particular, firm-level input choices can serve as valid IVs because they are not correlated with the unexpected shock u_{jt}. In the implementation, we use $Z_{jt} = (1, E_{Mjt}, E_{Ljt}, L_{jt}, K_{jt}/L_{jt})$ as IVs. Our results are robust to a set of alternative firm-level IVs.

Of course, (19) can only identify $\frac{\alpha_K}{\alpha_L}$ (rather than α_L, α_M, and α_K separately). As shown by Grieco et al. (2016), 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: $\alpha_L + \alpha_M + \alpha_K = 1$. 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 (12), which is implied by the first-order conditions of labor and material quantities, (8) and (9), of all firms.[23]

[23] As shown by Grieco et al. (2016), this constraint holds conditional on a normalization of the CES production function. Thus, we follow the same procedure to normalize the inputs using their corresponding industry-level geometric means as in the literature (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 (12) across firms.

As a result, β can be estimated as:

$$\begin{aligned} \hat{\beta} &= \operatorname{argmin}_{\beta} \left[\frac{1}{\mathbb{N}} \sum_{j,t} u_{jt} Z_{jt} \right]' W \left[\frac{1}{\mathbb{N}} \sum_{j,t} u_{jt} Z_{jt} \right], \qquad (20) \\ \text{subject to:} & \qquad \alpha_L + \alpha_M + \alpha_K = 1 \quad \text{and} \quad \frac{\alpha_M}{\alpha_L} = \frac{\overline{E}_M}{\overline{E}_L}, \end{aligned}$$

where W is a weight matrix and $\mathbb{N}$ is the number of firm-time observations.

As a summary of the full estimating approach, the first step is to estimate $\hat{b}_n = \frac{\eta_1 - 1}{\eta_n - 1}$, $n = 2, \ldots, N$ via Two-Stage Least Squares (2SLS) using the relationship imposed by the within-firm relative sales in (18). The second step is to estimate $(\hat{\rho}, \hat{\eta}_1, \hat{\alpha}_L, \hat{\alpha}_M, \hat{\alpha}_K, \hat{\gamma})$ using (19) via GMM (together with the constraints regarding the parameters). With these estimates, the demand elasticities can be recovered as $\hat{\eta}_n = \frac{1}{\hat{b}_n}(\hat{\eta}_1 - 1) + 1$. After that, we compute $\tilde{\xi}_{jtn}$, $\tilde{\omega}_{jtn}$, and P_{Mjt} via (11), (15), and (13), respectively. We demonstrate that our method is able to recover the true parameter values in the Monte Carlo exercises shown in Table A7 of Appendix B.

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.[24] These datasets are particularly useful for our analysis because they provide quantity and sales information at the firm-product level.

Next, we describe in more detail these two surveys and the variables we extract from them.[25] The EIA contains information on 6867 firms in 1994, but this number decreases over time due to attrition. It covers roughly 85 percent of all manufacturing output value based on information from the industrial census, but it excludes assembly plants, i.e., "maquiladoras". The EIA includes variables related to output indicators, inputs, and investment. These data make it possible to calculate the value of intermediate inputs and physical capital stock based on information on investment and the perpetual inventory method. The EIM runs in parallel with the EIA and covers

[24]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.

[25]More information on the EIA and EIM can be found in Caselli et al. (2017) and Caselli (2018).

the same firms. The EIM contains information on the number of workers and their wage bills so that the average wage at the firm level can be calculated. The EIM also contains output-related variables, in particular values and quantities of sales at the product level, so that an implicit average unit price can be calculated.[26]

Firms are classified 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 information product by product based on their industries.

In this paper, 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). These three industries were chosen because each industry is made up of more than 500 firm-year pairs, a number of observations large enough for our estimation strategy. More importantly, multi-product firms are particularly prevalent in these industries – 65% of firms in these three industries are multi-product producers and such firms account for 86% of total revenues and produce on average 6.7 products per year.[27] They also represent a diverse set of manufacturing industries with clear concepts/characteristics 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.

For the purpose of the production function estimation in Section 5, all products with fewer than 100 observations are aggregated together in a residual product category.[28] The prices and quantities of the aggregated residual product category are estimated following Diewert et al. (2009) and Caselli (2018). 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

[26] 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.

[27] Tables A1, A2 and A3 in the Appendix show how detailed the product-level information is by reporting the list of products with at least 100 observations for each of the three chosen industries.

[28] The residual product category is defined as "Others" (product code 99) in Tables A1, A2 and A3 in the Appendix.

relatively small share of products: the main (i.e., not aggregated) products account for between 74% and 92% of observations and 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 the estimation in Section 5.

There are a few patterns worth noticing. 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). Specifically, we construct an analog index of 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. In Figure 1, we aggregate the firm-level index with weights equal to the firms' total revenues, by firm-year pairs' product scope. 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.[29] 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.[30] Thus, multi-product firms need to be treated and modelled as such and they cannot be simplified as single-product producers.

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 1. The percentage of multi-product firms ranges from 21% in the footwear industry to 55% in printing and 85% in pharmaceuticals and they account for an even larger share of revenues (from 39% in the footwear industry to 94% in pharmaceuticals). The average product scope is larger in printing and pharmaceuticals (respectively, 5.9 and 7.9 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.

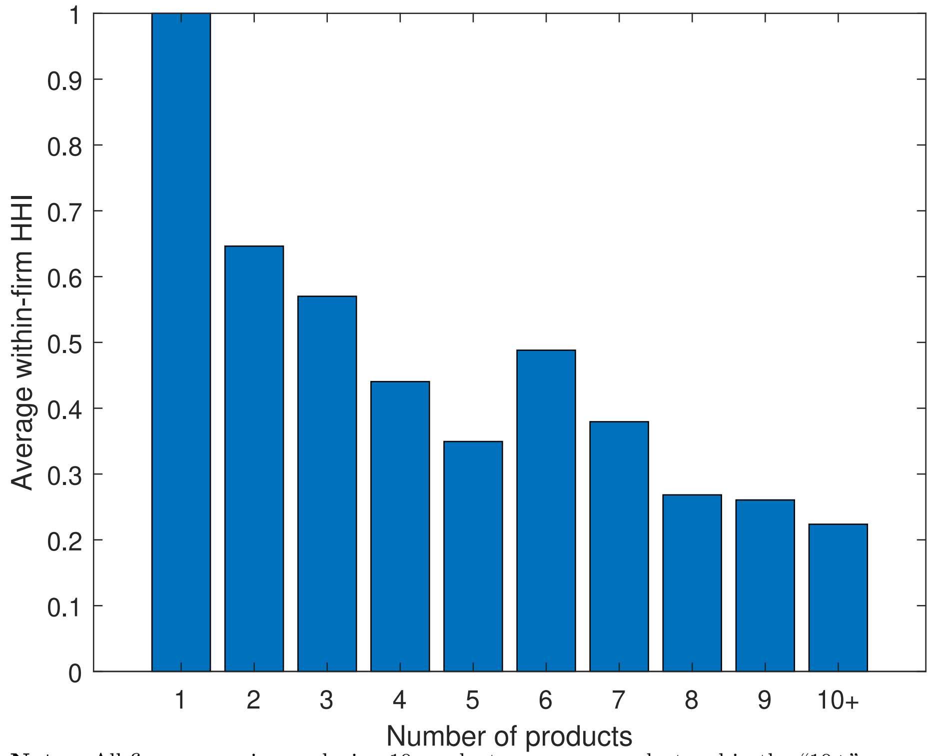
Second, the status of being a multi-product firm is quite persistent, and so is the product scope. In particular, using a simple autoregressive process of the number of products produced by each firm, we measure the persistence coefficients are 0.87, 0.95, and 0.98 in the three industries, respectively.[31] Thus, multi-product firms unequivocally dominate manufacturing production in

[29] 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 6.

[30] To confirm that firms rely heavily on all products for their total sales, Appendix Table A4 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 0.557 and the share of products with rank 5 and beyond is 0.146, on average.

[31] The entry of new products and the exit of old products only account for 3.8 and 4.4 percent of the observations, respectively.

Figure 1: 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.

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, intra-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 (15), 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 model's demand assumption. On average, about 19 to 43 firms are competing 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 at the product-year level ranges between 0.11 in the pharmaceutical industry and 0.26 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) are reasonably assumed as segmented. Of course, for each product, firms' outputs are vertically differentiated as evidenced by the large dispersion in prices. For example, the interquartile range of prices in logarithm is 1.4 (i.e., a 400% difference) within a product category, on average, across the three industries. Overall, these patterns support abstracting from demand cannibalization across products manufactured by the same firm and assuming that firms face monopolistic competition within each product category.

Table 1: Descriptive statistics

Variable	Footwear	Printing	Pharmaceutical
Revenue per product (R)	64.846	29.142	100.167
	(101.261)	(73.091)	(207.205)
Number of workers (L)	236.180	157.704	450.222
	(361.356)	(153.598)	(482.926)
Labor expenditure (E_L)	13.785	17.435	88.792
	(28.726)	(22.124)	(110.791)
Material expenditure (E_M)	50.446	65.568	263.033
	(77.265)	(90.175)	(382.666)
Capital stock (K)	3.603	21.491	22.534
	(8.444)	(47.314)	(31.437)
Product scope, all firms	1.289	3.708	6.858
	(0.627)	(3.752)	(3.740)
Product scope, MPFs only	2.388	5.891	7.925
	(0.602)	(3.841)	(3.024)
Share of MPFs	0.208	0.554	0.846
Revenue share of MPFs	0.389	0.599	0.940
Number of products	4	14	16
Number of firms	72	83	82
Average number of firms per product-year	21	19	43
Number of firm-year pairs	707	831	928

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); Product scope is the number of products manufactured by firm.

5 Estimation Results

In this section, we apply the empirical model to the data and estimate the production and demand function parameters by industry, which then allows us to compute firm-product level productivity and quality. Notably, our approach employs a novel method, and despite this novelty, the resulting structural parameter estimates align closely with existing literature. Moreover, the productivity and quality measures derived from these estimates exhibit economically meaningful properties. Because our empirical analysis relies on estimated variables, and to account for this, we employ bootstrapping with 250 samples to compute all standard errors presented in the subsequent tables, ensuring robustness and accuracy in our findings.

Table 2 presents the production function 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 (ranging from 1.4 to 2.6), France (ranging from 1.1 to 2.0), and China (ranging from 1.2 to 2.7), respectively. It is also close to the average estimate (around 1.4) of the elasticity of substitution among Chinese industries by Berkowitz et al. (2017) using a different method. Finally, the returns to scale parameter ρ of the three industries is larger than one, but it is not significantly different from one, implying that the production is close to constant returns to scale in these industries, except in the case of the footwear industry.

Table 2: Production Function Estimates

Parameter	Footwear	Printing	Pharmaceutical
α_L	0.202	0.229	0.227
	(0.014)	(0.015)	(0.022)
α_M	0.774	0.673	0.595
	(0.043)	(0.030)	(0.066)
α_K	0.023	0.099	0.178
	(0.054)	(0.038)	(0.084)
σ	1.518	1.244	1.168
	(0.587)	(0.151)	(0.250)
ρ	1.227	1.078	1.002
	(0.102)	(0.115)	(0.118)

Note: Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions).

Table 3 presents the estimates of the demand elasticities of substitution of different products in the three industries. These estimates generally fall within a similar range as those found in the existing literature (e.g., see Roberts et al., 2018; Grieco et al., 2016; Dubois and Lasio, 2018, for the Chinese footwear industry, the Colombian printing industry, and the French pharmaceutical industry, respectively). Notably, our estimation exploits the relationship between product revenues within the same firm, as described in (18). This approach is in contrast to the literature, which often relies on direct estimation of the demand function while assuming time-invariant product quality and/or using firm-level instrumental variables, such as capital stock. By leveraging the multi-product context, this approach capitalizes on the advantage of utilizing firm-level IVs that may be potentially correlated with the *level* of quality but are less likely to be correlated with the *difference* in production capabilities of any two products within the same firm.[32]

Table 3: Demand Function Estimates

Parameter	Footwear	Printing	Pharmaceutical
η_1	2.823	4.523	3.688
	(0.480)	(1.738)	(1.187)
η_2	2.455	8.661	3.037
	(0.570)	(2.664)	(1.470)
η_3	3.588	4.432	4.209
	(0.690)	(1.494)	(1.869)
η_4	3.250	7.321	3.999
	(0.695)	(2.251)	(2.037)
η_5		4.448	4.010
		(1.510)	(1.982)
η_6		4.769	2.712
		(2.124)	(0.812)
η_7		5.140	3.544
		(1.640)	(1.426)
η_8		6.157	3.210
		(2.306)	(1.168)
η_9		7.139	3.133
		(2.214)	(1.520)
η_{10}		4.838	3.263
		(1.528)	(1.281)
η_{11}		6.682	3.418
		(1.859)	(1.620)
η_{12}		5.588	3.047
		(1.761)	(1.054)
η_{13}		4.279	4.713
		(2.203)	(2.014)
η_{14}		5.379	7.279
		(1.480)	(2.360)
η_{15}			2.431
			(1.973)
η_{16}			2.809
			(1.673)

Note: Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions).

[32] As expected, when we estimate the demand function (2) directly using the same firm-level IVs, the estimated demand elasticities are significantly biased towards zero: the mean elasticities are -0.005, 1.941, and -0.395 for the footwear, printing, and pharmaceutical industries, respectively.

The variations in demand elasticities across products, as documented above, lead to differences in markups at the firm-year level. These markups can be calculated as the weighted average of product markups considering their respective shares within firms. Across the three industries, the average markup at the firm-year level is 1.40 with a standard deviation of 0.14. The interquartile range (using the logarithm of the markups) is 0.16. This dispersion in markups is smaller than the estimate reported by De Loecker and Warzynski (2012), who found a standard deviation of 0.5 across a broader range of industries. Presumably, this difference is due to the fact that our variation of firm-year-level markups only captures the heterogeneous revenue shares and sets of products (as well as their associated markups) manufactured by different firms. Despite this narrower focus, the dispersion of markups at the firm-year level remains significant.

After all model parameters are estimated, we compute the firm-product-time varying output quality and productivity according to (11) and (15) in logarithm, respectively.[33] Nonetheless, these two measures are not directly comparable across and within firms. This is because the varieties (in the same product category) are of different quality levels and the unit of measurement across different products can be also different (e.g., grams vs. liters). However, the quality-adjusted output is readily comparable across firms and products, as shown by Melitz (2000), Orr (2022), and Li et al. (2023). Thus, we follow the literature to construct a combined measure that takes both quality and productivity into account. In our context, given the setup of quality-adjusted output in (1), we define a quality-adjusted productivity (ATFP) measure as[34]

$$\mathrm{ATFP}_{njt} = \tilde{\omega}_{njt} + \frac{1}{\eta_n - 1}\tilde{\xi}_{njt}. \tag{21}$$

As expected, ATFP reflects significant dispersion across firms even within a specific product category.[35] The mean interquartile range within a product is about 2.8 (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 quality-adjusted productivity, the interquartile range of $\tilde{\omega}_{jtn}$ within a product has a mean of 2.8, while the interquartile range of $\frac{1}{\eta_n-1}\tilde{\xi}_{njt}$ within a product has a mean of 1.8.[36] This suggests that

[33]We also compute firm-level intermediate input prices according to (13). We find that there is significant heterogeneity in intermediate prices, as documented by Ornaghi (2006) using observed intermediate price data.

[34]This measure is similar to the conventionally defined revenue-based productivity (a.k.a., TFPR).

[35]The distributions of ATFP by product, as well as the distributions of its components, $\tilde{\omega}_{njt}$ and $\tilde{\xi}_{njt}$, are reported in Figures A1, A2 and A3, respectively.

[36]The interquartile range of $\tilde{\omega}_{jtn}$ is slightly larger than that of ATFP because the two components of ATFP, productivity, and quality, are negatively related, as will be clear in Section 6.

the dominant force that drives the ATFP dispersion is productivity.[37]

Overall, our estimation results reflect reasonable parameter estimates and productivity and quality measures at the firm-product level. In the following sections, we turn to use these measures to explore the role of productivity and quality in shaping intra-firm performance heterogeneity.

6 Intra-firm Heterogeneity: the Role of Productivity and Quality

Given that the structural parameters are reasonably estimated and the rich distributions of productivity and quality demonstrate remarkable heterogeneity even within narrowly defined product lines, we now explore the pivotal question of what crucial new understandings about multi-product firms emerge from our analysis.

We uncover three noteworthy insights about multi-product firms. First, in Section 6.1, we examine how the heterogeneity in productivity and quality at the firm-product level influences within-firm relative performance across various products. Second, in Section 6.2, we establish a negative relationship at the firm-product level between these two dimensions of heterogeneity by estimating a trade-off between productivity and quality (i.e., the cost of quality). Lastly, our investigation reveals that quality-adjusted productivity, which accounts for both the costs and benefits of quality, exhibits a positive correlation with product quality. This analysis profoundly enriches our understanding of how production is organized within firms and sheds light on the key factors that shape intra-firm heterogeneity.

6.1 Productivity, Quality, and Product Rank

The literature traditionally emphasizes the role of productivity in explaining the growth and performance of firms and industries (e.g., Jovanovic, 1982; Hopenhayn, 1992; Ericson and Pakes, 1995; Melitz, 2003). Recently, a growing literature shows that demand is equally, if not more, important for firm turnover and growth (e.g., Foster et al., 2008; Pozzi and Schivardi, 2016; Kumar and Zhang, 2019). However, this strand of the literature usually focuses on across-firm analysis using firm-level data.

Taking into account the joint nature of production in multi-product firms, our estimation method allows us to uncover rich, flexible dimensions of heterogeneity within firms and explore the role

[37] Note that the demand heterogeneity (i.e., $\tilde{\xi}_{jtn}$) enters ATFP as $\frac{1}{\eta_n - 1}\tilde{\xi}_{njt}$ according to the definition of consumer utility (1). As a result, the adjustment due to the demand elasticity (as in $\frac{1}{\eta_n - 1}$) lowers its contribution to the dispersion of ATFP. Nonetheless, $\tilde{\xi}_{jtn}$ itself demonstrates a much higher degree of dispersion than that of $\tilde{\omega}_{jtn}$ (interquantile range of 5.8 vs. 2.8, respectively). The relatively larger dispersion of $\tilde{\xi}_{jtn}$ is consistent with the findings emphasizing the importance of demand heterogeneity in firm performance in the literature (e.g., Eslava et al., 2023).

of productivity and demand at the firm-product level. Specifically, we estimate the following regression equation to explore the relationship between the within-firm product rank of sales and firm-product-level productivity and quality:

$$\text{Log product rank, sales}_{jtn} = \alpha_{\tilde{\omega}}\tilde{\omega}_{jtn} + \alpha_{\tilde{\xi}}\tilde{\xi}_{jtn} + D_{jn} + D_{jt} + D_{tn} + \epsilon_{jtn}, \tag{22}$$

where the product rank (in logarithm) is defined based on the sales of products within firm-year pairs.[38] The rank of the top product (i.e., with the largest sales) is 1. An increase in rank indicates a product further away from the core competency of a firm. We include D_{jn}, D_{jt}, and D_{tn} as firm-product, firm-year, and product-year fixed effects, respectively, to capture different characteristics other than productivity and quality that vary at these levels and influence the product rank. This rich set of fixed effects is feasible due to the multiple-product nature of our dataset and helps minimize the problem of endogeneity that may be caused by missing variables, which is more likely to occur in across-firm analyses in which a less flexible set of fixed effects can be used.[39]

Table 4: Product rank (by sales level), productivity and quality

Dep. var.: Log product rank, sales	(1) All	(2) Footwear	(3) Printing	(4) Pharmaceutical
Productivity	-0.602***	-0.577***	-0.773***	-0.628***
	(0.140)	(0.168)	(0.176)	(0.209)
Quality	-0.170***	-0.218***	-0.175***	-0.253***
	(0.049)	(0.027)	(0.063)	(0.051)
Firm-Product FE	yes	yes	yes	yes
Firm-Year FE	yes	yes	yes	yes
Product-Year FE	yes	yes	yes	yes
Observations	9638	398	2981	6259
R-squared	0.887	0.950	0.916	0.884

Note: The dependent variable is the log of product rank based on sales within firm-year pairs. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions). *** $p < 0.01$.

[38] Equation (22) examines the relationship between product rank on the one hand and productivity and quality on the other by using product rank as the dependent variable. The purpose of the regression is to study directly the relative importance of productivity and quality for differences in sales across products within firms. Alternative regressions, usually with the purpose of investigating the gaps in productivity across products within firms, use product rank as an explanatory variable (controlling for quality when using productivity as the dependent variable). We report the results of the alternative regressions in Table A5 in the Appendix, which are in line with the results in the literature (e.g., Eckel and Neary, 2010; Mayer et al., 2014; Orr, 2022).

[39] The importance of using a flexible set of fixed effects will be more apparent when we estimate the cost of quality in Section 6.2.

The regression results are presented in Table 4. As anticipated, products closer to firms' core competence (i.e., with a lower rank value) have both higher productivity and quality. This is consistent with the literature that has theoretically postulated cost (i.e., productivity) or demand (i.e., quality) as key determinants of such within-firm variation in sales (e.g., Berman et al., 2012; Chatterjee et al., 2013; Mayer et al., 2014, 2021; Eckel et al., 2015; Arkolakis et al., 2021). Our results provide empirical support to both of these hypotheses.

In a context similar to ours, Orr (2022) uses firm-product-level estimates to document revenue efficiency gaps across products based on product rank. We depart from Orr (2022) by exploring the relative importance of quantity-based productivity and quality as determinants of within-firm heterogeneity. We show that, both in the overall sample as well as in each industry, productivity has a stronger impact on intra-firm performance. In particular, Column (1) of Table 4 shows that an increase of 1 percent in productivity and quality moves the rank of the product up by 0.602 percent and 0.170 percent, respectively, conditional on all other factors. Complementing the greater emphasis on quality in shaping performance across firms (e.g., Roberts et al., 2018; Kumar and Zhang, 2019), our result highlights that intra-firm performance is more influenced by productivity than by quality.

Table 5: Product rank (by sales growth), and growth in productivity and quality

Dep. var.: Log product rank, growth	(1) All	(2) Footwear	(3) Printing	(4) Pharmaceutical
Δ Productivity	-1.232***	-0.591	-1.492***	-1.491***
	(0.308)	(0.377)	(0.316)	(0.490)
Δ Quality	-0.339***	-0.358***	-0.335***	-0.617***
	(0.117)	(0.075)	(0.120)	(0.132)
Firm-Year FE	yes	yes	yes	yes
Product-Year FE	yes	yes	yes	yes
Observations	8311	307	2448	5556
R-squared	0.496	0.697	0.676	0.521

Note: The dependent variable is the log of product rank based on changes (growth) in sales between t and $t-1$ within firm-year pairs. Δ represents changes between t and $t-1$. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions). *** $p < 0.01$.

The robust pattern of productivity as a more important (compared to quality) determinant of the relative performance of products within a firm also emerges when product rank is measured in terms of the growth of sales within firms. The underlying question is which dimension is the dominant driving force of sales growth within firms over time. To answer this question, we examine a regression equation similar to (22):[40]

$$\text{Log product rank, growth}_{jtn} = \beta_{\tilde{\omega}} \Delta\tilde{\omega}_{jtn} + \beta_{\tilde{\xi}} \Delta\tilde{\xi}_{jtn} + D_{jt} + D_{tn} + \epsilon_{jtn}, \tag{23}$$

where the product rank (in logarithm) is defined based on the growth rate of sales of products within firm-year pairs. A higher rank value means a lower sales growth rate over time. Accordingly, $\Delta\tilde{\omega}_{jtn}$ and $\Delta\tilde{\xi}_{jtn}$ are the growth rates of productivity and quality, respectively. D_{jt} and D_{tn} are firm-year and product-year fixed effects, respectively. The results of the regression are presented in Table 5. As expected, growth in productivity and quality is positively associated with growth in sales within a firm (as implied by the product rank of sales growth). However, consistent with the pattern in Table 4 in terms of the rank based on the sales level, the sales growth rate is more responsive to the growth of productivity than to the growth of quality. The footwear industry stands as the only exception, where the coefficient of productivity growth is not statistically different from zero (presumably due to the lower number of observations), despite its larger magnitude compared to the coefficient of quality growth.

Overall, our analysis shows that, although both productivity and quality drive the performance and sales growth of products within firms, productivity is a dominant determinant. This finding suggests that when firms can optimally allocate resources to produce the most profitable products, productivity, as an internal capability within firms, plays a more important role than quality (demand), which can be heavily influenced by external market conditions.

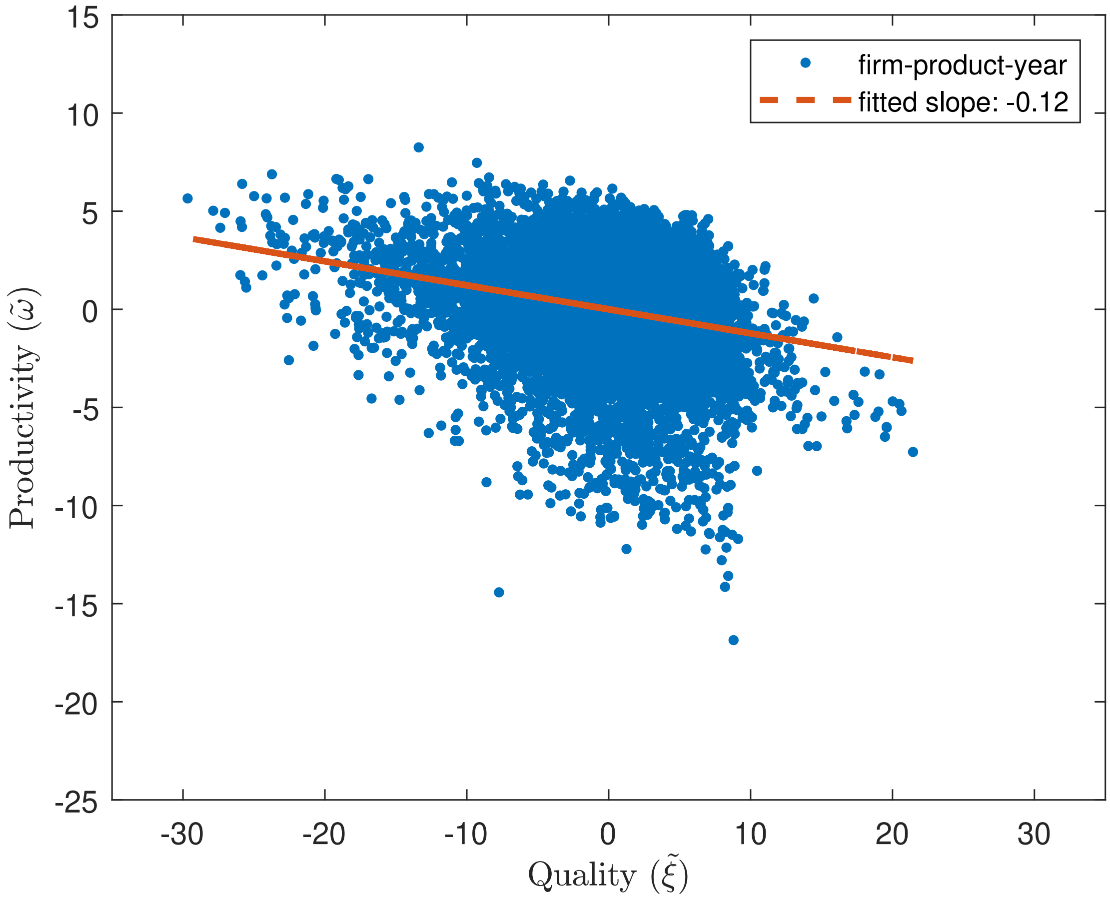
6.2 Trade-off between Productivity and Quality

As both productivity and quality influence the intra-firm performance of a product, it is natural to ask whether and how these different dimensions of within-firm heterogeneity are related. As a starting point, Figure 2 presents the raw relationship between our two key estimated measures of heterogeneity, i.e., (quantity-based) productivity ($\tilde{\omega}_{jtn}$) and quality ($\tilde{\xi}_{jtn}$).[41] This raw correlation is negative, consistent with the emerging literature (e.g., Grieco and McDevitt, 2017; Orr, 2022; Li et al., 2023) highlighting the trade-off between these two dimensions of firm heterogeneity.

[40]We use the rank of sales growth (rather than the sale growth itself) because our purpose is to investigate the intra-firm heterogeneity in growth performance. Also, note that this regression, (23), is not the first-order difference version of (22) because the dependent variable is defined as the product rank based on the growth rate of sales rather than as the change in the product rank.

[41]When we tease out the fixed effects at firm-product, firm-year, and product-year levels from $\tilde{\xi}_{jtn}$ to obtain a finer measure of quality (i.e., ξ_{jtn}) as defined in Section 2, the correlation is also negative. Of course, the firm-product fixed effects may contain parts of quality that only vary at the firm-product level. We find that the correlation is robustly negative when we include the firm-product fixed effects as a part of the quality measure.

Figure 2: The relationship between productivity and quality



Overall, this empirical pattern suggests that producing higher-quality products increases the marginal cost of production (by decreasing output quantity per unit of inputs, as highlighted in Grieco and McDevitt, 2017), which in turn reduces (quantity-based) productivity.[42] This is a relationship between productivity and quality that we allow for in (4) of the model but do not impose in our structural estimation. Consequently, this trade-off implies that (quantity-based) productivity should not be modelled by an auto-regressive process because it is strongly correlated with contemporaneous quality (or broadly, demand heterogeneity), which is endogenously influenced by firm choices based on a rich set of state variables.

To further estimate the trade-off between the two dimensions within firms, we propose to estimate a linear version of (4):

$$\tilde{\omega}_{jtn} = \omega_{jtn} - \gamma_{\xi}\xi_{jtn}, \tag{24}$$

where $\gamma_{\xi}\xi_{jtn}$ is interpreted as the cost (in terms of lowering productivity or raising marginal cost)

[42] Of course, this does not necessarily mean the correlation between (quantity-based) productivity and quality must be negative. But it does imply that the correlation is pushed downwards in the presence of a quantity-quality trade-off.

of increasing quality, holding inputs fixed. Formally, γ_ξ measures the elasticity of productivity with respect to the change in quality. We refer to it as the **cost-responsiveness** of quality.

Because our estimated measure of quality, $\tilde{\xi}_{jtn}$, comes as the residual demand from the demand function (2), its variation across products, firms, and over time may be naturally influenced by quality as well as demand heterogeneity such as demand conditions (e.g., macroeconomic conditions and market size), firm-brand image, product measurement unit (e.g., grams vs. liters), and firm-time specific measurement errors, as discussed in Section 2.1. In order to tease out the actual impact of quality (ξ_{jtn}) from those potential confounding factors and control for unobserved product and firm features, we include a flexible set of fixed effects in the regression analysis. More importantly, the rich set of fixed effects allows us to use the variation of productivity and quality over time *within firm-product* to avoid the potential bias caused by endogenous product selection (e.g., firms with high technical efficiency endogenously choosing to produce high quality products).

Specifically, the following regression equation is an empirical representation of (24):

$$\tilde{\omega}_{jtn} = -\gamma_\xi \tilde{\xi}_{jtn} + D_{jn} + D_{jt} + D_{tn} + \epsilon_{jtn}, \tag{25}$$

where D_{jn}, D_{jt}, and D_{tn} are firm-product, firm-year, and product-year fixed effects, respectively. ϵ_{jtn} contains technical efficiency (ω_{jtn}) and thus can be endogenously related to $\tilde{\xi}_{jtn}$ if firms choose to produce different quality products based on technical efficiency. In our implementation, we estimate (25) using different sets of IVs and the richest set of fixed effects available to jointly address the endogeneity problem.

The estimation results are presented in Table 6. The first three columns are Ordinary Least Squares (OLS) regressions of firm-product-level productivity on quality with increasingly rich specifications of the fixed effects. The comparison of the coefficient estimates in Columns (1)-(3) suggests that it is important to control for unobserved fixed effects to minimize the potential selection bias. In Columns (4)-(5), we use various instruments for the firm-product-level quality while allowing for the most flexible specification of firm-product, firm-year, and product-year fixed effects. In particular, in Column (4) the lagged quality of the same firm-product variety is used as the instrument, whereas in Column (5) the instrument is the lagged average quality of all other varieties (from other firms) of the same product. In Columns (6)-(9), we use both of these instruments together for all industries together and each industry separately, respectively. In all cases, the Kleibergen-Paap F test indicates a strong IV. In addition, the OLS result in Column (3), which controls for the most flexible set of fixed effects, is close to the IV result in Column (6), suggesting that leveraging the rich fixed effects offered by the multiple-product context is effective in controlling for unobservable covariates that influence both quality and productivity.[43]

Taken together, all the results in Table 6 indicate a negative trade-off between productivity and quality at the firm-product level. In particular, Column (6) shows that, on average, a 1-percent increase in quality lowers productivity (and thus increases marginal cost) by 0.198 percent, holding all other variables fixed.[44] This estimate is close to other estimates obtained using different approaches in various industries and countries. For example, Jaumandreu and Yin (2014) find strong negative correlations (ranging from -0.99 to -0.59, by industry) between their measures of cost advantage and demand advantage of exporters in the Chinese manufacturing industries. Grieco and McDevitt (2017) show that reducing a healthcare center's quality standards can increase its patient load, and they document a quality-quantity (number of patients) trade-off elasticity of -0.2 in the dialysis industry in the United States. Atkin et al. (2019) find that firms that make lower quality rugs produce more quickly among rug-makers in Egypt, demonstrating a reverse correlation between quantity productivity and quality productivity with an elasticity of -0.40. Orr (2022) estimates firm-product level measures of productivity and "product appeal" from the Indian machinery manufacturing industry and finds a negative correlation of about -0.28 between them. Using an objective output quality measure, Li et al. (2023) find that about half of the benefit created by quality is offset by the cost of producing the quality in the Chinese steel-making industry. Forlani et al. (2023) document an even stronger negative correlation (about -0.9) between demand and quantity-based productivity at the firm level in various Belgian manufacturing industries, suggesting a trade-off between the quality of a firm's products and their production cost.

Next, we further investigate the heterogeneity of the cost of quality across the product space and throughout the life cycle of products. First, we explore the relationship between the cost of quality and product differentiation, extending equation (25) to allow for variation in the cost-responsiveness of quality based on the elasticity of demand. The elasticity of demand (i.e., η), which varies at the product level, is estimated and presented in Table 3. A lower elasticity of demand indicates a higher degree of differentiation among product variety offered by different

[43]Firm-product fixed effects may contain parts of quality that only vary at the firm-product level. In an unreported regression, we exclude the firm-product fixed effects from the control variables and find that the results are quantitatively similar to the results in Table 6.

[44]The estimated cost of quality is significant, but not excessive. Taking into account the estimated demand elasticities, a simple accounting exercise shows that an increase in quality still leads to an improvement in revenue.

Table 6: Cost of quality

Dep. var.: Productivity	(1) OLS	(2) OLS	(3) OLS	(4) IV	(5) IV	(6) IV	(7) IV Footwear	(8) IV Printing	(9) IV Pharmaceutical
Quality	-0.122***	-0.183***	-0.204***	-0.197***	-0.204***	-0.198***	-0.236***	-0.182***	-0.216***
	(0.042)	(0.030)	(0.037)	(0.030)	(0.052)	(0.030)	(0.073)	(0.058)	(0.030)
Firm FE	no	yes	no	no	no	no	no	no	no
Product FE	no	yes	no	no	no	no	no	no	no
Year FE	no	yes	no	no	no	no	no	no	no
Firm-Product FE	no	no	yes	yes	yes	yes	yes	yes	yes
Firm-Year FE	no	no	yes	yes	yes	yes	yes	yes	yes
Product-Year FE	no	no	yes	yes	yes	yes	yes	yes	yes
Observations	11021	11020	9638	8160	8160	8160	292	2365	5503
R-squared	0.061	0.733	0.995	0.798	0.799	0.798	0.606	0.900	0.679
Kleibergen-Paap F				166.827	19.468	83.627	6.037	28.175	93.495
Hansen J						0.363	1.609	1.619	0.491

Note: The dependent variable is quantity-based productivity at the firm-product-year level. The instrumental variable in column (4) is lagged quality. The instrumental variable in column (5) is lagged average quality of all other varieties of the same product. The instrumental variables in columns (6)-(9) are lagged quality and lagged average quality of all other varieties of the same product. The Kleibergen-Paap F is a weak identification test. The Hansen J is a test of overidentifying restrictions. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.105$.

firms. As shown in Column (1) of Table 7, we observe that products with lower elasticities of demand face a more expensive trade-off between productivity and quality. This finding implies that the production of high-quality levels for more differentiated products entails increased costs. Specifically, Column (1) reveals that a 1-unit increase in the elasticity of demand results in a 2.3 percentage point rise in the cost-responsiveness of quality, which implies an increase by 7.4 percent in the overall impact of quality on productivity.[45] Considering that the cost of quality contributes to the marginal cost of production, our result suggests that firms producing high-quality products tend to charge higher prices, particularly if their products exhibit higher degrees of differentiation. This finding aligns with the conclusions drawn by Eckel et al. (2015), who demonstrate that product differentiation enhances the importance of quality in determining core competence.

Table 7: Cost of quality, by product differentiation and age

Dep. var.: Productivity	(1) IV	(2) IV	(3) IV
Quality	-0.310***	-0.223***	-0.328***
	(0.038)	(0.039)	(0.040)
Quality × η	0.023***		0.022***
	(0.007)		(0.007)
Quality × Age, log		0.013**	0.012**
		(0.007)	(0.005)
Age, log		-0.007	-0.039
		(0.105)	(0.082)
Firm-Product FE	yes	yes	yes
Firm-Year FE	yes	yes	yes
Product-Year FE	yes	yes	yes
Observations	8160	8160	8160
R-squared	0.823	0.801	0.826
Kleibergen-Paap F	21.056	40.989	14.186
Hansen J	9.042	1.457	11.100

Note: The dependent variable is quantity-based productivity at the firm-product-year level. The variable "Age" is calculated as the number of years since the first year in which a product variety appears in the data. The instrumental variables in column (1) are lagged quality, lagged average quality of all other varieties of the same product, and their interactions with the elasticity of substitution, η. The instrumental variables in column (2) are lagged quality, lagged average quality of all other varieties of the same product, and their interactions with log age. The instrumental variables in column (3) are all those included in columns (1) and (2). The Kleibergen-Paap F is a weak identification test. The Hansen J is a test of overidentifying restrictions. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions). *** $p < 0.01$, ** $p < 0.05$.

[45] This is calculated as the ratio of the coefficient on quality × η over that on quality, $\frac{0.023}{0.310}$.

Second, we investigate how the trade-off between productivity and quality evolves with the age of a product within a firm. To do so, we expand equation (25) to incorporate an interaction term between quality and the logarithm of product age, in addition to considering product age on its own. For each product-firm pair, product age is defined as the number of years since the initial appearance of the product variety in the sample. Column (2) of Table 7 presents the results of our analysis. The positive coefficient observed for the interaction between quality and product age indicates that the trade-off between productivity and quality becomes attenuated as a firm continues to produce a specific product for a longer duration. Essentially, firms that have accumulated substantial experience in manufacturing a particular product demonstrate improved production management capabilities, enabling them to achieve higher quality levels without compromising efficiency. Based on the estimate in Column (2), a simple calculation reveals that a 5-year experience in product manufacturing leads to approximately a 2.3 percentage point reduction in the impact of quality on productivity, which implies a decrease by 10.4 percent in the overall impact of quality on productivity.[46]

It is worth noting that the aforementioned results remain robust even when we simultaneously include the interactions between quality and the elasticity of demand, as well as between quality and product age, as depicted in Column (3) of Table 7. This robustness strengthens our confidence in the validity of the findings and underscores the significance of demand elasticity and product age in influencing the cost of quality.

Overall, the results obtained above demonstrate a robust negative relationship between quality and quantity-based productivity. However, when considering quality and quality-adjusted productivity (ATFP) which takes into account both the costs and benefits of quality as indicated by its definition in (21), a significantly positive relationship emerges. In Table 8, Columns (2)-(4) present the correlation coefficients between ATFP and quality at the firm-product level, which are 0.906, 0.334, and 0.592 for the three industries, respectively. This positive relationship is intuitive. While the cost of quality tends to lower ATFP as quality increases, the benefits of quality contribute to a positive association with ATFP. The dominance of the latter force results in an overall positive relationship between ATFP and quality. This finding aligns with previous analyses that emphasize firms with high production capability choose to produce high-quality output endogenously (e.g., Verhoogen, 2008; Kugler and Verhoogen, 2009, 2012; Feenstra and Romalis, 2014; Hottman et al., 2016; Fan et al., 2018). Our results not only highlight the positive sorting within firms but

[46] The calculation of the level of the impact is: $0.013 \times (\log(5+1) - \log(1)) = 0.023$. Relative to the overall impact of quality on productivity, the calculation is: $\frac{0.013\times(\log(5+1)-\log(1))}{0.223} = 0.104$.

also indicate that it is conditional upon acknowledging both the increasing cost and benefit of producing higher-quality products. This observation is consistent with the findings of Li et al. (2023), who utilize a firm-level objective quality measure from the Chinese steel industry, rather than using an estimated demand residual as a proxy for quality.

Table 8: Correlation between quality and ATFP

Quality	(1) All	(2) Footwear	(3) Printing	(4) Pharmaceutical
ATFP	0.440***	0.906***	0.334***	0.592***
	(0.084)	(0.053)	(0.116)	(0.102)

Note: The table reports the correlation coefficients are controlling for product and firm-year fixed effects. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions). *** $p < 0.01$.

Overall, our analysis on productivity and quality highlights the significance of considering the cost of quality and the relationship between these variables at the firm-product level. Importantly, a notable implication arises from the relationship: reducing the cost of quality (e.g., through long experience in production) not only contributes directly to an increase in the ATFP of a firm but also indirectly stimulates growth through intra-firm resource reallocation towards the production of higher-quality products, which subsequently enhances the firm's ATFP further. In the following section, we shift our focus to evaluating the cost of quality and study the role of product scope in firm growth through intra-firm resource reallocation resulting from a reduction in the cost of quality.

7 How Costly is Quality?

The results regarding the cost of quality are meaningful because they imply that a reduction in the cost-responsiveness of quality can lead to growth in ATFP. Intuitively, conditional on the underlying technical efficiency (i.e., ω) and product quality, a reduction of the cost-responsiveness of quality (i.e., γ_ξ) means a direct increase in quantity-based productivity (i.e., $\tilde{\omega}$) according to (24) and, thus, a corresponding increase in ATFP as defined in (21). More importantly, the impact on higher-quality products is larger for a given reduction in the cost responsiveness of quality. Thus, in the short run, multi-product firms can endogenously reallocate resources towards high-quality and high-productivity products, which consequently improves ATFP at the firm level. In addition, considering that product quality is endogenously chosen by firms based on productivity as emphasized in the literature (e.g., Verhoogen, 2008; Kugler and Verhoogen, 2009, 2012; Feenstra and Romalis, 2014; Fan et al., 2018), a cost of quality reduction implies an incentive for quality upgrading, thus increasing ATFP even further in the long run.

Our static empirical model does not allow us to estimate the endogenous reaction of quality choices in the long run, thus we focus on investigating the short-run effect of a reduction in the cost-responsiveness of quality, keeping the quality choices fixed.[47] To this end, we highlight the role of product scope in shaping productivity gains due to resource reallocation within firms. We conduct a counterfactual exercise by reducing the cost-responsiveness of quality and comparing the resulting ATFP at the firm level against that from the baseline scenario (i.e., with no reduction in the cost of quality). The improvement in ATFP is decomposed into a direct increase of ATFP due to the reduced cost of quality and intra-firm resource reallocation.

Specifically, in the counterfactual scenario, we reduce the cost-responsiveness of quality (γ_ξ) by 1 percent for all firm-product pairs. This leads to a direct improvement in quantity-based productivity: $\tilde{\omega}'_{jtn} = \tilde{\omega}_{jtn} + 0.01 \times \gamma_\xi \xi_{jtn}$, where $\tilde{\omega}'_{jtn}$ is the counterfactual productivity and $\tilde{\omega}_{jtn}$ and ξ_{jtn} are the baseline quantity-based productivity and quality, respectively. Here γ_ξ denotes the estimated cost-responsiveness of quality specific to each industry and reported in Columns (7)-(9) of Table 6. The direct improvement in quantity-based productivity drives an increase in ATFP at the firm-product level according to (21).[48]

More interestingly, there is an indirect improvement in firm-level ATFP due to intra-firm resource reallocation across products for multi-product firms. To see this mechanism, note that the 1-percent decline in γ_ξ leads to a differential improvement in the counterfactual productivity across products within firms, depending on the baseline quality level (ξ_{jtn}). For a product with higher quality, the resulting productivity improvement due to the reduction in the cost responsiveness of quality is larger. As a result, firms can react to the differential productivity improvement by re-optimizing their intra-firm allocation of inputs and outputs. Because ATFP and quality are positively related as documented in Section 6, multi-product firms tend to reallocate more production resources to products with higher ATFP and higher quality. Consequently, this reallocation leads to an indirect improvement in firm-level ATFP.

Both the direct and indirect improvements contribute to the increase in firm-level ATFP. To understand their magnitude and relative importance, we aggregate firm-level ATFP from firm-

[47] Thus, our results regarding the evaluation of the cost of quality in this section should be interpreted as the lower bound of the actual impact on firm performance.

[48] Throughout the analysis, we treat all the dynamic decisions (i.e., product quality, scope, and investment) described in Section 2.4 as fixed.

product-level ATFP using sales as weights. We apply the within-industry across-firm decomposition proposed by Olley and Pakes (1996) to compute the intra-firm decomposition. That is, for each firm j in period t,

$$\mathrm{ATFP}_{jt} = \overline{\mathrm{ATFP}}_{jt} + \sum_{n \in \Lambda_{jt}} (s_{jtn} - \overline{s}_{jt})(\mathrm{ATFP}_{jtn} - \overline{\mathrm{ATFP}}_{jt}), \tag{26}$$

where $\overline{\mathrm{ATFP}}_{jt}$ is the simple average of the exponent of quality-adjusted productivity, ATFP_{jtn}, across products produced by the same firm. s_{jtn} is the within-firm sales share of product n by firm j in period t. $\overline{s}_{jt}$ is the simple average of the sales shares (that is, the inverse of the product scope). Intuitively, an increase in firm-level ATFP can be caused by an increase in ATFP of all products as well as a reallocation of resources towards more productive products. Accordingly, intra-firm resource reallocation is defined as the difference of the covariance term (the second term on the right-hand side) in (26) between the counterfactual scenario and the baseline scenario. To obtain the overall improvement in ATFP at the industrial level, we aggregate firm-level ATFP improvement using firms' total sales as weights. The relative contribution of intra-firm resource reallocation to the firm-level ATFP improvement is aggregated to the industry level in the same way.

Table 9: Impact of 1-percent reduction in cost-responsiveness of quality on ATFP

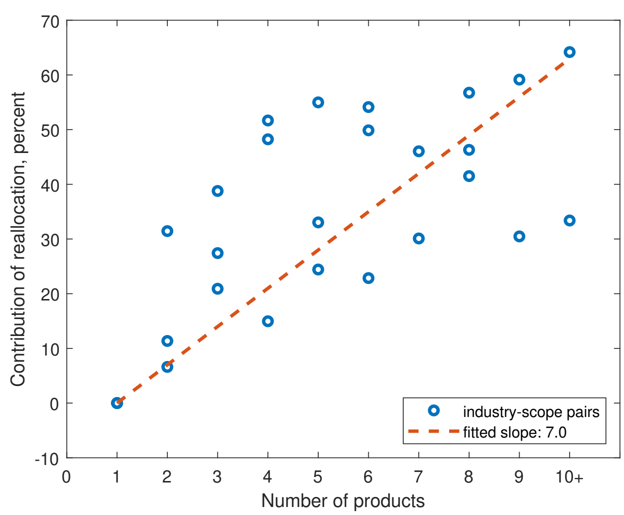
	All firms				MPF only
Industry	All	Footwear	Printing	Pharmaceutical	All
Total improvement, percent	2.562	1.202	2.711	2.670	2.653
	(0.249)	(0.357)	(0.412)	(0.301)	(0.269)
Intra-firm reallocation,					
percent	0.777	0.121	0.434	0.890	0.907
	(0.103)	(0.036)	(0.119)	(0.126)	(0.119)
percentage relative to total	30.3	10.0	16.0	33.3	34.2
	(4.2)	(1.9)	(3.9)	(5.0)	(4.8)

Note: The improvement in ATFP at the industry level is measured in percentage and calculated as the weighted average of the improvements in ATFP at the firm-year level with firms' total sales in the baseline scenario as weights. Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions).

Table 9 reports the overall improvement in firm-level ATFP as well as the contribution from the intra-firm resource reallocation of multi-product firms in the three industries. A 1-percent decline in the cost-responsiveness of quality leads to an improvement in ATFP by more than 1.2, 2.7, and 2.6 percent for the footwear, printing, and pharmaceutical industries, respectively. This is a sizable magnitude. More importantly, the contribution of the within-firm resource reallocation accounts for roughly 10 percent to 33 percent of the overall improvement in ATFP across the three industries, as reported in Table 9. This is essentially a lower bound of the contribution because the calculation is based on all firms including the single-product firms that experience, by definition, zero within-firm reallocation. When focusing on multi-product firms only, the contribution is on average approximately 34 percent across the three industries. This result establishes the economic significance of the cost of quality within multi-product firms as a channel impacting overall quality-adjusted productivity.

A large literature on resource reallocation focuses on across-firm analysis and shows that much of the aggregate productivity growth is attributable to the resource reallocation towards 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). Complementary to the literature, our firm-product-level analysis shows that the contribution of within-firm resource reallocation is also sizable. Interestingly, compared to the footwear industry, the relatively higher intra-firm contribution in the printing and pharmaceutical industries is consistent with the relatively larger number of products in these industries. Indeed, as shown in Table 1, firms in the printing and pharmaceutical industries produce 3.7 and 6.9 products on average, respectively, while firms in the footwear industry produce 1.3 products. Intuitively, a larger product scope allows for a greater potential to reallocate resources across products.

Figure 3: Contribution of within-firm resource reallocation to ATFP growth



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

To unpack such a heterogeneous pattern, we group firms by the number of products produced and compute the sales-weighted average contribution of intra-firm resource reallocation to firm-level ATFP improvement (due to the reduction in the cost of quality). This computation is conducted for each industry. We plot the relationship between product scope and the contribution of intra-firm reallocation (in percentage) in Figure 3.[49] Each dot represents the average contribution of within-firm reallocation by product scope and industry. The dashed line represents the fitted line obtained from a simple OLS regression of within-firm reallocation against product scope. The upward-sloping fitted line establishes that, on average, the role of within-firm reallocation increases in firms with a larger scope with more room for within-firm adjustment. The slope of the fitted line suggests that producing one more product can increase the contribution of within-firm reallocation in improving ATFP by 7 percent. In sum, our results highlight that multi-product firms with larger scope experience larger productivity gains when the cost of quality is lower. This reveals a new mechanism for enhancing the performance of multi-product firms.

8 Conclusion

Multi-product firms account for a significant share of our economy. Yet, the traditional firm-level analysis in the literature masks the intra-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 the production of different products within firms nor restrictions on how productivity evolves over time. 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. We find that, in contrast to the emphasis on the role of quality (demand) in explaining the across-firm performance heterogeneity in the literature, productivity is a dominant force that drives the intra-firm revenue heterogeneity. However, firms face a trade-off between upgrading quality and productivity, which we define as the cost of quality. Such cost of quality is highly heterogeneous across products and changes over time. The impact of producing high-quality products is larger for products with higher degrees of differentiation; but, over time, the impact decreases when firms produce the products for a longer time. After taking both the costs and benefits of quality into account, the recovered quality-adjusted productivity shows a strong positive intra-firm correlation with quality.

[49] The relationship is similar when the figure is plotted by industry. This suggests that although product scope is not directly comparable across industries, the pattern reflected in the figure is robust.

To understand how costly quality is for productivity growth and intra-firm resource allocation, we conduct a counterfactual exercise where we reduce the cost-responsiveness of quality by 1 percent. We show that the reduction can lead to substantial productivity gains, especially for multi-product firms. Importantly, a sizable portion of the productivity gain of multi-product firms is due to the within-firm reallocation of resources towards more-productive and higher-quality products. In particular, we show that a larger product scope allows more room for intra-firm resource reallocation, leading to a higher productivity gain when there is a reduction in the cost of quality. This result establishes the quantitative significance of intra-firm resource reallocation in enhancing the performance of multi-product firms that dominate manufacturing production. This channel, thus, has strong potential implications for aggregate productivity growth.

References

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

Arkolakis, C., S. Ganapati, and M.-A. Muendler (2021). The extensive margin of exporting products: A firm-level analysis. *American Economic Journal: Macroeconomics 13*(4), 182–245.

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. In *AEA Papers and Proceedings*, Volume 109, pp. 444–49.

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.

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, 10). Recasting the Iron Rice Bowl: The Evolution of China's State Owned Enterprises. *The Review of Economics and Statistics 99*(4), 735–747.

Berman, N., P. Martin, and T. Mayer (2012). How do different exporters react to exchange rate changes? *The Quarterly Journal of Economics 127*(1), 437–492.

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 (2023). Multi-product markups.

Caselli, M. (2018, August). Do all imports matter for productivity? Intermediate inputs vs capital goods. *Economia Politica: Journal of Analytical and Institutional Economics 35*(2), 285–311.

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.

Chatterjee, A., R. Dix-Carneiro, and J. Vichyanond (2013). Multi-product firms and exchange rate fluctuations. *American Economic Journal: Economic Policy 5*(2), 77–110.

Chen, Y., M. Igami, M. Sawada, and M. Xiao (2021). Privatization and productivity in china. *The RAND Journal of Economics 52*(4), 884–916.

Chen, Z. and M. Liao (2022). Production Function Estimation for Multi-Product Firms. Working Paper 3968432, SSRN.

Collard-Wexler, A. and J. De Loecker (2015). Reallocation and technology: Evidence from the us steel industry. *American Economic Review 105*(1), 131–71.

Das, S., M. J. Roberts, and J. R. Tybout (2007, May). 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 Vol. 79, No. 5*, pp. 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.

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.

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

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

Eckel, C., L. Iacovone, B. Javorcik, and J. P. Neary (2015). Multi-product firms at home and away: Cost-versus quality-based competence. *Journal of International Economics 95*(2), 216–232.

Eckel, C. and J. P. Neary (2010, January). Multi-product firms and flexible manufacturing in the global economy. *Review of Economic Studies 77*(1), 188–217.

Ericson, R. and A. Pakes (1995, January). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies 62*(1), 53–82.

Eslava, M., J. Haltiwanger, and N. Urdaneta (2023). The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand, and Wedges. *The Review of Economic Studies forthcoming*.

Fan, H., Y. A. Li, and S. R. Yeaple (2018). On the relationship between quality and productivity: Evidence from china's accession to the wto. *Journal of International Economics 110*, 28–49.

Feenstra, R. C. and J. Romalis (2014, May). International prices and endogenous quality. *The Quarterly Journal of Economics 129*(2), 477–527.

Forlani, E., R. Martin, G. Mion, and M. Muûls (2023, 04). Unraveling Firms: Demand, Productivity and Markups Heterogeneity. *The Economic Journal forthcoming*.

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, May). 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, February). Techies, Trade, and Skill-Biased Productivity. CEPR Discussion Papers 15815, C.E.P.R. Discussion Papers.

Hopenhayn, H. A. (1992, September). Entry, Exit, and Firm Dynamics in Long Run Equilibrium. *Econometrica 60*(5), 1127–1150.

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.

Jaumandreu, J. and H. Yin (2014). Cost and product advantages: A firm-level model for the chinese exports and industry growth. working paper, Boston College.

Jovanovic, B. (1982, May). Selection and the Evolution of Industry. *Econometrica 50*(3), 649–670.

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

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.

Kugler, M. and E. Verhoogen (2009). Plants and imported inputs: New facts and an interpretation. *American Economic Review 99*(2), 501–07.

Kugler, M. and E. Verhoogen (2012). Prices, plant size, and product quality. *Review of Economic Studies 79*(1), 307–339.

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)*(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 (2023). Output Quality, Productivity, and Demand Advantage: Evidence from the Chinese Steel Industry. Unsw 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, February). Does External Monitoring from the Government Improve the Performance of State-Owned Enterprises? *The Economic Journal 132*(642), 675–708.

Mayer, T., M. J. Melitz, and G. I. Ottaviano (2014). Market size, competition, and the product mix of exporters. *American Economic Review 104*(2), 495–536.

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.

Melitz, M. J. (2003, November). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica 71*(6), 1695–1725.

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), 000–000.

Panzar, J. C. and R. D. Willig (1977). Economies of scale in multi-output production. *The Quarterly Journal of Economics 91*(3), 481–493.

Panzar, J. C. and R. D. Willig (1981). Economies of scope. *The American Economic Review 71*(2), 268–272.

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

Roberts, M., D. Y. Xu, X. Fan, and S. Zhang (2018, 11). The Role of Firm Factors in Demand, Cost, and Export Market Selection for Chinese Footwear Producers. *Review of Economic Studies 85*(4), 2429–2461.

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

Valmari, N. (2022). Estimating production functions of multiproduct firms. *Review of Economic Studies 130*(11), 000–000.

Verhoogen, E. A. (2008). Trade, quality upgrading, and wage inequality in the mexican manufacturing sector. *The Quarterly Journal of Economics 123*(2), 489–530.

Appendices

A Additional Figures and Tables

Table A1: Product list, manufacturing of footwear, mainly of leather (class 324001)

Industry	Product description	Code
324001	Cow leather, for men	1
324001	Cow leather, for women	2
324001	Cow leather, for kids	3
324001	Others	99

Table A2: Product list, printing and binding (class 342003)

Industry	Product description	Code
342003	Printing of Calendars and almanacs	5
342003	Folding boxes	6
342003	Labels and prints	13
342003	Brochures and catalogs	14
342003	Continuous forms	15
342003	Accounting, administrative and tax forms	16
342003	Telephone directories	17
342003	Books	18
342003	Journals	19
342003	Checks	21
342003	Commemorative and business cards	23
342003	Commercial flyers	24
342003	Posters	25
342003	Others	99

Table A3: Product list, manufacturing of pharmaceutical products (class 352100)

Industry	**Product description**	**Code**
352100	Medicinal products, for human use with a specific action, anti-infectious: Bactericides	11
352100	Antiparasitics	13
352100	Dermatological	15
352100	Other products with specific action not included in other categories	19
352100	Medicinal products for human use for specialties with action on: Circulatory system	21
352100	Digestive system and metabolism	22
352100	Human musculoskeletal system	23
352100	Respiratory system	24
352100	Sensory organs	25
352100	Genitourinary organs, except hormones	26
352100	Blood and hematopoietic organs	27
352100	Central nervous system	28
352100	Hormones	32
352100	Vitamins and Vitamin Compounds	43
352100	Non-therapeutic products	59
352100	Others	99

Table A4: Within-firm product shares by product scope

	Product rank (by sales level)				
Product scope	1	2	3	4	5+
1	1.000				
2	0.783	0.217			
3	0.675	0.238	0.087		
4	0.560	0.283	0.117	0.040	
5+	0.443	0.204	0.124	0.083	0.146

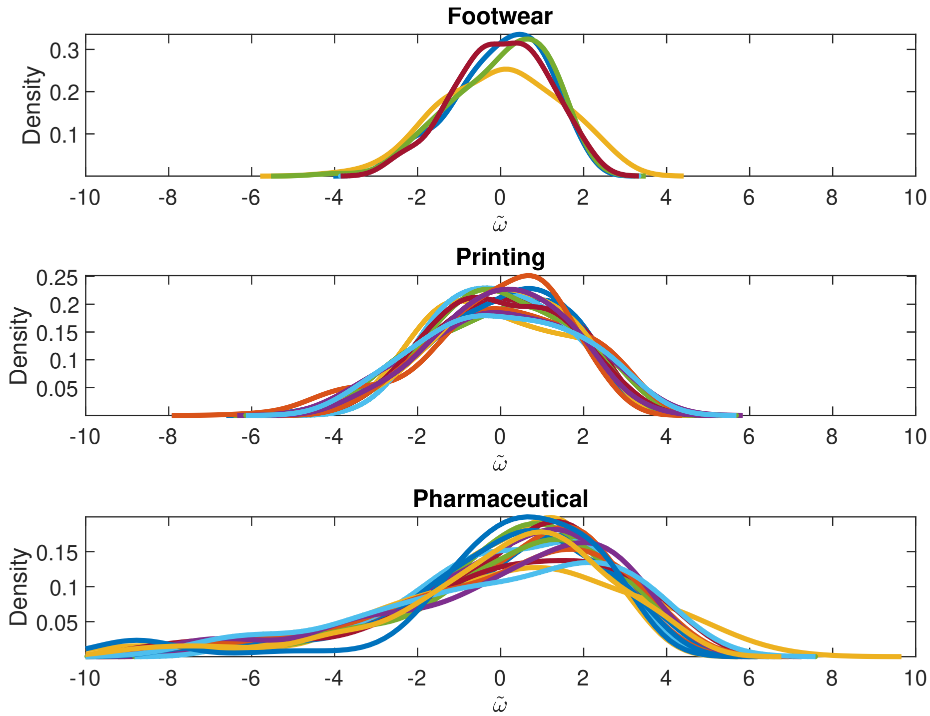
Note: All firm-year pairs producing 5 products or more are clustered in the "5+" group. All products ranked 5 or lower are clustered in the "5+" group.

Table A5: Productivity, quality, and product rank (by sales level)

	(1)	(2)	(3)	(4)
Dep. var.:	Quality	Productivity	Productivity	ATFP
Product rank	-0.489***	0.039***	-0.069***	-0.128***
	(0.043)	(0.006)	(0.020)	(0.033)
Quality			-0.221***	
			(0.052)	
Firm-Product FE	yes	yes	yes	yes
Firm-Year FE	yes	yes	yes	yes
Product-Year FE	yes	yes	yes	yes
Observations	9638	9638	9638	9638
R-squared	0.892	0.974	0.996	0.997

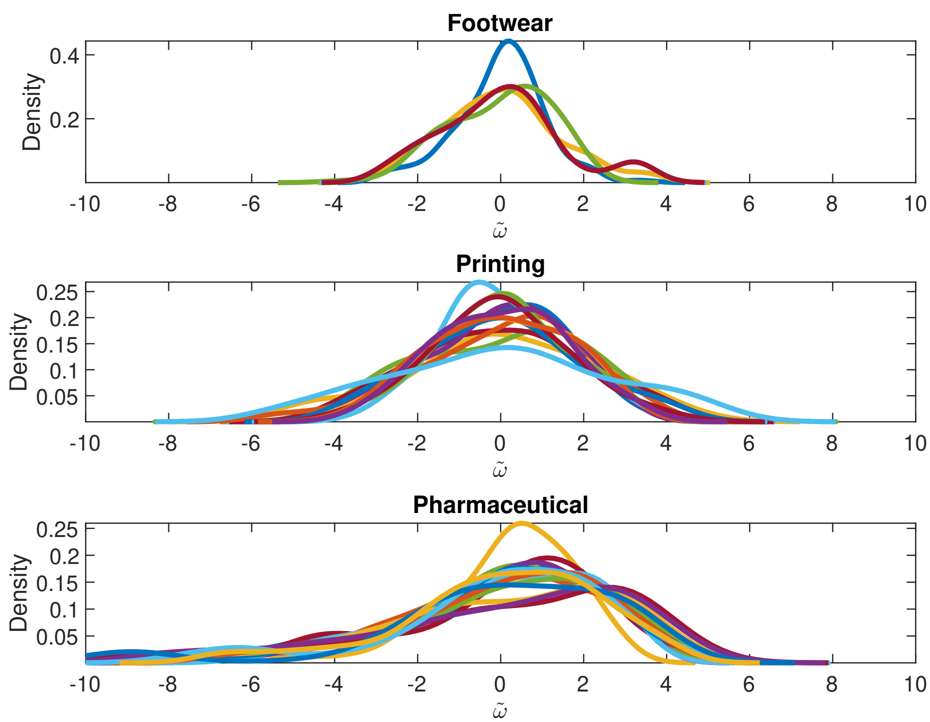
Note: The dependent variable is the log of quality at the firm-product-year level in Column (1), the log of productivity at the firm-product-year level in Columns (2)-(3), and quality-adjusted productivity (ATFP) at the firm-product-year level in Column (4). Bootstrapped standard errors clustered at the firm level and stratified by industry and scope are shown in parentheses (250 repetitions). *** $p < 0.01$.

Figure A1: Distribution of quality-adjusted productivity, ATFP



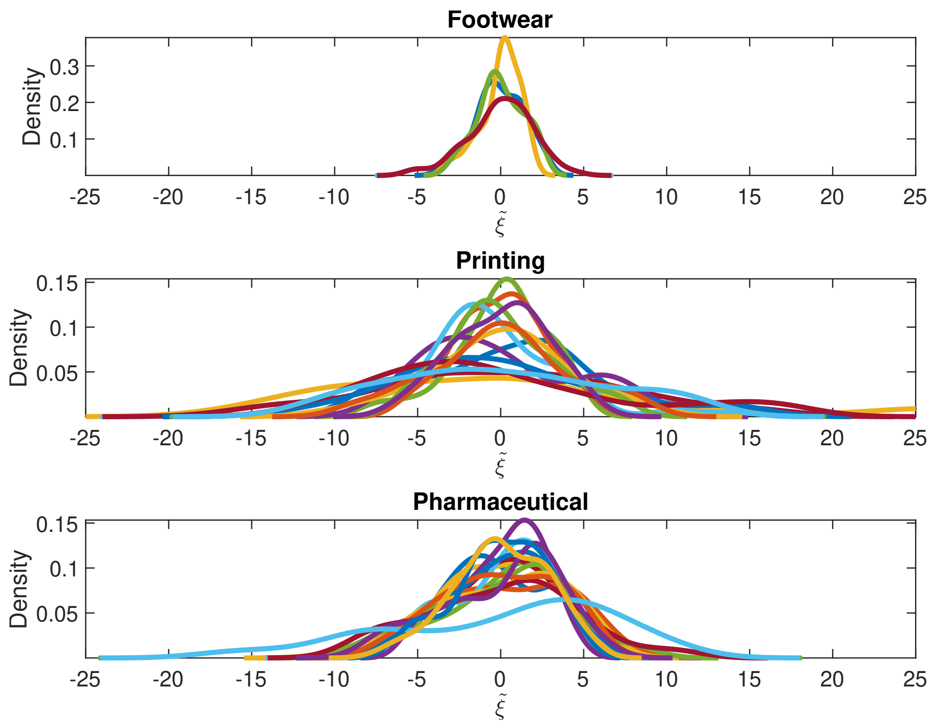
Notes: ATFP is demeaned, and only products with at least 100 observations are included.

Figure A2: Distribution of productivity, $\tilde{\omega}$



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

Figure A3: Distribution of quality, $\tilde{\xi}$



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

B Monte Carlo Exercises

In this appendix, we present the results of Monte Carlo exercises to demonstrate the performance of our estimation method.

In this Monte Carlo setting, the choice of product sets is exogenous and random. Wage rate, material prices, and capital stock are serially correlated.[50] The levels of productivity and quality of any given product are not only serially correlated over time but also negative correlated with each other. With this setting, the Monte Carlo exercises consist of N replications of simulated data sets of J firms in T years, given a set of true parameters of interest for 5 products, namely, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \alpha_L, \alpha_M, \alpha_K, \sigma, \rho)$.

Specifically, in each replication, we simulate productivity ($\tilde{\omega}_{jtn}$) and quality ($\tilde{\xi}_{jtn}$) for each product n, firm j, and time t. We also simulate the wage rate ($P_{L_{jt}}$), the material price ($P_{M_{jt}}$) and the capital stock (K_{jt}) for each firm j and time t. All of these variables are serially correlated. In addition, we simulate the negative relationship between productivity and quality as documented in the paper by allowing for a negative correlation r between the shocks in their evolution processes. Specifically, the evolution process of each of these variables for each firm follows an AR(1) process:

$$\tilde{\omega}_{jtn} = g^n_{0\omega} + g^n_{\omega}\tilde{\omega}_{jt-1n} + \varepsilon^{\omega}_{jtn}, \quad \forall n,$$

$$\tilde{\xi}_{jtn} = g^n_{0\xi} + g^n_{\xi}\tilde{\xi}_{jt-1n} + \varepsilon^{\xi}_{jtn}, \quad \forall n,$$

$$\ln(P_{Ljt}) = g_{0\ell} + g_{\ell}\ln(P_{Ljt-1}) + \varepsilon^{\ell}_{jt},$$

$$\ln(P_{Mjt}) = g_{0m} + g_{m}\ln(P_{Mjt-1}) + \varepsilon^{m}_{jt},$$

$$\ln(K_{jt}) = g_{0k} + g_{k}\ln(K_{jt-1}) + \varepsilon^{k}_{jt},$$

where ε is the innovation shock realized in period t, which is assumed to be a normally distributed error term with zero mean and standard deviation $sd(\varepsilon)$. While the shocks in the processes of P_{Ljt}, P_{Mjt}, and K_{jt} are i.i.d., those of $\tilde{\omega}_{jtn}$ and $\tilde{\xi}_{jtn}$ are correlated with a coefficient of r. Although the evolution of the capital stock is exogenous in this setup, the Monte Carlo result is similar if investment (and hence the capital stock) depends on productivity and quality levels.

Given these variables, we use the firm's static profit maximization problem to derive a sequence of optimal choices of labor and material inputs (L_{jt} and M_{jt}), the optimal output quantity (Q_{jtn}) and price (P_{jtn}) for firm j and product n in each period t.

In this way, we generate a data set of variables for the Monte Carlo experiments. Among them, we use the following variables for the estimation procedure (including the sets of IVs) described in Section 3: $\{Q_{jt1}, \ldots, Q_{jt5}, R_{jt1}, \ldots, R_{jt5}, K_{jt}, L_{jt}, E_{L_{jt}}, E_{M_{jt}}\}$. The values of the parameters used for the data generation process are reported in Table A6. The mean estimates of the key parameters, together with their corresponding standard errors, are reported in Table A7. Overall, the result shows that our estimation recovers the true parameters of the production and demand functions well.

[50] The Monte Carlo result is similar if the evolution of capital stock depends on an investment rule which is a function of capital stock and the levels of productivity and quality.

Table A6: Monte Carlo Parameter Values

Parameter	Description	Value
$\eta_1, \eta_2, \eta_3, \eta_4, \eta_5$	Demand elasticities	7, 6, 5, 4, 3
σ	Elasticity of substitution	2
α_L	Distribution parameter of labor	0.2
α_M	Distribution parameter of material	0.6
α_K	Distribution parameter of capital	0.2
g_ω^1, g_ω^2, g_ω^3,g_ω^4, g_ω^5	Persistence parameters in productivity evolution	0.75, 0.7, 0.65, 0.6, 0.55
g_ξ^1, g_ξ^2, g_ξ^3,g_ξ^4, g_ξ^5	Persistence parameter in quality evolution	0.75, 0.7, 0.65, 0.6, 0.55
g_l	Persistence parameter in wage rate evolution	0.8
g_m	Persistence parameter in material price evolution	0.8
g_k	Persistence parameter in capital evolution	0.8
r	Correlation between productivity and quality shocks	-0.2
$sd(\varepsilon^\omega)$	Standard deviation of productivity shock	0.02
$sd(\varepsilon^\xi)$	Standard deviation of quality shock	0.02
$sd(\varepsilon^\ell)$	Standard deviation of wage rate shock	0.1
$sd(\varepsilon^m)$	Standard deviation of material price shock	0.1
$sd(\varepsilon^k)$	Standard deviation of capital stock shock	0.1
$sd(u)$	Standard deviation of revenue measurement error (u)	0.01
T	Number of periods	15
J	Number of firms	400
N	Number of Monte Carlo replications	300

Table A7: Monte Carlo Estimates of Production and Demand Function Parameters

Parameter	True	Estimate
$\frac{\eta_1-1}{\eta_2-1}$	1.200	1.199
		(0.021)
$\frac{\eta_1-1}{\eta_3-1}$	1.500	1.499
		(0.027)
$\frac{\eta_1-1}{\eta_4-1}$	2.000	1.999
		(0.037)
$\frac{\eta_1-1}{\eta_5-1}$	3.000	3.002
		(0.053)
α_L	0.200	0.200
		(0.002)
α_M	0.600	0.600
		(0.001)
α_K	0.200	0.200
		(0.002)
σ	2.000	2.000
		(0.010)
ρ	1.100	1.101
		(0.009)
η_1	7.000	7.000
		(0.350)
η_2	6.000	6.003
		(0.254)
η_3	5.000	5.001
		(0.204)
η_4	4.000	4.002
		(0.157)
η_5	3.000	2.998
		(0.102)

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