The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand and Wedges.*

Marcela Eslava† John Haltiwanger‡ Nicolas Urdaneta §

February 10, 2023

Abstract

What determines the distribution of establishments in terms of size and life-cycle growth? How are those determinants related to aggregate productivity? We provide novel answers by developing a framework that uses price and quantity information on establishments' outputs and inputs to jointly estimate the demand and production parameters, and subsequently, establishments' quality-adjusted productivity, deriving both micro-level and aggregate implications. We find that the dominant source of variation in establishment size is variation in quality/product appeal but that variation in technical efficiency plays an important supporting role. Multiple factors dampen dispersion in establishment size including dispersion in input (quality-adjusted) prices, markups, and residual wedges. Relatively moderate dampening factors induce large aggregate allocative efficiency losses relative to their absence. We show that joint estimation of the parameters of the demand and production function crucially affects inferences on the determinants of the size distribution of firms and their implications for aggregate productivity.

Keywords: size distribution of manufacturing plants; productivity; allocative efficiency; quality; markups.

JEL codes: L11;O14;O47

*We thank Alvaro Pinzón for superb research assistance, and Innovations for Poverty Action, CAF and the World Bank for financial support for this project. We also thank DANE for permitting access to the microdata of the Annual Manufacturing Survey, as well as DANE's staff for advice in the use of these data. The use and interpretation of the data are the authors' responsibility. We gratefully acknowledge the comments of David Atkin, David Baqaee, Irene Brambilla, Ariel Burstein, Steven Davis, Manuel García-Santana, Peter Klenow, Virgiliu Midrigan, Stephen Redding, Diego Restuccia, Robert Shimer, Chad Syverson, Gabriel Ulyssea, Venky Venkateswaran, Daniel Xu and those from participants at the 2017 Society for Economic Dynamics Conference; 2017 meeting of the European chapter of the Econometric Society; 2017 NBER Productivity, Development, and Entrepreneurship workshop; the 2016 Trade and Integration Network of LACEA; the 2020 Ridge Forum in Growth and Development in Macroeconomics; and seminars at Universitat Pompeu Fabra, Universitat Autónoma de Barcelona, the University of Chicago, UCLA, LSE, Oxford and Universidad de Los Andes.

†Universidad de Los Andes, Bogotá. meslava@uniandes.edu.co

‡University of Maryland at College Park. haltiwan@econ.umd.edu

§Duke University and Universidad de Los Andes, Bogotá. n.urdaneta@duke.edu

1 Introduction

A prevalent feature of market economies is heterogeneity of firm and establishment size, growth, and a host of establishment attributes correlated with size (e.g., productivity, exports, survival). What are the sources of such heterogeneity, how does the answer matter for aggregate productivity and welfare? A crucial insight from the macro misallocation literature is that there are wedges (often referred to as distortions) impacting establishment size relative to what would be implied by establishment true productivity, and that this leads to aggregate productivity losses, especially in developing economies. Contributions in trade and IO have focused on how firm/establishment size is impacted by attributes such as demand (quality/appeal), markups, or costs, finding that idiosyncratic demand-side factors dominate.[1]

How do these findings relate to each other? Do wedges lie mainly on the cost or demand sides? What sources of heterogeneity across productive units are most harmful for aggregate activity and which are most enhancing, and how is that harm reflected in the size distribution of firms? We examine these questions by developing a unified conceptual, measurement, and estimation structure that accounts for a uniquely rich set of establishment (plant) attributes, and taking it to detailed data on manufacturing establishments. Our framework takes advantage of data on output and input prices and quantities to measure and estimate the role of these detailed establishment attributes. We consider establishment-level quality shifters, markups, and two distinct dimensions of idiosyncratic marginal costs: technical efficiency and quality-adjusted input prices, including wages, material prices, and, in an extension, idiosyncratic user cost of capital inclusive of factor-biased distortions. Residual wedges help account for the differences between the size distribution implied by theory and the data even after incorporating all of the components separately measured in our analysis.[2]

In the face of data constraints, assessing the roles of each of these different margins simultaneously has not been possible. True productivity (best interpreted as a composite of technical efficiency and quality/appeal) and wedges are typically identified from structures that exploit micro data on revenue and input expenditures, while structures that use product-level data on output prices and quantities have been used to separately identify quality, costs, and markups. We use detailed product-level data on quantities and prices for outputs and inputs from the Colombian Annual Manufacturing Survey. This is a uniquely rich census of non-micro manufacturing establishments with data on quantities and prices, at the detailed

[1]The misallocation literature is extensive. Prominent examples are Restuccia and Rogerson (2008, 2017); Hsieh and Klenow (2009, 2014); Guner, Ventura and Xu (2008); Midrigan and Xu (2014); Bartelsman, Haltiwanger and Scarpetta (2013); Bento and Restuccia (2017); Adamopoulos and Restuccia (2014). Quality is the focus in Brooks (2006); Fieler, Eslava and Xu (2018); Hallak and Schott (2011); Khandelwal (2010); Kugler and Verhoogen (2011); Manova and Zhang (2012). Hottman, Redding and Weinstein (2016) recently integrated demand, markups, and residual costs into an estimation framework, but not wedges (i.e. departures from the model) Technical efficiency vs. demand is emphasized in Foster, Haltiwanger and Syverson (2008, 2016); Jaumandreu and Mairesse (2010). De Loecker and Warzynski (2012); De Loecker, Eeckhout and Unger (2020) have focused on markups using an indirect approach with only revenue and expenditure data.

[2]Our approach whittles down the contribution of unexplained residual wedges considerably relative to the literature. As has been emphasized in the literature, such residual wedges might reflect a host of factors including policy and institutional distortions, adjustment costs, information frictions, financial frictions, and labor market frictions (see, e.g. Asker, Collard-Wexler and De Loecker, 2014; David and Venkateswaran, 2019; Midrigan and Xu, 2014; Guner, Ventura and Xu, 2008)

product class, for outputs and inputs. It follows individual plants for up to thirty years, allowing us to investigate the role of different attributes over medium- and long-term life cycle growth.

In the model, which nests the Hsieh and Klenow (2009) model on the production side and that proposed by Hottman, Redding and Weinstein (2016) on the demand side, consumers value both the quantities and qualities consumed of goods produced by (multi-product) establishments. The scale of an establishment is its choice, as a function of a set of attributes: quality/appeal to consumers of its bundle of products, efficiency of its production process, the input prices it faces, its markup, and other characteristics known to the establishment but unmeasured by the econometrician. The model delivers an expression that allows decomposing variation in establishment size in the cross section and over the life cycle into the contribution of each of these attributes, and another that relates aggregate allocative efficiency to each of these sources of establishment heterogeneity.

As in Hsieh and Klenow (2009) (HK henceforth), establishment size in an efficient world with quality differentiation would be determined solely by a composite of technology and quality (the attributes valued by consumers). We denominate that composite as quality-adjusted productivity. There are wedges between efficient and actual size. Relative to their seminal work, we unpack efficient size into its quality vs. efficiency components, and wedges into those linked to idiosyncratic markups, dispersion in input prices, and other factors captured in residual wedges. We quantify the role of each in the distribution of establishment size and in aggregate efficiency. As in Hottman, Redding and Weinstein (2016) (henceforth HRW), in turn, we determine how plant size is impacted by quality/appeal and the markup, while also disentangling the residual "marginal cost" (as labeled in HRW) into the contributions of technical efficiency in production, input price dispersion and residual wedges that the econometrician cannot appropriately attribute to cost or demand factors.

The measurement of these attributes of establishments requires, and the richness of the data permits, estimating the parameters of the production and demand functions. We introduce an estimation technique that jointly estimates the two functions for each sector, bringing together insights from recent literature on estimating production functions based on output and input use data and proxy methods, and literature on estimating demand functions using P and Q data for outputs.[3] We do not impose constant returns to scale. In contrast to much of the literature estimating demand functions, we allow technical efficiency and quality/appeal to be correlated, even within establishments over time.

Quality-adjusted productivity accounts for about 114% of the cross sectional dispersion in size in levels and 122% in growth. About 93% of this variation is accounted for quality/appeal but 7% by technical efficiency. Composite wedges dampen dispersion in establishment size relative to their absence in an exact compensating amount (e.g., wedges dampen dispersion in levels by about 14% and 22% in growth). This dampening is even more extreme in the tails of the distribution. Plants in the bottom(top) productivity quartile are 42% larger (24% smaller) than efficient. Dispersion in input prices accounts for about half of this dampening component, more from wages than prices of material inputs. Markups play a negligible role on average–but explain a sizable wedge for the most highly productive plants–and the remaining

[3]For production function estimation using proxy methods, see, e.g. Ackerberg, Caves and Frazer (2015); De Loecker et al. (2016). For demand function estimation see, e.g., Hottman, Redding and Weinstein (2016); Foster, Haltiwanger and Syverson (2008).

residual wedge captures mostly revenue (rather than factor-biased) unobserved distortions.

Even though composite wedges dampen dispersion in size by a relatively modest 14%, the composite wedges imply large aggregate productivity losses of 37.6% with respect to efficiency. This contrast stems from the important role of wedges in the tails of the size distribution. Markups fit this narrative well, with dispersion in markups having a negligible impact on the size distribution but leading to a sizable 10.5% aggregate efficiency loss. This contrast is because markups affect precisely the highest productivity (and largest) plants. Similar contrasts apply to other components of wedges. Input price heterogeneity dampens the size distribution by 7% but alone implies a 32% efficiency loss, most of it driven by quality-adjusted wage heterogeneity. Residual sales wedges dampen the size distribution by 5% but alone imply 16% efficiency losses.[4] As we demonstrate, these wedges especially impact the tails of the distribution of size.

The dominant role of quality-adjusted productivity (rather than wedges) in accounting for the dispersion in size has implications for aggregate productivity. Efficient aggregate productivity (i.e., aggregate productivity in the absence of composite wedges) is 152% larger than it would be in the absence of dispersion in quality-adjusted productivity. The aggregate productivity "gain" from dispersion in quality-adjusted productivity is entirely driven by dispersion in quality/appeal, with a negligible role of dispersion in technical efficiency.

Our joint estimation procedure yields more pronounced concavity of the revenue function than implied by methods based solely on data for revenue and input use. These differences have important implications for the quantification of the role of determinants of the size distribution as well as aggregate allocative efficiency. For example, using traditional methods yields that dampening composite wedges account for -24% rather than -14% of the sales variability. Traditional methods also don't permit decomposition of quality-adjusted productivity into its quality/appeal and technical efficiency components nor the decomposition of composite wedges into input price dispersion, idiosyncratic markups, and residual wedges.

There are antecedents of some of our results in the literature. First, the dominant role of demand/quality/appeal in accounting for variation in plant heterogeneity has been featured in Foster, Haltiwanger and Syverson (2008), Foster, Haltiwanger and Syverson (2016), and Hottman, Redding and Weinstein (2016). However, our approach is based on joint estimation of the production and demand system in a rich environment with multiproduct producers that use multiple intermediate inputs, both of which are subject to product turnover. This allows us to contrast the role of demand to that of technical efficiency and show that the latter is non-negligible in explaining establishment performance. Second, the decomposition of composite wedges into multiple components is a novel feature of our analysis. Much of the literature has focused on either composite wedges (e.g. HK) or on specific individual components such as markups but without integrating them with other sources to assess their relative roles. For markups, moreover, we discuss below the critical question is how such markups are identified. Third, using a composite or residual marginal cost approach from the HRW framework masks a non-negligible positive contribution of technological improvements by lumping them together with negatively correlated residual wedges and input prices, which can only be uncovered by bringing together price/quantity data on both outputs and inputs.

[4]The efficiency costs from these three sources individually add to more than the overall efficiency loss due to the interactions between them.

In particular, the -6.5% contribution of the composite HRW "cost" residual to the variance of sales growth in our data reflects a positive contribution of 9.1% of cost factors (14.2% of technical efficiency and -5.1% from input prices), and an additional drag of -15.6% from residual wedges, which are not inherently a cost/supply side factor.

A novel finding relative to the literature is our identification of variation in quality-adjusted input prices as an important source of variation. Plant-specific input prices have been difficult to measure especially taking into account quality adjustment. The variation in quality-adjusted input prices that play an important dampening factor in size dispersion might reflect many factors, including the geographic segmentation of markets as well as institutional barriers or other frictions in the market. Such segmentation and frictions might be present in both intermediate input and labor markets. We don't identify these frictions but our analysis takes an important step forward by highlighting this variation. Most of the literature only identifies composite wedges indirectly through revenue productivity dispersion with any input price dispersion implicitly reflected in the measured revenue productivity dispersion.

Our analysis provides new insights into the ongoing debate about the role of markups as a drag on aggregate productivity. Instead of the indirect estimation approach of De Loecker, Eeckhout and Unger (2020), we jointly estimate the production and demand structure of the economy with heterogeneous markups that emerge from the assumed oligopolistic structure. In this respect, our structural approach to markups is similar to Hottman, Redding and Weinstein (2016) and Edmond, Midrigan and Xu (2018). By integrating this approach with our rich data and estimation, we provide guidance on the contribution of markups in composite wedges. Some recent papers investigating the drag on productivity from markups use all of the dispersion in revenue productivity to identify markups (e.g., Baqaee and Farhi (2020) and De Loecker, Eeckhout and Mongey (2021)).[5] Our findings do not support this strong assumption as we find a relatively minor role for markups in accounting for revenue productivity dispersion (i.e., equivalent to composite wedges in our framework) and in turn the unweighted size distribution of activity. However, we find a substantial role for markups in accounting for the drag on aggregate productivity from wedges given that we find that markups are the highest for the plants with the highest quality-adjusted productivity (and in turn largest size).

Our application is to an economy where wedges/distortions arguably play a larger role than in the US. Our results on the allocative efficiency effects of composite wedges for the Colombian manufacturing sector are in the broad range found by the literature that applies the HK method to developing countries, including those in Latin America. We thus see as likely that the relative role we find for Colombia for the different components of composite wedges (input price variability, markups, and residual wedges) applies more widely to similar countries. At the same time, we also find quantitatively similar results for the relative role of cost vs. demand and markup components to those found with data for the US by HRW, which is an indication that our results on the decomposition of the relative role of demand vs. efficiency and cost factors shed light on that role for a variety of environments.[6]

[5]De Loecker, Eeckhout and Mongey (2021) use an oligopolistic structural model but target the sales-weighted change in markups from the indirect De Loecker, Eeckhout and Unger (2020) that uses the dispersion in cost shares of revenue of variable factors to identify markups.

[6]Panel A of Table X of HRW shows that demand (combining appeal/scope) accounts for 107% of firm

The paper proceeds as follows. Section 2 presents our framework. We then explain the data used in our empirical work and the approach we use to measure attributes, including the joint estimation of the parameters of production and demand, respectively in sections 3 and 4. Our results on the drivers of size and growth dispersion are presented in section 5, while 6 presents implications for aggregate efficiency. Section 7 examines the value added of our joint estimation approach, by contrasting our results with those obtained with our framework and alternative estimation methods. Section 8 concludes by providing a more comprehensive view on the implications of our analysis, and on open questions for future research.

2 Theoretical framework

We build a model of plant optimal behavior given plant attributes in the context of multiproduct plants and a nested CES demand. The model nests the HK and HRW frameworks to derive the theoretical relationship between size and underlying attributes for multiproduct plants that exhibit appeal/quality differences within and between plants. The attributes we measure are: 1) the efficiency of the establishment's productive process (which we term $TFPQ$ as in Foster, Haltiwanger and Syverson (2008), though we generalize the concept to producers of heterogeneous goods); 2) appeal/quality;[7] 3) unit prices for inputs, in particular, material inputs and labor; 4) markups. The conceptual framework below defines each of these components. We also permit wedges between the theoretical prediction of a plant's size given its observed attributes and its size observed in the data.[8] (We use the words "plant" and "establishment" interchangeably.)

Measuring $TFPQ$ and appeal/quality in the context of multiproduct producers requires defining output at the level of the plant. In this multiproduct-establishment context, it is not possible to define real output without assumptions about demand. The concept of real output "in theory equals nominal output divided by a price index, but the choice of price index is not arbitrary: it is determined by the utility function" (Hottman, Redding and Weinstein, 2016, page 1349). We thus present our framework starting with the demand side and its implications for price and output measurement. We then move to the plant's problem. Next, we show how our framework nests those by HK and HRW. Finally, we examine how aggregate productivity is impacted by the different plant attributes we measure, including wedges between theory and data.

sales growth in their data, compared to our finding of 107% in the Colombian data (averaging across the life cycle). Combined cost factors are a drag of -7% in HRW's application, while if we combine the contributions of efficiency, input prices, and residual wedges that we find for Colombia, we account for about -7% as well.

[7] Hsieh and Klenow (2009, 2014) use the term $TFPQ$, to refer to a composite productivity measure that lumps together technical efficiency and demand shocks. We refer to this composite concept further below as $TFPQ\_HK$, as a reference to Hsieh and Klenow (2009, 2014). Haltiwanger, Kulick and Syverson (2018) explore properties of $TFPQ\_HK$ using U.S. data.

[8] Compared to HK's distortions, these wedges are narrower because we observe variation in input prices and markups not present in HK's data. As we show below, HK's distortions are a composite of input prices, markups, and our residual wedges.

2.1 Demand

Establishments produce multiple products and face demand for each of the products that depends on their quality. Taking into account multiproduct producers is important in our context, where two-thirds of observations correspond to multiproduct producers. The theoretical structure is such that we can measure the output of a multiproduct producer as revenue deflated with an appropriate establishment-level price index. As long as different products within an establishment are not perfect substitutes, that price index reflects product turnover and changing product appeal across existing products. To take this theory to the data we use the CUPI approach developed by Redding and Weinstein (2020) and also build on insights of Hottman, Redding and Weinstein (2016).

Consumers derive utility from a composite CES utility function, with a CES layer for establishments, indexed by f, and another for products (j) within establishments. Consumer's utility in this general CES structure in period t is given by:

$$U\left(Q_{1t},...,Q_{Nt}\right)=Q_t = \left(\sum_{I_t} d_{ft}Q_{ft}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}} \quad (1)$$

$$\text{where } Q_{ft} = \left(\sum_{\Omega_t^f} d_{fjt}q_{fjt}^{\frac{\sigma_w-1}{\sigma_w}}\right)^{\frac{\sigma_w}{\sigma_w-1}} \quad (2)$$

where p_{fjt} is the price of q_{fjt}, I_t is the set of establishments in period t, and $\sum_{I_t}\sum_{\Omega_t^f} p_{fjt}q_{fjt}$ $= E_t$.

An establishment f's real output, Q_{ft}, is a CES composite of individual products $Q_{ft} = \left(\sum_{\Omega_t^f} d_{fjt}q_{fjt}^{\frac{\sigma_w-1}{\sigma_w}}\right)^{\frac{\sigma_w}{\sigma_w-1}}$, where q_{fjt} is period t sales of good j produced by establishment f, the weights d_{fjt} reflect consumers' relative preference for different goods within the basket offered by establishment f, σ_w is the elasticity of substitution between products within f, and Ω_t^f is the basket of goods produced by f in year t. Products within establishments are not perfect substitutes so that tracking product turnover and changing product appeal within establishments is critical for measuring establishment-level output.

d_{fjt} and d_{ft} correspond to the weight, in consumer preferences, of product fj in establishment f's basket of products, and of establishment f in the set of establishments. We impose the following normalizations:

$$\prod_{\Omega_{t,t-1}^f} d_{fjt}^{\frac{1}{\left\|\Omega_{t,t-1}^f\right\|}}=1; \qquad \prod_{I_t} d_{it}^{\frac{1}{\|I_t\|}}=1. \quad (3)$$

where $\Omega_{t,t-1}^f$ is the set of products produced by f in both t and $t-1$. Given normalizations in equation (3), we refer to d_{fjt} and d_{ft} as, respectively, product (within establishment) and establishment appeal/quality or demand shocks. Product appeal d_{fjt} captures the valuation of attributes specific to good fj relative to other goods produced by establishment f. Establishment appeal d_{ft} captures attributes that are common to all goods provided by establishment

f, such as the establishment's customer service and the average quality of establishment f's products. Both establishment and product appeal may vary over time besides varying across establishments.[9] Consumer optimization implies that the period t demand for product fj and the establishment revenue are, respectively, given by

$$q_{fjt} = d_{ft}^{\sigma} d_{fjt}^{\sigma_w} \left(\frac{P_{ft}}{P_t}\right)^{-\sigma} \left(\frac{p_{fjt}}{P_{ft}}\right)^{-\sigma_w} \frac{E_t}{P_t} \tag{4}$$

$$R_{ft} = Q_{ft} P_{ft} = d_{ft}^{\sigma} P_{ft}^{1-\sigma} \frac{E_t}{P_t^{1-\sigma}} \tag{5}$$

where

$$P_t = \left(\sum_{I_t} d_{ft}^{\sigma} P_{ft}^{1-\sigma}\right)^{\frac{1}{(1-\sigma)}} \tag{6}$$

Dividing (5) by P_{ft} and solving for P_{ft}, we obtain

$$P_{ft} = D_{ft} Q_{ft}^{-\frac{1}{\sigma}} = D_t d_{ft} Q_{ft}^{-\frac{1}{\sigma}} \tag{7}$$

where $D_t = \left(\frac{E_t}{P_t^{1-\sigma}}\right)^{\frac{1}{\sigma}}$ and the establishment-level price index is given by:

$$P_{ft} = \left(\sum_{\Omega_t^f} d_{fjt}^{\sigma_w} p_{fjt}^{1-\sigma_w}\right)^{\frac{1}{(1-\sigma_w)}} \tag{8}$$

Given the nested CES demand, the establishment will charge the same markup on all products.[10]

Using equation 5, establishment appeal (d_{ft}) can be measured as sales holding prices constant: $d_{ft} = \frac{R_{ft}^{\frac{1}{\sigma}} P_{ft}^{\left(\frac{\sigma-1}{\sigma}\right)}}{D_t}$. This is akin to quality as defined by Hottman, Redding and Weinstein (2016); Khandelwal (2010); Hallak and Schott (2011); Fieler, Eslava and Xu (2018), and others. Foster, Haltiwanger and Syverson (2016) interpret establishment appeal as capturing the strength of the business' client base.

[9] We follow Redding and Weinstein (2020) in our treatment of product entry and exit. They don't formally model the decisions to add and subtract products but rationalize the entry and exit of products through assumptions on the patterns of product specific demand shocks. That is, they assume products enter when the product specific demand shock switches from zero to positive and exits when the reverse occurs. We rationalize product entry and exit in the same manner. We consider multi-product plants mostly for the purpose of obtaining a plant-level price deflator that takes into account changing multi-product activity.

[10] See Appendix S2 of Hottman, Redding and Weinstein (2016). In this nested environment the producer's optimization problem can be decomposed into two steps. The producer first chooses the composite index of products. It then chooses individual products to minimize the composite total cost subject to the optimal level of producer-level output. It is optimal for the producer to equate the ratio of marginal costs across products to the ratio of marginal utilities. Since consumer maximization yields that the ratio of marginal utilities across products is equal to the ratio of prices this implies the markups must be the same across products. One important difference with Hottman, Redding and Weinstein (2016) is that we don't permit product-specific random cost shocks.

Equation (8) defines establishment-level prices. Since (8) depends on unobservable σ_w and d_{fjt}, and thus cannot be measured readily from observables, we use Redding and Weinstein's (2020) CES Unified Price Index (CUPI) approach to express (the annual change in) this price index in terms of observables. Redding and Weinstein (2020) and Appendix A show that the CUPI provides the appropriate empirical analogue of our theoretical price index (8). The CUPI adjusts prices to take into account the evolution of the distribution of in-plant product appeal shifters d_{fjt}, emanating both from changes in appeal for continuing products and the entry/exit of products. This is crucial in our setting, since we define real output as deflated revenue, and thus our deflator needs to properly take into account changes in appeal from these sources.

In particular, the CUPI log change in $f's$ price index is given by:

$$\ln \frac{P_{ft}}{P_{ft-1}} = \sum_{\Omega^f_{t,t-1}} \ln \left(\frac{p_{fjt}}{p_{fjt-1}} \right)^{\frac{1}{\left\| \Omega^f_{t,t-1} \right\|}} + \frac{1}{\sigma_w - 1} \left(\ln \lambda^{QRW}_{ft} + \ln \lambda^{QF}_{ft} \right) \tag{9}$$

Defining as s_{fjt} the share of $f's$ period t revenue represented by product j ($s_{fjt} = \frac{p_{fjt} q_{fjt}}{R_{ft}}$), $\lambda^{QF}_{ft} = \frac{\sum_{\Omega^f_{t,t-1}} s_{fjt}}{\sum_{\Omega^f_{t,t-1}} s_{fjt-1}}$ is Feenstra's (1994) adjustment for within-plant appeal changes from the entry/exit of products, allowing us to take product entry and exit into account. Similarly, defining s^*_{fjt} as the share that product j represents in the revenue that f obtains in t from products that belong to the bundle $\Omega^f_{t,t-1}$, $\lambda^{QRW}_{ft} = \prod_{\Omega^f_{t,t-1}} \left(\frac{s^*_{fjt}}{s^*_{fjt-1, \Omega^f_{t,t-1}}} \right)^{\frac{1}{\left\| \Omega^f_{t,t-1} \right\|}}$ is Redding and Weinstein's (2020) adjustment for changes in relative appeal for continuing products within the plant, which deals with the consumer valuation bias that affects traditional approaches to the empirical implementation of theory-motivated price indices.[11] The derivation of the CUPI price index from our theoretical price index 8 (Appendix A) requires imposing the normalization that $\sum_{\Omega^f_{t,t-1}} \ln d_{fjt}^{\frac{1}{\left\| \Omega^f_{t,t-1} \right\|}} = 0$. That is, the CUPI adjusts for relative appeal changes within the plant, while average appeal changes for the plant are captured by d_{ft}.

Building recursively from a base year B and denoting $\overline{P^*_{ft}} = \prod_{l=B+1}^{t} \left[\prod_{\Omega_{t,t-1}} \left(\frac{p_{fj}}{p_{fjt-1}} \right)^{\frac{1}{\left\| \Omega_{t,t-1} \right\|}} \right]$, $\Lambda^{QRW}_{ft} = \prod_{l=B+1}^{t} \left[\left(\lambda^{QRW}_{fl} \right) \right]$ and $\Lambda^{QF}_{ft} = \prod_{l=B+1}^{t} \left[\left(\lambda^{QF}_{fl} \right) \right]$, we obtain the empirical price index in levels:

[11] Sato (1976) and Vartia (1976) show how the theoretical price index can be implemented empirically under the assumption of invariant firm appeal shocks and constant baskets of goods. Feenstra (1994) derives an empirical adjustment of the Sato-Vartia approach that takes into account changing baskets of goods, keeping the assumption of a constant firm appeal distribution for continuing products. It is this last assumption that the UPI relaxes.

$$
\begin{aligned}
P_{ft} &= P_{fB} * \overline{P^*_{ft}} * \left(\Lambda^{QRW}_{ft}\Lambda^{QF}_{ft}\right)^{\frac{1}{\sigma_w - 1}} \\
&= P_{fB} * \overline{P^*_{ft}} * \left(\Lambda^{Q}_{ft}\right)^{\frac{1}{\sigma_w - 1}}
\end{aligned} \tag{10}
$$

where P_{fB} is the plant-specific price index at the plant's base year B (see Appendix A).

2.2 Plant Optimization

We now specify the establishment-level problem. We specify a framework such that establishments that produce multiple products matter, and this occurs in three dimensions. First, we specify our cost/production structure directly at the establishment-level, rather than setting up product specific cost/production functions as in Hottman, Redding and Weinstein (2016). We make this assumption for more than the convenience that our data on input use are at the establishment level. Our view is that, if one queried establishments (plants) to specify input costs (capital, labor, materials, and energy) on a product specific basis, most would be unable to do so since multiple costs are shared across products (i.e., there is joint production). That is, an establishment is not simply a collection of separable lines of production. It is, in itself, an empirically relevant object. A second sense in which establishments matter in our framework is that we depart from monopolistic competition: some establishments may be large enough in the market that they don't take the sectoral output price as given. Third, there may be cannibalization between products of the same establishment.[12]

In the model, the establishment chooses its size optimally given technical efficiency, quality, input prices, and residual wedges. In the spirit of an accounting exercise, the framework remains silent about the sources of these attributes and rather asks how the establishment adjusts its size given those attributes at time t, and contingent on survival to that time. Further below we discuss our explorations of endogenous innovation and exit.[13]

Consider an establishment indexed by f, that produces output Q_{ft} using a composite input X_{ft} to maximize its profits, with technology

$$
Q_{ft} = A_{ft}X^{\gamma}_{ft} = a_{ft}A_t X^{\gamma}_{ft} \tag{11}
$$

[12]It is potentially of interest to also consider the firm which may consist of multiple establishments. One practical reason to focus on establishments is that, while aggregation across establishments of the same firm is possible in the manufacturing survey, ownership changes affect longitudinal linkages for firms, while establishments in the survey keep their identifiers over time independent of ownership changes. We also think that establishment is an appropriate level of aggregation for decision-making for a number of reasons. First, more than 90% of firms in Colombia are single-establishment firms. Second, even for multi-units, profit maximization of the firm would typically involve profit maximization at each establishment. There are some issues that also involve the firm such as financing and there may be interactions between establishments of the same firm worth considering. These are interesting issues we leave for future work.

[13]For instance, the seminal models of Hopenhayn (2016); Melitz (2003), and much of the work that has since followed in Macroeconomics and Trade. Endogenous productivity-quality growth has made its way to these models more recently (e.g. Atkeson and Burstein, 2010; Acemoglu et al., 2018; Hsieh and Klenow, 2014; Fieler, Eslava and Xu, 2018) . The firm's efforts to strengthen demand may include investments in building its client base (Foster, Haltiwanger and Syverson, 2016), and adding new products and/or improving the quality of its pre-existing product lines. Those to strengthen $TFPQ$ may include better management of the production process (e.g. Bloom and Reenen, 2007) or acquiring better machines.

A_{ft} is the establishment's technical efficiency, which we term $TFPQ$ following Foster, Haltiwanger and Syverson (2008). A_{ft} has an aggregate and an idiosyncratic component (A_t and a_{ft}). γ is the returns to scale (in production) parameter. Equation (11) defines a_{ft} as the (idiosyncratic) efficiency of the productive process: how much output the establishment obtains from a unit of a basket of inputs.

The establishment faces demand given by (7). Multiplying (7) by Q_{ft} we obtain:

$$R_{ft} = D_{ft} Q_{ft}^{1-\frac{1}{\sigma}} \tag{12}$$

The establishment chooses its scale X_{ft} to maximize profits

$$\underset{X_{it}}{Max} \ (1-\tau_{ft})\, R_{ft} - C_{ft} X_{ft} = (1-\tau_{ft})\, D_{ft} A_{ft}^{1-\frac{1}{\sigma}} X_{ft}^{\gamma\left(1-\frac{1}{\sigma}\right)} - C_{ft} X_{ft}$$

taking as given A_{ft}, D_{ft}, and unit costs of the composite input, denoted C_{ft}. For developing our theoretical predictions, we treat input prices as exogenous and potentially idiosyncratic for the composite input.

There may be idiosyncratic residual wedges τ_{ft}, that lead to a gap between an establishment's observed scale and that which would be implied by the static model given its measured attributes.[14] Such wedges capture, for instance, adjustment costs to changing the scale, changing the mix of inputs, or building up a customer base; product-specific tariffs; financing constraints; information frictions; and size-dependent regulations or taxes. Adjustment costs break the link between actual adjustment and the "desired adjustment"[15]. Financing constraints may similarly limit the ability of the establishment to undertake optimal investments, and force it to remain smaller than optimal and even potentially exit the market during liquidity crunches even if its present discounted value is positive.[16] In the HK framework wedges can also arise from idiosyncratic variability in input prices and markups (see subsection 2.3), but we explicitly account for these sources of heterogeneity, so that they are not lumped into τ_{ft} in our framework.

The resulting τ_{ft} may be correlated with plant attributes themselves. By their very nature, adjustment costs and financing constraints are typically correlated with plant attributes. Size-dependent regulations are another prominent example of correlated wedges.[17]

We impose Cournot competition to allow establishments to hold market power, so that an establishment's market share may be non-negligible.[18] This also implies that, in choosing its optimal scale, an establishment does not take as given the aggregate price index, P_t. Under these conditions and our CES demand structure, variability in markups across establishments stems from market power:

[14] As in Restuccia and Rogerson (2008); Hsieh and Klenow (2009). Further below, we also consider factor-specific distortions that, for a given choice of X_{it}, affect the relative choice of a given input with respect to others.

[15] See, for instance, Caballero, Engel and Haltiwanger (1995, 1997); Eslava et al. (2010); Asker, Collard-Wexler and De Loecker (2014)

[16] Gopinath et al. (2017); Fieler, Eslava and Xu (2018)

[17] E.g. García-Santana and Pijoan-Mas (2014); Garicano, Lelarge and Van Reenen (2016).

[18] Alternatively assuming Bertrand competition has little effect on our model and our results. Under both Cournot and Bertrand competition, the markup is a function of the plant's market share (see equation 13). The functional form for the markup is the only change between the two cases in the theory

$$\mu_{ft} = \frac{\sigma}{(\sigma-1)}\frac{1}{(1-s_{ft})} \tag{13}$$

where μ_{ft} is the establishment's markup and $s_{ft} = \frac{R_{ft}}{E_t}$ (proof: Appendix D). As in Hsieh and Klenow (2009, 2014), marginal cost is defined inclusive of residual wedges, so that

$$P_{ft} = \mu_{ft} * \frac{\partial CT_{ft}}{\partial Q_{ft}}(1-\tau)^{-1} \tag{14}$$

where CT is total cost.

In our application, the demand function and production function parameters are constant across establishments within sectors (at the three digit level of the ISIC revision 2 classification for Colombia, of which there are 23 manufacturing sectors). All technological differences across plants within sectors are thus lumped into A_{ft} dispersion. An establishment's relevant market, for the purpose of calculating its market share and markup, is defined as the group of producers of the plant's most important CPC 3-digit product, of which there are 112 such groups, so that s_{ft} is f's revenue share in its CPC 3-digit group.

Profit maximization yields optimal input demand $X_{ft} = \left(\frac{D_{ft}A_{ft}^{1-\frac{1}{\sigma}}\gamma}{C_{ft}\mu_{ft}\left(1-\tau_{ft}\right)^{-1}}\right)^{\frac{1}{1-\gamma\left(1-\frac{1}{\sigma}\right)}}$. Plugging into 11 and then into 12, we obtain optimal sales and life-cycle growth of sales as functions of measured attributes (D_{ft}, A_{ft}, μ_{ft}, and C_{ft}), wedges τ_{ft}, and parameters:

$$\begin{aligned} R_{ft} &= d_{ft}^{\kappa_1}a_{ft}^{\kappa_2}pm_{ft}^{-\phi\kappa_2}w_{ft}^{-\beta\kappa_2}\mu_{ft}^{-\gamma\kappa_2}\left(\widehat{\chi}_t\chi_{ft}\right)^{1-\frac{1}{\sigma}} \\ \frac{R_{ft}}{R_{f0}} &= \left(\frac{d_{ft}}{d_{f0}}\right)^{\kappa_1}\left(\frac{a_{ft}}{a_{f0}}\right)^{\kappa_2}\left(\frac{pm_{ft}}{pm_{f0}}\right)^{-\phi\kappa_2}\left(\frac{w_{ft}}{w_{f0}}\right)^{-\beta\kappa_2}\left(\frac{\mu_{ft}}{\mu_{f0}}\right)^{-\gamma\kappa_2}\left(\frac{\widehat{\chi}_t}{\chi_0}\frac{\chi_{ft}}{\chi_{f0}}\right)^{1-\frac{1}{\sigma}} \end{aligned} \tag{15}$$

where $\kappa_1 = \frac{1}{1-\gamma\left(1-\frac{1}{\sigma}\right)}$, $\kappa_2 = \left(1-\frac{1}{\sigma}\right)\kappa_1$, and we have further assumed $X_{ft} = K_{ft}^{\frac{\beta}{\gamma}}L_{ft}^{\frac{\alpha}{\gamma}}M_{ft}^{\frac{\phi}{\gamma}}$, so that C_{ft} is the corresponding Cobb-Douglas aggregate of different input prices, inclusive of any relative distortions across the prices of the different inputs (factor-biased distortions). d_{ft}, a_{ft}, pm_{ft} and w_{ft} are, respectively, the idiosyncratic components of D_{ft}, A_{ft}, Pm_{ft} and W_{ft}.The second line is obtained by dividing each optimal outcome in period t by its optimal level at birth ($t = 0$)(see Appendix B).[19] Aggregate components, D_t, A_t and C_t, as well as other factors that affect all plants equally, are lumped into χ_t and $\widehat{\chi}_t$. The focus of our empirical analysis is on idiosyncratic attributes and behavior, so these aggregate components are later differenced out.

Among input prices, two are observed in the data: the price of material inputs, Pm_{ft}, and average wage per worker, W_{ft}. Empirically we consider multiple material inputs and make efforts to take into account material input heterogeneity through quality-adjusting prices. In particular, we use a plant-level price index for materials, pm_{ft}, using information on prices and quantities of material inputs at the detailed product class level. We construct pm_{ft} using an analogous approach to that used to construct output prices, which takes into

[19]There is some slight abuse of notation here as t is used for calendar time and, in this expression, to refer to age when we express the ratio of these variables at *age t* to *age at birth* ($t = 0$) to express growth over the life cycle.

account quality heterogeneity across different material inputs used by the plant, changes in the mix of inputs, and changes over time in the relative quality of existing inputs. We also measure material inputs deflating material expenditure by pm_{ft}. We similarly attempt to quality-adjust wages using information on different types of workers.

Crucially, $\chi_{ft} = (1-\tau_{ft})^{\gamma\kappa_1} * \left(r_{ft}\chi^{\mathrm{K}}{}_{ft}\right)^{-\alpha\kappa_1}$ includes revenue distortions τ_{ft} and the unobserved idiosyncratic user cost of capital, inclusive of distortions with respect to the prices of other inputs (factor-biased distortions). χ_{ft} thus captures idiosyncratic wedges from different sources, including factor-biased and factor-unbiased. We refer to $\chi_{ft}^{1-\frac{1}{\sigma}} = (1-\tau_{ft})^{\gamma\kappa_1\left(1-\frac{1}{\sigma}\right)}\left(r_{ft}\chi^{\mathrm{K}}{}_{ft}\right)^{-\alpha\kappa_1\left(1-\frac{1}{\sigma}\right)}$ as a "residual wedge".

Equation system (15) is the focus of our analysis of the distribution of establishment revenue and establishment revenue growth. We start with the level and growth of (idiosyncratic) attributes that we can measure: quality/appeal or "demand shocks" (d_{ft}), measured as a residual from equation 5; technical efficiency or $TFPQ$ (a_{ft}) measured as a production function residual; markups (μ_{ft}) measured using equation 13; and wages and material input prices (w_{ft}, pm_{ft}). The residual wedges χ_{ft} that an establishment faces are measured as an age-specific residual. Although we do not directly measure the (adjusted) user cost of capital to be able to dissect χ_{ft} into its revenue and factor-biased components, in the Appendix I we implement an indirect approach to this further decomposition.

We estimate χ_{ft} taking observed attributes as given. However, we do explore the empirical cross-sectional relationship between those attributes and wedges. And, although we assume the plant takes its attributes as given in choosing its size, we also explore the relationship between proxies for investment in innovation, D_{ft} and A_{ft}, to shed light on the determinants of the latter. This is done in Appendix E. Finally, we focus on decomposing the determinants of size and growth of surviving establishments up to any given age, but include robustness analysis separating survivors from exits in Appendix H. We conclude that our findings for plants that survive up to age t are largely driven by the establishments that survive at least one more year, despite exiting plants exhibiting much worst quality-adjusted productivity and more marked negatively correlated wedges from input prices.

Notice also that, although we don't explicitly model dynamic frictions, we take the shortcut in recent literature on misallocation to permit wedges or distortions between frictionless static first-order conditions and actual behavior (e.g. Hsieh and Klenow, 2009). Such distortions and wedges might capture factors that induce dynamic behavior such as adjustment costs, information frictions, and distortions arising from the business climate.[20] This shortcut enables us to use a simple static model of optimal input determination to frame our analysis of size and growth between birth and any given age. We permit the wedges or distortions to vary by establishment age.

[20]This shortcut has limitations as the idiosyncratic distortions that we permit don't provide the discipline that formally modeling dynamic frictions imply. See, e.g., Asker, Collard-Wexler and De Loecker (2014); Decker et al. (2020); David and Venkateswaran (2019). But it has the advantage in subsuming in a simple measure different types of frictions and distortions, including those that capture dynamic considerations.

2.3 Decomposing HK's $TFPR_{ft}$ and HRW's marginal cost

Our framework nests HK's on the supply side and HRW's on the demand side. We now explain how the HK and HRW approaches are nested into ours, and how our integration of the two frameworks allows us to further decompose the residual heterogeneity, shedding light on the factors that determine those residuals and on their empirical role.

In absence of data on input and output prices, HK decompose revenue into a measure of quality-adjusted productivity that combines our $TFPQ_{ft}$ and D_{ft} shocks, which we label as $TFPQ\_HK_{ft}$, and a residual wedge that captures all determinants of size other than technical efficiency and quality.[21] Define a plant's quality-adjusted productivity as $TFPQ\_HK_{ft} = A_{ft}D_{ft}^{\frac{1}{1-\frac{1}{\sigma}}}$. Starting from revenue in equation 12, $TFPQ\_HK_{ft}$ can be measured using solely revenue and input data, as long as estimates of γ and σ are available:

$$TFPQ\_HK_{ft} = R_{ft}^{1/(1-\frac{1}{\sigma})}/X_{ft}^{\gamma} = A_{ft}D_{ft}^{\frac{1}{1-\frac{1}{\sigma}}} \tag{16}$$

A widely used implication of HK's framework is that wedges can be estimated from the idiosyncratic component of $TFPR_{ft} = \frac{R_{ft}}{X_{ft}}$. Replacing into (12) optimal input demand $X_{ft} = \left(\frac{D_{ft}A_{ft}^{1-\frac{1}{\sigma}}\gamma}{C_{ft}\mu_{ft}\left(1-\tau_{ft}\right)^{-1}}\right)^{\frac{1}{1-\gamma\left(1-\frac{1}{\sigma}\right)}}$ we obtain

$$TFPR_{ft} = \frac{C_{ft}\mu_{ft}}{\gamma\left(1-\tau_{ft}\right)} \tag{17}$$

$TFPR_{ft}$ variability reflects variation not only in τ_{ft}, but also in markups and input prices.[22] Notice also that, plugging X_{ft} into 12 and using 16 and 17 to label composite terms, revenue can then be expressed as:

$$R_{ft} = \left[\frac{TFPQ\_HK_{ft}}{\left(\gamma * TFPR_{ft}\right)^{\gamma}}\right]^{\frac{1-\frac{1}{\sigma}}{1-\gamma\left(1-\frac{1}{\sigma}\right)}} \tag{18}$$

The HK wedge $TFPR_{ft}$ is a *composite* measure of wedges, just as quality-adjusted productivity $TFPQ\_HK$ is a *composite* measure of efficiency and demand/quality/appeal. A crucial insight from HK is that $TFPR_{ft}$ heterogeneity induces large distortions in the size distribution of plants and associated large aggregate efficiency losses. Our ability to measure μ_{ft} and part of C_{ft} (specifically pm_{ft} and w_{ft} allows us to decompose the composite $TFPR_{ft}$ into the contribution of those measurable components and a remaining residual sales wedge,

[21] See the appendix to Hsieh and Klenow (2009) where they extend their model to account for D shocks. What we label $TFPQ\_HK$ is what is called $TFPQ$ by HK, but in their extended framework comprises quality/appeal besides technical efficiency. Haltiwanger, Kulick and Syverson (2018) also explore properties of $TFPQ\_HK$ constructed from revenue and input data compared to $TFPQ$ and demand shocks constructed from price and quantity data.

[22] If, as originally defined in Foster, Haltiwanger and Syverson (2008), we rather defined $TFPR$ as $\frac{R_{ft}}{X_{ft}^{\gamma}}$, $TFPR$ dispersion would also reflect A_{ft} and D_{ft} dispersion. Their definition of $TFPR_{ft} = P_{ft}A_{ft}$ applies to the $\gamma = 1$ case. The relevant definition for a measure of wedges/distortions is equation 17.

thus shedding light on the role played, in the losses identified by the HK framework, by idiosyncratic markups and distortions in the labor market and other input markets.

In turn, the differential contribution of demand vs. cost-side shocks to plant sales is explored by Hottman, Redding and Weinstein (2016). Using the same CES demand structure on which we rely, they decompose idiosyncratic sales as captured by equation (5) into the contributions of observed plant prices and demand shocks obtained using the estimated elasticity of substitution. They subsequently use 14 to decompose price into the contributions of markups–computed as in equation (13)–and residual marginal costs. These residual marginal costs, given by $\frac{\partial CT_{ft}}{\partial Q_{ft}}(1-\tau_{ft})^{-1}$, thus capture idiosyncratic variation in costs from both input price variability and technical efficiency, as well as residual wedges. However, these wedges are not inherently driven by cost/supply side factors. For example, they could reflect the adjustment costs associated with building up a customer base. Our ability to measure input prices and technical efficiency ($TFPQ_{ft}$) allows us to decompose their residual marginal cost measure into these different sources to assess the role played by wedges vs. factors that truly lie on the cost side, and to decompose the latter into its components related to observed input prices and efficiency in production. HRW find a negligible role for their residual marginal cost term; our approach allows us to uncover non-trivial roles for efficiency in production and input prices in determining the size distribution of plants, offset by residual wedges that mask the true importance of cost factor if not dissected. See Appendix J for greater details on the relationship between $HRW's$ framework and ours.

2.4 Aggregate Productivity and Aggregate Efficiency

We now derive an expression for aggregate productivity, TFP_t, and show how TFP_t and efficiency relate to cross sectional variability in the components of $TFPQ\_HK_{ft}$ and $TFPR_{ft}$ (equations 16 and 17).

Aggregate TFP_t is given by $\frac{Q_t}{X_t} = \frac{E_t}{X_t}\frac{1}{P_t} = \frac{E_t}{X_t}\left(\sum_{I_t} d_{ft}^{\sigma} P_{ft}^{1-\sigma}\right)^{\frac{1}{\sigma-1}}$. As shown in Appendix K, replacing P_{ft} by its equilibrium value, TFP_t can be written as

$$TFP_t = D_t^{-\frac{\sigma}{\sigma-1}}\left(\sum_{I_t}\left(\left(\frac{TFPQ\_HK_{ft}}{TFPR_{ft}^{\gamma}}\right)^{\frac{1}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}\overline{TFPR_t}\right)^{\sigma-1}\right)^{\frac{1}{\sigma-1}} \tag{19}$$

where we have defined $\overline{TFPR_t} = \frac{P_t Q_t}{X_t} = \sum\left(\frac{P_{ft}Q_{ft}}{X_{ft}}\frac{X_{ft}}{X_t}\right)$ and have used the fact that $TFPQ\_HK_{ft} = D_{ft}^{\frac{\sigma}{(\sigma-1)}} A_{ft}$.

Also notice that, normalizing all attributes of the plant around a sector*year mean (in logs), we can write $TFPR_{ft} = \frac{C_{ft}\mu_{ft}}{\gamma(1-\tau_{ft})} = \frac{\overline{C_t}\overline{\mu_t}}{\gamma(\overline{1-\tau_t})} tfpr_{ft}$, where we use upper bars for aggregate components and $tfpr_{ft}$ contains only the idiosyncratic components of C_{ft}, μ_{ft} and $(1-\tau_{ft})$.

We define efficiency as a situation where, using lower bars to denote idiosyncratic components, $\underline{c_{ft}} = 1$, $\underline{\mu}_{ft} = 1$ and $\underline{1-\tau_{ft}} = 1$, and denote TFP_t evaluated at this situation as TFP_t^{eff}. Further defining $AE_t = \frac{TFP_t}{TFP_t^{eff}}$, where AE stands for "aggregate efficiency" , and using 19, we obtain, after some manipulation

$$AE_t = \left(\frac{1}{N_t} \sum_{I_t} \left[\left(\frac{\Delta_{ft}^{\frac{1}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\widetilde{\Delta}_t} \right) \left(\frac{tfpr_{ft}^{\frac{\gamma}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\overline{tfpr_t}} \right)^{-1} \right]^{\sigma-1} \right)^{\frac{1}{\sigma-1}} \tag{20}$$

where $\Delta_{ft} = d_{ft}^{\frac{\sigma}{(\sigma-1)}} a_{ft} = \frac{TFPQ\_HK_{ft}}{D_t^{\frac{\sigma}{\sigma-1}} A_t}$; $\widetilde{\Delta}_t = \left(\frac{1}{N_t} \sum_{I_t} \Delta_{ft}^{\frac{\sigma-1}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}} \right)^{\frac{1}{\sigma-1}}$; and $\overline{tfpr_t} = \left(\sum_{I_t} tfpr_{ft} \frac{X_{ft}}{X_t} \right)$.

An implication of 20 is that aggregate efficiency depends on the covariance between (a function of the idiosyncratic components of) $TFPQ\_HK_{ft}$ and $TFPR_{ft}$, and not only on the dispersion of $TFPR_{ft}$. In particular, using $E \left(\frac{\Delta_{ft}^{\frac{1}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\widetilde{\Delta}_t} \right)^{\sigma-1} = 1$, equation 20 can be written as:[23]

$$AE_t = \left[cov \left(\left(\frac{\Delta_{ft}^{\frac{1}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\widetilde{\Delta}_t} \right)^{\sigma-1}, \left(\frac{tfpr_{ft}^{\frac{\gamma}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\overline{tfpr_t}} \right)^{1-\sigma} \right) + E \left(\frac{tfpr_{ft}^{\frac{\gamma}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\overline{tfpr_t}} \right)^{1-\sigma} \right]^{\frac{1}{\sigma-1}} \tag{21}$$

Since $1-\sigma < 0$, (composite) distortions that are negatively correlated with $TFPQ\_HK_{ft}$ and more disperse have negative effects on welfare.

[23] This decomposition is similar to that in Blackwood et al. (2021), who analyzed the sensitivity of measured allocative efficiency to parameters that can be estimated from traditional revenue and input expenditure data. Blackwood et al. (2021) also analyze the general case in which allocative efficiency depends on the covariance between functions of $TFPQ\_HK_{ft}$ and $TFPR_{ft}$. The special case of $\gamma = 1$ reproduces HK's result that, under a joint lognormal distribution for $TFPR_{ft}$ and $TFPQ\_HK_{ft}$, and if factor-specific distortions are absent, aggregate TFP depends on the dispersion of $TFPR$:

$$\begin{aligned} \log TFP_t &= \frac{1}{\sigma-1} \log \sum_{I_t} TFPQ\_HK_{ft}^{\sigma-1} - \frac{\sigma}{2} var(\log TFPR_{ft}) \\ &= \frac{1}{\sigma-1} \log(\|I_t\|) + E\left(\log TFPQ\_HK_{ft}\right) + \frac{(\sigma-1)}{2} var\left(\log TFPQ\_HK_{ft}\right) - \frac{\sigma}{2} var(\log TFPR_{ft}) \end{aligned}$$

3 Data

3.1 Annual Manufacturing Survey

We use data from the Colombian Annual Manufacturing Survey (AMS) from 1982 to 2012 (DANE, 1982-2013). The survey, collected by the Colombian official statistical bureau DANE, covers all manufacturing establishments (=plants) belonging to firms that own at least one plant with 10 or more employees, or those with annual production value exceeding a level close to US$100,000. Our sample contains 19,326 plants over the whole period, with 5,434 plants in the average year. Over 90% of plants in the AMS (i.e. over 90% manufacturing plants in Colombia with size over the inclusion threshold) belong to single-plant firms, so that the distinction between plants and firms is not as crucial in our context as it is in others.

Surveyed establishments are asked to report their level of production and sales, as well as their use of employment and other inputs, their purchases of fixed assets, and the value of their payroll. We construct a measure of plant-level wage per worker by dividing payroll into number of employees and obtain the capital stock using perpetual inventory methods, initializing at book value of the year the plant enters the survey. Sector IDs are also reported, at the 3-digit level of the ISIC revision 2 classification.[24] Appendix Table C2 lists the 23 sector classes in our data.

A unique feature of the AMS, crucial for our ability to measure a variety of plant attributes, is that inputs and products are reported at a detailed level. Plants report separately each material input used and product produced, at a level of disaggregation corresponding to seven digits of the ISIC classification (close to six digits in the Harmonized System, a level such that refined soy oil is distinguished from unrefined soy oil and from refined sunflower oil). For each of these detailed inputs and products, plants report separately quantities and values used or produced, so that plant-specific unit prices can be computed for both individual inputs and individual outputs. The average (median) plant produces 3.77 (3) products per year and employs 12.05 (10) inputs per year (Table 2).

By taking advantage of product-plant-specific prices, we produce plant-level price indices for both inputs and outputs, and as a result, generate measures of productivity based on output, estimate demand shocks, and consider the role of input prices in plant size. Details on how we go about these estimations are provided in section 4. Our product level data are not at the detailed UPC code level used by Hottman, Redding and Weinstein (2016), which implies the limitations discussed in the introduction, but we observe them at the plant-by-product-by-year level, which offers key advantages relative to other data sources. Unlike UPC codes, our product-level information is available by plant (physical location of production) rather than the aggregate firm, and is jointly observed with input use by that plant. And, unlike transactions data for imports (used, for instance by Feenstra, 1994; Broda and Weinstein, 2006), we observe them not only at the product level (at similar levels of disaggregations with respect to imports transactions data) but by producer at a physical location. Compared to analysis based on UPC codes, the higher aggregation of our data implies that quality adjustments are likely captured as within product d_{fjt} changes rather

[24]The ISIC classification in the survey changed from revision 2 to revision 3 over our period of observation. The three-digit level of disaggregation of revision 2 is the level at which a reliable correspondence between the two classifications exists.

than through the introduction of a new product code.

Each establishment is assigned a unique ID that allows us to follow it over time. Since a plant's ID does not depend on an ID for the firm that owns the plant, it is not modified with changes in ownership, and such changes are not mistakenly identified as plant births and deaths.[25] There is exit in our sample, at a rate of approximately 7% per year. Our analysis includes both continuers and exiters, which we examine separately in Appendix H.

The plant's initial year of operation is recorded–again, unaffected by changes in ownership. We use that information to calculate an establishment's age in each year of our sample. Though we can only follow establishments from the time of entry into the survey, we can determine their correct age, and follow a subsample from birth. About a third of plants in our sample are observed from birth. Based on that restricted subsample, we generate measurement adjustment factors that we then use to estimate life-cycle growth even for plants that we do not observe from birth.[26] Our decomposition results are in general robust to using the subsample observed from birth rather than the full sample, although estimated with less precision and for a shorter lifespan.

4 Estimating $TFPQ$ and quality/appeal

Measuring $TFPQ$ and quality/appeal requires estimating the production and demand functions, (11) and (7). Once the coefficients of these functions have been estimated, $TFPQ$ is the residual from (11) and quality/appeal is the residual from (7).

We implement a joint estimation procedure of (7) and (11). Jointly estimating the two equations allows us to take full advantage of the information to which we have access to separate supply from demand in the data. As a result, we can estimate production rather than revenue elasticities, even for multiproduct plants, and simultaneously obtain unbiased estimates of σ and σ_w. We impose a set of moment conditions that require less structure overall, and weaker restrictions on the covariance between $TFPQ$ and demand shocks, than other usual estimation methods of the demand-supply system in multiproduct settings. This is in part possible thanks to the fact that we have access to price and quantity information for both inputs and outputs. Data on inputs inform the estimation directly about the production side, thus allowing us to separate it from demand under weaker restrictions than if we only used information on prices and quantities for outputs (as in, for instance, Broda and Weinstein, 2006; Hottman, Redding and Weinstein, 2016). Data on prices allows us to properly estimate both production and revenue elasticities.

Beyond the usual simultaneity biases and restrictions on supply vs demand, the estimation of (11) and (7) faces the problem that, until we have an estimate of σ_w, we are unable to

[25] Plant IDs in the survey were modified in 1992 and 1993. To follow establishments over that period, we use the official correspondence that maps one into the other. The correspondence seems to be imperfect (as suggested by the apparent high exit in 92 and high entry in 93), but even for actual continuers that are incorrectly classified as entries or exits, our age variable is correct (see further below).

[26] If B is the age of plant f when we first observe it in the survey, then for variable Z we estimate size at age a relative to birth as $Z_{fa}/Z_{f0} = (Z_{f,a}/Z_{f,B})(Z_B/Z_0)_{restricted}$ where $(Z_B/Z_0)_{restricted}$ is average growth from birth to age B in the restricted subsample observed from birth. This helps us deal with selection bias without losing valuable information from plants first observed years after birth. Unadjusted average growth is generally biased downwards, since $(Z_{f,a}/Z_{f,B}) = 1$ at age a=B, while generally in the sample $(Z_{f,a}/Z_{f,B}) > 1$.

properly construct P_{ft}, and thus $Q_{ft} = \frac{R_{ft}}{P_{ft}}$ (see section 2.1). We therefore rely on P_{ft}'s two separate components from equation 10: $\overline{P^*_{ft}}$ and Λ^Q_{ft}.[27] We define

$$Q^*_{ft} = \frac{R_{ft}}{P_{fB}\overline{P^*_{ft}}} = Q_{ft} * \left(\Lambda^Q_{ft}\right)^{\frac{1}{\sigma_w - 1}} \tag{22}$$

and proceed in three steps to address this limitation:

1. (This step is only sketched here, details are provided in the following subsection) Jointly estimate the coefficients of the production function (11), the demand function (7), and σ_w using $Q^*_{ft} = \frac{R_{ft}}{P_{fB}\overline{P^*_{ft}}} = Q_{ft} * \left(\Lambda^Q_{ft}\right)^{\frac{1}{\sigma_w - 1}}$ and $\overline{P^*_{ft}} = \frac{P_{ft}\left(\Lambda^Q_{ft}\right)^{\frac{-1}{\sigma_w - 1}}}{P_{fB}}$ as the respective dependent variables of these two functions. We carry Λ^Q_{ft} as a separate regressor in each equation to deal with potential biases induced by the–at this point–still partial estimation of revenue deflators. In particular, not explicitly accounting for changes in product quality and variety within the plant leads to "quality" and "variety" biases in the estimation of production function coefficients, as described by de Roux et al. (2021). Λ^Q_{ft} explicitly accounts for those changes, freeing our estimates from such biases. We similarly introduce separately M^*_{ft} and Λ^M_{ft} in the production function (where $M^*_{ft} = \frac{\text{materials expenditure}}{PM_{fB}\overline{PM^*_{ft}}}$, and Λ^M_{ft} is the adjustment factor for the prices of materials analogous to Λ^Q_{ft} see Appendix A). The joint estimation is conducted separately for each three-digit sector. The parameters $\{\alpha,\ \beta,\ \phi,\ \sigma,$ and $\sigma_w\}$ used in the analysis are those estimated in this step.

2. Use the estimated elasticity $\widehat{\sigma_w}$ for the respective three-digit sector to obtain $P_{ft} = P_{fB} * \overline{P^*_{ft}} * \left(\Lambda^Q_{ft}\right)^{\frac{1}{\widehat{\sigma_w} - 1}}$ and subsequently $Q_{ft} = \left(\frac{R_{ft}}{P_{ft}}\right)$. Proceed in an analogous way to obtain a quantity index for materials, M_{ft}.

3. Using $P_{ft},\ Q_{ft}, M_{ft}$ (now properly estimated) and the estimated coefficients of the production and demand functions, obtain residuals $TFPQ_{ft}$ and D_{ft}. In estimating $TFPQ_{ft}$ and D_{ft} as residuals at this stage, we first regress P_{ft}, Q_{ft}, M_{ft}, L_{ft} and K_{ft} on sector*year effects and use only the residuals from those regressions, so that from this stage on, only idiosyncratic variation in $TFPQ_{ft}$ and D_{ft} is considered. More generally, our application only considers idiosyncratic (within sector-year) variation.

We now explain step 1 in detail.

4.1 Joint production-demand function estimation

We want to jointly estimate the log production and demand functions:

$$\ln Q_{ft} = \alpha \ln K_{ft} + \beta \ln L_{ft} + \phi \ln M_{ft} + \ln A_{ft} \tag{23}$$

[27] We initialize each plant's price index at P_{fB}, which takes into account the average price level in year B and the deviation of plant $f's$ product's prices from the average prices in the respective product category in that year. Details are provided in Appendix A.

and

$$\ln P_{ft} = -\frac{1}{\sigma}\ln Q_{ft} + \ln D_{ft} \tag{24}$$

where $Q_{ft} = \left(\frac{R_{ft}}{P_{ft}}\right)$. Using (10) and (22), the system can be rewritten:

$$\ln Q^*_{ft} = \alpha \ln K_{ft} + \beta \ln L_{ft} + \phi \ln M^*_{ft} + \frac{1}{\sigma_w - 1}\ln \Lambda^Q_{ft} - \frac{\phi}{\sigma_w - 1}\ln \Lambda^M_{ft} + \ln A_{ft} \tag{25}$$

and

$$\ln(\overline{P^*_{ft}} P_{fB}) = -\frac{1}{\sigma}\ln Q^*_{ft} - \left(\frac{1}{\sigma_w - 1}\right)\left(\frac{\sigma - 1}{\sigma}\right)\ln \Lambda^Q_{ft} + \ln D_{ft} \tag{26}$$

In practice, we estimate the parameters of (25) and (26), which are natural transformations of the original production and demand functions, rather than those original forms. This transformation enables us to specify the residuals from a log-linear specification which in turn permits specifying the moment conditions and estimating the parameters from linear GMM.

The usual main concern in estimating these functions is simultaneity bias. In the production function, this is the problem that factor demands are chosen as a function of the residual A_{ft}. A standard approach to deal with this problem is the use of proxy methods as in Olley and Pakes (1996); Levinsohn and Petrin (2003); De Loecker and Warzynski (2012); Ackerberg, Caves and Frazer (2015, ACF henceforth) and many others. In the demand function, simultaneity arises because both price and quantity respond to demand shocks. Usual demand estimation approaches rely on assumptions regarding orthogonality between demand and supply shocks at some particular level. Foster, Haltiwanger and Syverson (2008, 2016); Eslava et al. (2004, 2013) impose orthogonality between the levels of $TFPQ$ and demand shocks, while in Broda and Weinstein (2006); Hottman, Redding and Weinstein (2016) double-differenced demand and marginal cost shocks are assumed orthogonal. We build on these approaches, but take advantage of prices and quantities for both inputs and outputs, and the consequent possibility of jointly estimating (25) and (26), to relax the assumptions about covariance between demand and supply shocks that identify the elasticities of substitution across and within establishments.

We assume that $TFPQ$ and D_{ft} follow the following flexible laws of motion:

$$\begin{aligned}
\ln A_{ft} &= \pi^A_0 + \pi^A_1 \ln A_{ft-1} + \pi^A_2 \ln A^2_{ft-1} + \pi^A_3 \ln A^3_{ft-1} + \xi^A_{ft} \\
\ln D_{ft} &= \pi^D_0 + \pi^D_1 \ln D_{ft-1} + \pi^A_2 \ln D^2_{ft-1} + \pi^A_3 \ln D^3_{ft-1} + \xi^D_{ft}
\end{aligned}$$

That is, ξ^A_{ft} and ξ^D_{ft} are, respectively, the stochastic component of the innovation to $TFPQ$ and D_{ft}. Given this structure, our identification of production and demand elasticities (α, β, ϕ, σ, σ_w) uses standard linear GMM procedures, imposing the following set of moment

conditions (further details provided in Appendix F):

$$E\begin{bmatrix} \ln M^*_{ft-1} \times \xi^A_{ft} \\ \ln L_{ft} \times \xi^A_{ft} \\ \ln K_{ft} \times \xi^A_{ft} \\ \ln D_{ft-1} \times \xi^A_{ft} \\ \ln A_{ft-1} \times \xi^D_{ft} \\ \ln A_{ft} \\ \ln D_{ft} \end{bmatrix} = 0 \tag{27}$$

To write the moment conditions for M_{ft}, L_{ft} and K_{ft} in 27, we assume that materials are freely adjusted while the demand for capital and labor is assumed quasi-fixed. As traditional in ACF-based methods, inputs respond to stochastic innovations to $TFPQ$ contemporaneously or with a lag if, respectively, they are freely adjusted or quasi-fixed.[28] Thus, in (27) we require lagged materials demand to be orthogonal to current $TFPQ$ innovations, while L and K are required to be contemporaneously orthogonal to ξ^A_{ft}. The assumption that K is quasi-fixed is standard, as is that indicating that M adjusts freely.[29] L is also assumed quasi-fixed in our context because important adjustment costs have been estimated for the Colombian labor market (e.g. Eslava et al., 2013). We thus follow De Loecker et al. (2016) in treating L as quasi-fixed for purposes of estimation.

Meanwhile, the conditions that D_{ft-1} must be orthogonal to ξ^A_{ft} while A_{ft-1} must be orthogonal to ξ^D_{ft} identify σ and σ_w, following the logic that the slope of the demand function can be inferred taking advantage of shocks to supply.[30] Foster, Haltiwanger and Syverson (2008, 2016); Eslava et al. (2013) also relied on the logic that shocks to production identify the demand curvature, but imposed orthogonality between demand and technology shocks in levels (A_{ft} and D_{ft}). This effectively precludes the possibility that establishments endogenously invest in quality when they perceive better returns (as would be the case with higher $TFPQ$), or that they acquire technologies that increase production costs to produce better quality.[31] Hottman, Redding and Weinstein (2016); Broda and Weinstein (2006, 2010) address these concerns by imposing orthogonality between double-differenced demand and supply shocks (double differencing over time and varieties). Orthogonality between the double-differenced shocks may still be a strong assumption if, even within product groups, changes in quality require changes in production technologies.[32] Given our ability to specify demand and production separately using the price and quantity data of both output and inputs, we impose

[28] We also follow standard proxy methods and purge measurement error in a first stage of the estimation (Appendix F).

[29] For $\ln M_{ft-1}$ to be useful in the identification of ϕ, it must be the case that input prices are highly persistent. The AR1 coefficient for log materials prices is 0.95 in our sample.

[30] Production elasticities are initialized at MCO estimates, while σ is initialized at the estimate from an IV regression where $TFPQ_{ft}$ is used as an instrument in the demand equation, as in (e.g. Eslava et al., 2013). Using this initial estimate for σ for each sector, σ_w is initialized at a level such that $\frac{\sigma_w}{\sigma}$ equals the $\frac{\sigma_w}{\sigma}$ ratio for the median sector in HRW.

[31] R&D decisions that are endogenous to current profitability and affect future profitability, for instance, are present in Aw, Roberts and Xu (2011). Their framework does not separately identify the demand and technology components of profitability, but both could plausibly respond dynamically. In turn, the idea that quality is more costly to produce appears in Fieler, Eslava and Xu (2018), to characterize cross sectional correlations between quality and size.

[32] This is more of an issue for the earlier papers using harmonized trade data and not an issue for the recent

$E(\ln D_{ft-1} \times \xi^A_{ft}) = 0$ and $E(\ln A_{ft-1} \times \xi^D_{ft}) = 0$ which permit a correlation between $TFPQ$ and demand even over time within the plant. While we are still taking advantage of shocks to the supply curve to identify elasticities on the demand side, we only require that *innovations* in technical efficiency in period t be orthogonal to demand in levels in $t-1$, and that *innovations* in demand in period t be orthogonal to $TFPQ$ in levels in $t-1$, where these innovations come from a very flexible law of motion for $TFPQ_{ft}$ and D_{ft}.

Notice also that $TFPQ$ obtained as a residual from quality-adjusted Q is stripped of apparent changes in productivity related to within-establishment appeal changes, eliminating a source of correlation between appeal and technical efficiency stemming from measurement error. Moreover, since we use plant-specific deflators for both output and inputs, our estimation is not subject to the usual bias stemming from unobserved input prices (De Loecker et al., 2016).[33]

We implement this estimation separately for each three digit sector of ISIC revision 2, adapted for Colombia (CIIU-AC by its acronym in Spanish). There are 23 manufacturing sectors at this level. The estimated factor and demand elasticities are summarized in Table 1 and listed in Appendix C. Our results reveal close to constant returns to scale in production on average, but with non-negligible variation across three-digits sectors. The estimated elasticities of substitution across products within the establishment and across establishments stand at averages (over sectors) of 3.53 and 1.95, respectively, with substantial cross-sector variation (see also Appendix C). The revenue function curvature parameter stands at an average of 0.48, ranging between 0.18 and 0.69. For almost all sectors, this stands in sharp contrast to the 0.67 curvature parameter implied by usual assumptions of CRS in production, CES demand, and an elasticity of substitution of 3 (e.g. in HK). A correct estimation of the level of σ is crucial in adequately determining the effect of wedges both on the size distribution of plants and on aggregate efficiency, as will be clear in our results.

It is encouraging that we obtain plausible factor elasticities for all sectors at the three digits sector level. Proxy methods for the estimation of production functions are usually implemented in estimations at higher levels of aggregation, and frequently yield implausible results–in particular negative estimated factor coefficients for several sectors–at finer levels of disaggregation such as the one in our estimation.

Table 1: Factor and Demand Elasticities

	β	α	ϕ	σ_w	σ	σ_w/σ	γ	$\gamma(1-1/\sigma)$
Average	0.37	0.14	0.52	3.53	1.95	1.81	1.03	0.48
Min	0.15	0.03	0.23	2.15	1.20	1.77	0.93	0.18
Max	0.65	0.28	0.69	4.98	2.75	1.88	1.20	0.69

Note: Estimated factor and demand elasticities for 23 different sectors.

We use the within-plant estimated demand elasticity $\widehat{\sigma_w}$ to construct $\ln P_{ft} = \ln\left(P_{fB}\overline{P^*_{ft}}\right) +$

papers using barcode data such as Hottman, Redding and Weinstein (2016). Quality changes that require higher costs would show up between but not within products with barcode data.

[33] De Loecker et al. (2016), use plant-level deflators for output but not for inputs. This induces a bias stemming from unobserved input price heterogeneity.

$\frac{1}{\widehat{\sigma_w}-1}\ln\Lambda^{Q}_{ft}$ and subsequently recover $Q_{ft}=\frac{R_{ft}}{P_{ft}}$. We proceed in an analogous way to construct pm_{ft} and M_{ft}.[34] We then use Q_{ft}, M_{ft} and P_{ft} to obtain the residuals A_{ft} and D_{ft}. We use the estimated σ (at the three digit level of ISIC revision 3) to obtain the markup $\mu_{ft}=\frac{\sigma}{(\sigma-1)}\frac{1}{\left(1-s_{ft}\right)}$. For markup estimation, we use plant f's market share s_{ft} as its revenue share in its relevant product group, defined at the three digit group of the product classification. Products are classified according to the international CPC classification. There are 111 product groups at the CPC three digit level (while our "sectors" classification, defined using Colombian ISIC three digit level, has 23 classes), with an average number of plants close to 49, and a median of 24. To illustrate the level of aggregation of product groups and compare them to sectors, Appendix Table C3 lists 15 of the 111 product groups and Table C4 provides the distribution of number of plants by product groups, while Table C2 lists the 23 sectors for which we estimate separate production and demand parameters.

From this point, we work only with the within-sector variability of all variables of interest. In particular, we regress all outcome variables (revenue, employment, capital, materials, output prices, and input prices) against sector*year effects, and from this point use only residuals from those regressions. Also, as previously stated, when building $TFPQ$, D, and μ we only exploit idiosyncratic (i.e. within sector*year) variation in the levels of outcomes. It is these variables deviated from sector*year effects that we use when building life cycle growth for any variable ($\frac{Z_{ft}}{Z_{0t}}$ for each variable Z for each variable Z).[35]

5 Results: Size distribution in Levels and Growth

5.1 Plant attributes

Table 2 presents basic summary statistics for (the idiosyncratic component of the log of) sales, output, output prices, A_{ft}, D_{ft}, the residual wedge, markups, and input prices. The residual wedge, $\chi_{ft}^{1-\frac{1}{\sigma}}=(1-\tau_{ft})^{\gamma\kappa_1\left(1-\frac{1}{\sigma}\right)}\left(r_{ft}\chi^{\mathrm{K}}{}_{ft}\right)^{-\alpha\kappa_1\left(1-\frac{1}{\sigma}\right)}$, is obtained as a residual from equation 15, since we have measures of all other components.[36] We note that we have adjusted materials prices for quality, but have not done the same for wages as yet due to data constraints. Further below we quality-adjust wages for a subperiod for which this is possible.

Table 2 shows that quality/appeal and technical efficiency are negatively correlated in levels, consistent with Forlani et al. (2021), but weakly in our data. Also especially interesting is the negative and strong correlation of residual wedges with $TFPQ$ (-0.419) and demand shocks (-0.156), indicating that the most highly productive plants face greater barriers, i.e. correlated wedges. These basic correlation patterns are echoed in the role of different size determinants below.

[34] I.e. we use the same measurement approach incorporating multi-materials inputs to construct the plant-level deflator for materials, and use it to deflate expenditures in materials to arrive at materials inputs. For each plant, we use for materials the same elasticity of substitution used for outputs.

[35] We also winsorize life cycle growth for each variable at 1% and 99% to eliminate outliers that may drive the results of our decompositions.

[36] χ_t is no longer relevant once we focus solely on within sector*year variation.

It is also worth noting that sales, quality/appeal, and $TFPQ$ exhibit an important degree of persistence, above 0.9 in all cases (column 2). Residual wedges are also persistent, but with a much lower AR(1) coefficient of 0.727, which would be consistent, for instance, with these wedges reflecting adjustment costs which imply gaps in revenue with respect to efficiency for reasons that correlate with persistent revenue in a lumpy fashion.

Table 2: Descriptive Statistics

Panel A. Number of plants, number of products and materials per plant-year

Number of plants		Number of products per plant				Number of materials per plant			
Total	Avg. year	Avg.	P25	P50	P75	Avg.	P25	P50	P75
19,326	5,434	3.77	1	3	5	12.05	6	10	15

Panel B.Standard deviations and correlation coefficient for outcomes and fundamentals (within sector*year, all variables in logs, average sector)

	Standard Deviation	AR(1) coefficient	Sales	Output	Output prices	TFPQ	D (quality / appeal)	Material prices	Average wage	Markup	Structural Wedge
Sales	1.560	0.981	1.000								
Output	1.691	0.980	0.921	1.000							
Output prices	0.632	0.958	0.000	-0.380	1.000						
TFPQ	0.746	0.905	0.191	0.458	-0.728	1.000					
D (quality/appeal)	0.892	0.977	0.930	0.724	0.342	-0.071	1.000				
Input prices	0.593	0.949	-0.047	-0.102	0.155	0.274	0.011	1.000			
Average wage	0.455	0.847	0.657	0.586	0.055	0.132	0.633	-0.017	1.000		
Markup	0.074	0.971	0.487	0.451	-0.009	0.091	0.447	-0.020	0.338	1.000	
Residual wedge	0.408	0.727	-0.143	-0.107	-0.066	-0.419	-0.156	0.001	0.046	0.109	1.000
Lagged D (quality/appeal)	0.884	0.974	0.898	0.704	0.319	-0.088	0.962	0.007	0.632	0.446	-0.091

Note: Descriptive statistics are restricted to a sample of plants observations which have information on all measured plant attributes.

Markups display modest variation across plants in absolute terms and relative to the variability in size. A few sector*year observations have dominant plants with very large markups (Appendix C, Table C5). Markups are positively correlated with $TFPQ$, D, and wages, besides plant size. They are also positively correlated with our residual wedge $\chi_{ft} = (1-\tau_{ft})^{\gamma\kappa_1} * \left(r_{ft}\chi^{\mathrm{K}}{}_{ft}\right)^{-\alpha\kappa_2}$.[37]

A few other attributes of $TFPQ_{ft}$ and D_{ft}, explored in Appendix E, are telling of the economic content of these residuals. D_{ft} is positively correlated with different efforts for marketing and product innovation. Conditional on sales, there is a strong positive correlation between D_{ft} and product innovation, as captured by the introduction of new high-price products or increases in quality of existing products, reflected in large increases in their prices. D_{ft} is also positively correlated with efforts to build a client base, such as advertisement expenditures and the use of the internet to provide customer support. $TFPQ_{ft}$, in turn, exhibits a strong positive correlation with investment in fixed assets, stronger than that

[37]The structural markup in our approach corresponds to a wedge-adjusted version of the widely used empirical estimation of the markup in De Loecker and co-author's work (2012; 2016; 2020). Denoting the markup calculated with such approach as μ^{DL}, Appendix B shows that under the demand and production structure in this paper the following relationship holds: $\mu_{ft} = \mu^{DL}_{ft} * (1-\tau_{ft}) = \frac{\theta^v_{ft}}{\frac{C^v_{ft}V_{ft}}{P_{ft}Q_{ft}}} * (1-\tau_{ft})$ where V_{ft} is some variable input, C^v_{ft} its unit cost and θ^v_{ft} is the output elasticity for the variable input. We find that $\widehat{\mu}^{DL}_{ft} = \frac{\phi}{\frac{p^m{}_{ft}M_{ft}}{R_{ft}}}$, compared with μ_{ft}, $\widehat{\mu}^{DL}_{ft}$ displays a much larger variance and much weaker correlations with sales and demand shifters, and a stronger correlation with $TFPQ$.

between D_{ft} and investment once revenue has been controlled for. Conditional on sales, technical efficiency ($TFPQ_{ft}$) is negatively correlated with the introduction of high price products, which suggests that producing high quality requires process efforts that induce a trade-off between producing quantity and quality.

5.2 Plant attributes vs. size

We now explore the role of these different plant attributes in determining establishment size. Equations 18 and 15 imply that revenue depends on quality-adjusted productivity $TFPQ\_HK_{ft}$ and composite distortions captured by (an inverse function of $TFPR_{ft}$) :

$$R_{ft} = \left[\frac{TFPQ\_HK_{ft}}{(\gamma * TFPR_{ft})^{\gamma}}\right]^{\kappa_2} = d_{ft}^{\kappa_1} a_{ft}^{\kappa_2} pm_{ft}^{-\phi\kappa_2} w_{ft}^{-\beta\kappa_2} \mu_{ft}^{-\gamma\kappa_2} \left(\widehat{\chi}_t \chi_{ft}\right)^{1-\frac{1}{\sigma}} \tag{28}$$

where $\kappa_1 = \frac{1}{1-\gamma\left(1-\frac{1}{\sigma}\right)}$, $\kappa_2 = \left(1-\frac{1}{\sigma}\right)\kappa_1$, and γ and σ have been estimated as explained above.

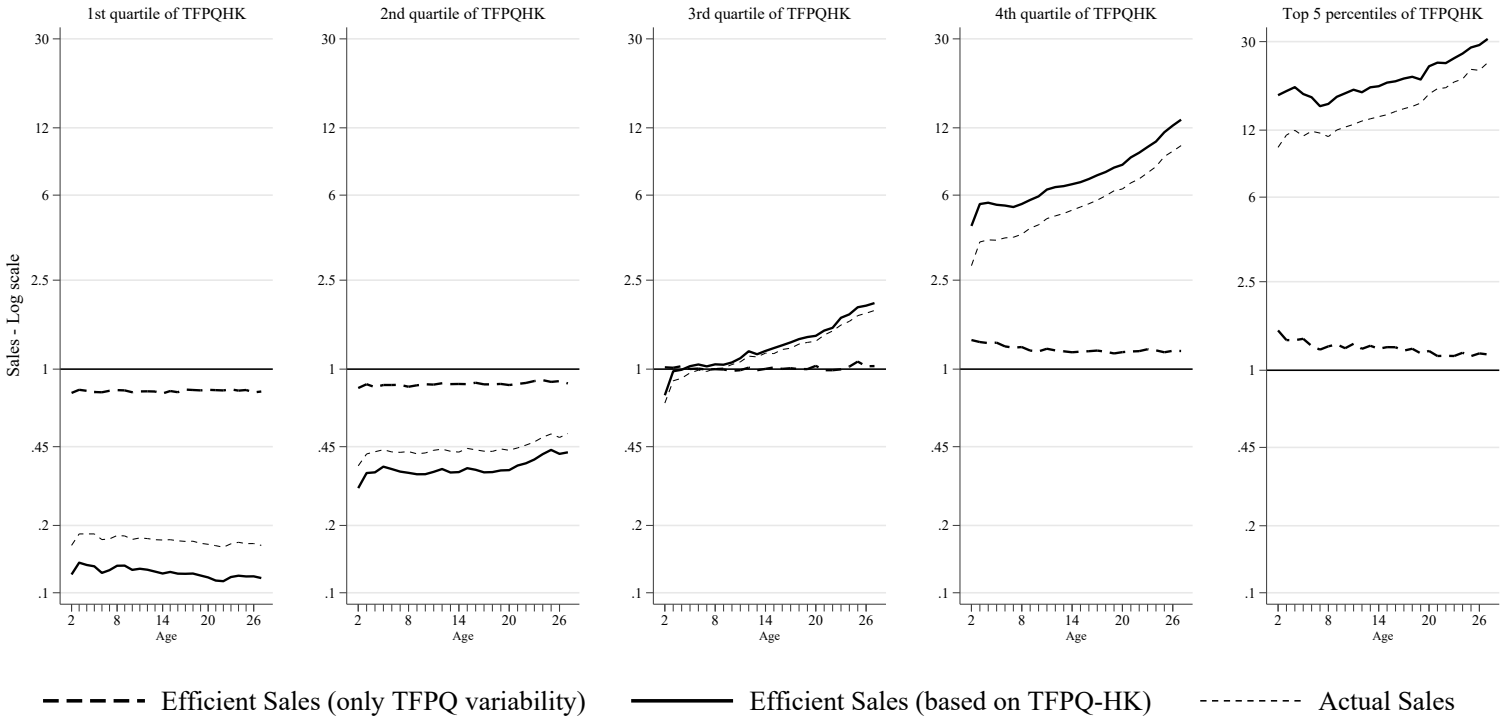
Figure 1 shows the evolution, over the life cycle of plants with different levels of quality-adjusted productivity $TFPQ\_HK_{ft}$, of actual sales and sales predicted based on the different components of R_{ft} in equation 28 (on a log scale, given large differences). Each panel represents establishments in a given section of the distribution of $TFPQ\_HK_{ft}$. The dotted line shows actual (average) sales, while the solid line corresponds to efficient sales, i.e. sales as would be determined by the quality-adjusted productivity $TFPQ\_HK_{ft}$ in the absence of wedges: $\widehat{R_{ft}} = \left[\frac{TFPQ\_HK_{ft}}{\gamma^{\gamma}}\right]^{\kappa_2}$. The gap between the two captures the composite wedge emanating from $TFPR_{ft}$, which includes the effects of input prices, idiosyncratic markups, and residual wedges. Efficient sales $\widehat{R_{ft}}$ are further decomposed in Figure 1 into those based only on $TFPQ_{ft}$ ($\widehat{R'_{ft}} = \left[\frac{TFPQ_{ft}}{\gamma^{\gamma}}\right]^{\kappa_2}$), corresponding to the dashed line, and the component based on D_{ft}, corresponding to the gap between the solid and dashed lines.

Three main messages emanate from Figure 1. First, though plants with higher levels of quality-adjusted productivity are larger than lower productivity ones (right vs. left panels), sizable wedges that are negatively correlated with $TFPQ\_HK_{ft}$ imply dampened size differentials over the productivity distribution with respect to the efficient levels. That is, the distance between low and high $TFPQ\_HK$ plants in their efficient size (solid lines) is larger than the distance in actual size (dotted lines), and plants in the first quartile of $TFPQ\_HK_{ft}$ face large positive composite wedges, while the opposite is true in the upper sections of the productivity distribution. These composite wedges are large: plants in the bottom $TFPQ\_HK$ quartile are on average 42% larger than efficient, and those in the top quartile are 24% smaller with respect to that benchmark. Especially notable is the impact in the tails as illustrated in the top 5%, where the dampening effect of wedges is very large.

Second, younger plants face larger composite wedges. This is especially the case in the top sections of the productivity distribution: young high productivity plants are especially undersized vis-a-vis their efficient size. [38]

[38] In the top quartile of $TFPQ\_HK$, the size gap vs. efficiency of establishments up to two years old exceeds 34%, going up to 42% for the 95th percentile, while across all ages the average gap is 23%, as mentioned above.

Figure 1: Sales by Age. Actual and Efficient Based on $TFPQ\_HK$



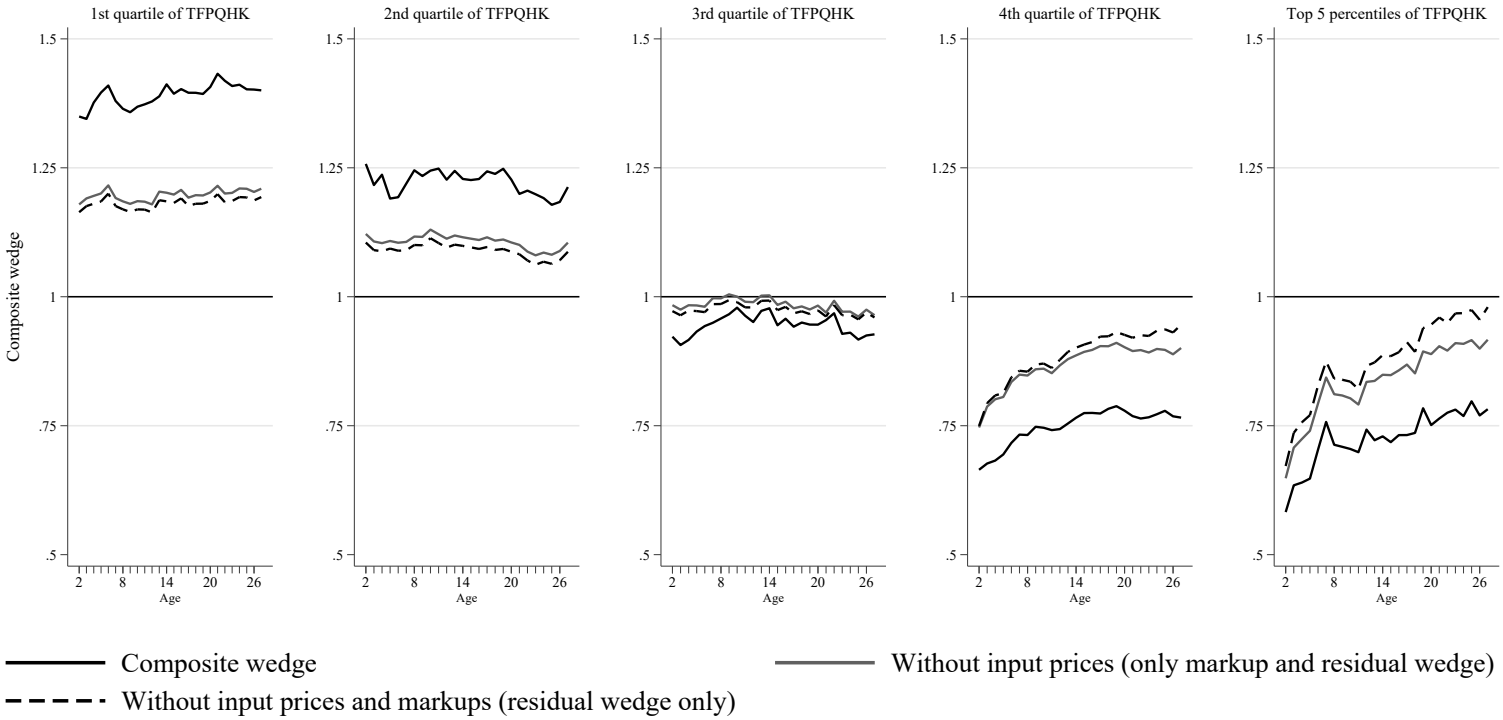
Note: Lines depict the average sales predicted by considering $TFPQ$ and $TFPQ\_HK$ and the actual observed average sales based only on idiosyncratic variation. Information from the sample of plant observations for which all measured attributes are observable. Quartiles are calculated for each age

Finally, the comparison between the solid and dashed lines (based, respectively on $TFPQ\_HK_{ft}$ and $TFPQ_{ft}$) shows that efficient size differentials between the bottom and top segments of the distribution are mostly due to D_{ft} differentials. Although higher productivity plants are also more technically efficient, size differentials based only on $TFPQ$ dwarf overall differences in efficient size. Also, demand growth is the driver of life cycle growth, while $TFPQ$ is basically flat over the life cycle in all panels of the figure.

Figure 2 further decomposes the composite wedge into its components related to input prices, markups, and residual wedges. The solid black line represents the composite wedge and is equal to the gap between efficient and actual sales in Figure 1. The role of each of its components is observed by shutting them down progressively: the grey solid line corresponds to counterfactual composite wedges in the absence of input price heterogeneity and the dashed line to a version where only residual wedges remain (no idiosyncratic markups or input prices).

Of the 42% (-24%) average composite wedge for plants in the bottom (top) quartile of $TFPQ\_HK$, 19% (-13%) corresponds to input prices, as seen in the reduction of (absolute) wedges when input price heterogeneity is shut down (grey solid line). We show further below that almost all of the dampening effect of input prices on revenue variation is explained by wages, with a minimal role for the prices of material inputs. Subsequently shutting down idiosyncratic markups makes little difference in general, adding no more than 3 p.p. (grey vs. dashed lines). That is, idiosyncratic markups play a minor role, explaining one percentage

Figure 2: Composite Wedges by Age: The Role of $TFPR$ and its Components



Note: Lines depict average composite HK wedges and its components based only on idiosyncratic variation. Information from the sample of plant observations for which all measured attributes are observable.

point or less of the composite wedge in most cases. The most important exception is the case of establishments at the very top of the productivity distribution, where some plants display very high markups that keep them undersized. In the 95th percentile, five out of the 23 points of composite wedges correspond to markups across all ages. This is especially the case for older plants: while for startups of up to five years only two points of the composite wedge correspond to markups, for plants aged 20 and older markups explain over five points of the composite wedge. This is despite composite wedges being much larger for younger plants: 42% for startups of at most two years vs. 20% for those aged 20 and above in the 95th percentile of $TFPQ\_HK_{ft}$.

The residual wedge (dashed line in Figure 2), $\chi_{ft} = (1-\tau_{ft})^{\gamma\kappa_1} * \left(r_{ft}\chi^{\mathrm{K}}{}_{ft}\right)^{-\alpha\kappa_1}$, is large and positive for low productivity plants and large but negative for high productivity plants, especially the youngest. That is, these residual wedges also contribute to making low (high) $TFPQ\_HK$ plants larger (smaller) than efficient. Though we cannot directly measure the user cost of capital and thus decompose χ_{ft} into its revenue and factor-biased components, Appendix I implements an indirect approach to this additional decomposition. Figure I1, which replicates Figure 2 dissecting the role of the (distortions adjusted) user cost of capital, shows that most of the residual wedge in Figure 2 in fact corresponds to a revenue (or factor-unbiased) component.

We quantify the overall roles of these factors in explaining size variability through a vari-

ance decomposition of R_{ft}. We follow a two-stage procedure, similar to that in Hottman, Redding and Weinstein (2016), the details of which are provided in Appendix G. The contribution of each (log) plant attribute to the variance of (log) sales is given by the product between its coefficient in equation 28, its correlation coefficient with sales, and the ratio of its standard deviation to that of sales. For instance, the contribution of $TFPQ$ to the variance of sales is given by the product: $\kappa_2 * corr\left(a_{it}, R_{it}\right) * \frac{std(a_{it})}{std(R_{it})}$. The contributions of the different components add up to 1. We implement the variance decomposition by age (see Appendix G for details).

Table 3: Variance Decomposition of Sales

	Panel A: Unweighted							
	Levels decomposition				Growth decomposition			
	Weighted avg. ages	Age 3	Age 10	Age 20	Weighted avg. ages	Age 3	Age 10	Age 20
TFPQ-HK	**1.139**	**1.184**	**1.148**	**1.129**	**1.216**	**1.317**	**1.247**	**1.194**
TFPQ	0.081	0.131	0.087	0.074	0.142	0.252	0.152	0.112
Demand	1.058	1.053	1.061	1.055	1.074	1.065	1.095	1.082
Composite (HK) wedge	**-0.139**	**-0.184**	**-0.148**	**-0.129**	**-0.216**	**-0.317**	**-0.247**	**-0.194**
Material prices	0.003	0.009	0.001	0.005	-0.005	-0.011	-0.009	-0.005
Wages	-0.073	-0.072	-0.069	-0.078	-0.046	-0.053	-0.056	-0.047
Markup	-0.019	-0.011	-0.014	-0.018	-0.009	-0.006	-0.006	-0.008
Residual wedge	-0.049	-0.110	-0.066	-0.038	-0.156	-0.248	-0.175	-0.134
Marginal cost HRW	**-0.039**	**-0.042**	**-0.047**	**-0.037**	**-0.065**	**-0.059**	**-0.088**	**-0.074**
	Panel B: Revenue weighted							
	Levels decomposition				Growth decomposition			
	Weighted avg. ages	Age 3	Age 10	Age 20	Weighted avg. ages	Age 3	Age 10	Age 20
TFPQ-HK	**1.141**	**1.183**	**1.167**	**1.140**	**1.286**	**1.384**	**1.349**	**1.207**
TFPQ	0.085	0.116	0.120	0.109	0.206	0.325	0.209	0.167
Demand	1.056	1.067	1.047	1.031	1.080	1.059	1.140	1.040
Composite (HK) wedge	**-0.141**	**-0.184**	**-0.167**	**-0.141**	**-0.287**	**-0.385**	**-0.350**	**-0.208**
Material prices	-0.003	-0.002	-0.007	0.003	-0.003	0.025	-0.030	0.019
Wages	-0.073	-0.067	-0.075	-0.078	-0.043	-0.087	-0.039	-0.036
Markup	-0.077	-0.023	-0.062	-0.085	-0.031	-0.021	-0.022	-0.025
Residual wedge	0.012	-0.092	-0.024	0.020	-0.210	-0.302	-0.258	-0.166
Marginal cost HRW	**0.021**	**-0.044**	**0.014**	**0.053**	**-0.050**	**-0.039**	**-0.118**	**-0.015**

Note: Weighted average across ages corresponds to the weighted average up to and including age 50. We estimate a decomposition for each sector and then estimate weighted averages by sector revenue. $TFPQ\_HK$ values correspond to the sum of the contributions of D and $TFPQ$; Composite (HK) wedge is the sum of the contributions of input prices, markups, and residual wedges; Marginal cost HRW is the sum of the contributions of $TFPQ$, input prices, and residual wedges.

Results are presented in the upper left panel of Table 3. Variation in $TFPQ\_HK_{ft}$ accounts for more than 100% of the variance of sales and sales growth across plants within a sector. In the levels decomposition, this contribution is 114% averaged across ages. That is, composite wedges dampen the variability of sales by 14% (corresponding rows in bold in Table 3).

Focusing first on what contributes most to the $TFPQ\_HK$-based sales variability, we find that quality/appeal contributes the most–by far– to it, but $TFPQ$ is not negligible.

Averaging over ages, the contributions to the variance of sales are 1.058 for the demand shifter and 0.081 for $TFPQ$. For the macro misallocation literature, which has placed such strong emphasis on isolating quantity productivity, this implies that quality is crucial in a proper conceptualization of productivity, so that quantity productivity measures should be quality-adjusted. Compared to the literature on demand vs. cost factors as determinants of business size, in turn, these results are in consonance with findings pointing at the dominant role of demand (quality/appeal), but they also show that the role of technical efficiency ($TFPQ$) is far from negligible, explaining 7% of the variability of sales. This highlights the limitations of estimating the role of cost factors from a residual, as in HRW. Such an approach has led to the interpretation that cost dimensions play a negligible role, since $TFPQ$, input prices, and residual wedges are aggregated into this residual marginal cost component. As is also the case in HRW's application for the US, we find that the marginal cost composite makes a negligible contribution to sales variability, of just -3.9% in the levels decomposition (bottom row in Table 3). But, Table 3 also shows that this small contribution obscures the 8.1% contribution of technical efficiency by lumping it together with those of input prices (-7%) and residual wedges (-5%), where the latter is not necessarily attributable to cost side considerations.

Moving to the question of what explains the 14% contribution of composite wedges in the levels decomposition, we find that input prices account for 7 of those 14 points, fully explained by wages. Meanwhile, markups explain 1.9 points, and the residual wedges the remaining 5 points (to which the factor-biased component deducts 2.9 so that the τ_{ft} component actually represents 7.9 points, see Appendix I). Results also indicate that markups play minor roles for most plants, but play an important role at the very top end of the sales distribution, where the top performers also display high markups that dampen their size. Interestingly, the contribution of markups to sales scales up by a factor of four when the decomposition weighs plants based on their revenue to -7.7% (lower left panel). The high weight of markups in the weighted decomposition (by contrast to the unweighted one) reflects the fact that a few large plants have very large market shares which significantly influence their size.

The variation in wages that plays such an important role to explain size differentials across plants might reflect many factors, including the geographic segmentation of labor markets as well as institutional barriers or other frictions in the labor market, from search costs to regulatory distortions to bargaining/monopsony power and labor allocation. For example, canonical search and matching models of the labor market (see, e.g., Mortensen and Pissarides (1994) yield dispersion in wages positively correlated with productivity induced by the bilateral bargaining in the face of frictions. Relatedly, on-the-job search models (e.g., Burdett and Mortensen (1998) and the many subsequent papers) imply a form of dynamic monopsony in the labor market so that workers of similar skills may be paid very different wages across firms. Wage dispersion might also rather reflect unmeasured quality differences since, by contrast to material inputs prices we are unable to quality adjust wages for our entire sample period since the data does not break labor into skill categories for the full extent of our estimation period. To address the relative importance of quality heterogeneity for labor, we now take advantage of data on broad skill categories available for 2000-2012. The available skill categories are production workers without tertiary education, production workers with tertiary education and administrative workers. We construct, for that subperiod, quality-adjusted wages using an approach analogous to the one we use to build quality-adjusted

materials and output prices.[39] Table 4 presents the results of this adjusted exercise (third column, to be compared to the second column, which presents the unadjusted decomposition for the period for which quality adjusting wages is possible).

Adjusting wages for labor quality reduces the contribution of wage dispersion in accounting for sales heterogeneity, and also that of composite wedges, while it decreases the contribution of TFPQ. This is not surprising as adjusting for labor quality impacts the measurement of technical efficiency. The effect of quality adjusting wages, however, is not large even for $TFPQ$ and wages, and importantly does not affect other components. In particular, the contribution of wages falls from 6.1% to 4.8%. The main message that wage dispersion plays a crucial role in explaining the magnitude of the contribution of (composite) wedges to the distribution of plant sizes remains after quality adjusting wages, and other components are not impacted. Thus, even if further labor quality adjustment is warranted, the lack of sensitivity of the contribution of other components suggests the inferences we draw for other components are robust. We, thus, proceed with our main full sample results as a baseline that provides robust inferences.

Table 4: Variance Decomposition of Sales With Quality Adjusted Wages

	Levels			Life cycle growth		
	Unadjusted wage		Q-adj. Wage	Unadjusted wage		Q-adj. Wage
	1982-2012	2000-2012	2000-2012	1982-2012	2000-2012	2000-2012
TFPQ-HK	**1.139**	**1.146**	**1.133**	**1.216**	**1.235**	**1.219**
TFPQ	0.081	0.095	0.083	0.142	0.175	0.158
Demand	1.058	1.051	1.051	1.074	1.060	1.060
Composite (HK) wedge	**-0.139**	**-0.146**	**-0.133**	**-0.216**	**-0.236**	**-0.219**
Material prices	0.003	0.006	0.006	-0.005	-0.008	-0.008
Wages	-0.073	-0.061	-0.048	-0.046	-0.034	-0.016
Markup	-0.019	-0.012	-0.012	-0.009	-0.006	-0.006
Residual wedge	-0.049	-0.079	-0.079	-0.156	-0.188	-0.189
Marginal cost HRW	**-0.039**	**-0.038**	**-0.038**	**-0.065**	**-0.054**	**-0.054**

Note: Weighted average across ages corresponds to the weighted average up to and including age 50. We estimate a decomposition for each sector and then estimate weighted averages by sector revenue. $TFPQ\_HK$ values correspond to the sum of the contributions of D and $TFPQ$; Composite (HK) wedge is the sum of the contributions of input prices, markups, and residual wedges; Marginal cost HRW is the sum of the contributions of $TFPQ$, input prices, and residual wedges.

Back to the top-left panel of table 3, the contributions of the different attributes to sales vary depending on the plant's age. Quality/appeal becomes increasingly important compared to $TFPQ$ for older plants. The ratio of contributions to sales of D relative to $TFPQ$ is close to 8 at age 3, but by age 20 it is close to 14. This is because the correlation between sales and

[39] That is, labeling quality-adjusted wages as $\widehat{w}_{ft}$ and denoting the set of the three skill categories in the data as Ω^w, the wage index is given by $\ln \frac{\widehat{w}_{ft}}{\widehat{w}_{ft-1}} = \sum_{j \in \Omega^w} \ln \left(\frac{w_{fjt}}{w_{fjt-1}} \right)^{\frac{1}{3}} + \frac{1}{\sigma_w - 1} \ln \lambda_{ft}^{w,QRW}$

where $\lambda_{ft}^{w,QRW} = \prod_{j \in \Omega^w} \left(\frac{s_{fjt}^w}{s_{fjt-1}^w} \right)^{\frac{1}{3}}$ and s_{fjt}^w is the share of skill class j in f's payroll at time t. We then build a quality-adjusted labor input given by the payroll deflated with our adjusted wages. $TFPQ$ is also re-calculated using this quality-adjusted input.

$TFPQ$ decreases for older plants, while that between sales and demand remains fairly stable. In other words, technical efficiency is particularly important for (composite) productivity for younger plants. Residual wedges, χ_{ft}, also play a more important dampening role at the youngest ages. The top quartile of productivity is crucial in understanding the decreasing importance of composite wedges for sales variability over the life cycle (2.

We conduct an analogous decomposition for the life cycle growth of sales, $\frac{R_{ft}}{R_{f0}}$.[40] We emphasize that we can measure life cycle growth directly using longitudinal data for each plant, rather than relying on cross-cohort comparisons. This approach addresses the usual selection concern in the literature on business' life cycle growth (Eslava, Haltiwanger and Pinzón, 2022). Results, presented in the right panel of Table 3, are broadly consistent with those in levels. However, the contributions of technical efficiency and residual wedges to the variability of sales are larger (in absolute value) for life cycle growth than the level of sales. The decomposition of residual marginal costs, HRW, also has interesting implications over the life cycle in the growth decomposition. At age 3, the residual marginal cost of 5.9% is accounted for a 25.2% technical efficiency component, a residual wedge component of -24.8% and -6.4% from input prices. By age 20, the residual marginal cost of -7.4% is accounted for by an 11.2% technical efficiency component, a residual wedge component of -13.4%, and -5.2% from input prices.

We only characterize and decompose life cycle size and growth for survivors up to any given age. Appendix H contrasts patterns for plants that survive into the future vs. exits-to-be (those for which year t is their last period in activity), showing that average patterns are mainly driven by plants that will survive, so that the exit bias is small. Appendix H also shows that: 1) Exits-to-be are much smaller than the average survivor and display markedly lower quality, and slightly lower $TFPQ$, but also pay lower wages; 2) Results for the sales decomposition are in general robust to selection, in the sense of being similar for survivors-to-be and exits-to-be. However, $TFPQ$ plays a relatively more important role vis-a-vis demand for exiters compared to survivors, suggesting that technical efficiency is important for survival. Wedges are also negative for exits-to-be, and play a more important role for them than survivors. That is, even though exits-to-be are relatively small, they remain significantly oversized compared to efficiency up to the time of their exit.

6 Results: aggregate productivity

Our results for the cross section imply that the distribution of quality-adjusted productivity, $TFPQ\_HK$, is dominated by the dispersion of quality/appeal, and that size dispersion across plants is dampened by the presence of composite wedges that are negatively correlated with quality-adjusted productivity. This section assesses: 1) the efficiency loss generated by distortions to the size distribution, and how each of the components of composite wedges contributes to that loss; and 2) the value that the dispersion in quality-adjusted productiv-

[40] We build the growth of variable Z over the life cycle of a plant at a given age as $\frac{Z_{f,age}}{Z_{f,0}}$ where Z_{f0} is the level of Z at f's birth, calculated as the average for ages 0 to 2. By averaging over the plant's first few years in operation we deal with measurement error coming, for instance, from partial-year reporting (e.g. if the plant was in operation for only part of its initial year). A plant's age in year t is the difference between the current year, t, and the initial year of operation.

ity ($TFPQ\_HK$), and each of its two components, efficiency and quality/appeal, have for aggregate productivity.

Aggregate Efficiency, AE_t, as characterized by equation 20, expresses the gap between actual aggregate productivity and productivity in a world with no $tfpr_{ft} = \frac{C_{ft}\mu_{ft}}{\gamma(1-\tau_{ft})}$ dispersion (where underlines were previously introduced to denote the idiosyncratic components of $TFPR$). That is, the benchmark is a world without idiosyncratic composite wedges. We calculate AE_t for each sector in an average year and then aggregate across sectors using sector revenue weights. Results are displayed in Table 5.

The top row of Column (1) shows the value that this aggregate object takes in our data. There is an aggregate productivity loss of 37.4% with respect to the efficient benchmark. Equation 21 shows that this loss depends on the covariance between (functions of) the composite wedge and $TFPQ\_HK$, as well as the expected value and dispersion of composite wedges. As shown in Appendix L, the estimated loss in our application in fact stems from a strong negative correlation between the composite wedge and productivity, in an environment with high dispersion of these wedges.

This estimated 37.4% efficiency loss corresponds to the combined effect of input price dispersion, idiosyncratic markups, and residual wedges. To assess the impact of these different distortions on aggregate efficiency, we progressively shut down the dispersion coming from each of them and their combinations. This is shown in the remaining rows of Column (1). Eliminating sources of $tfpr$ dispersion brings aggregate productivity closer to its efficient level. All three dimensions that we measure separately play non-negligible roles in explaining efficiency losses, with input price dispersion having the most bite, but markups and residual wedges also being important. If input prices were the only source of dispersion the efficiency loss would be 31.9% (AE=0.679). The corresponding efficiency loss from markup dispersion alone would be close to 10.5% and the one from residual wedges alone 16% ($AE = 0.895$ and $AE = 0.84$, respectively). The fact that input prices, markups, and residual wedges are closely interconnected implies that the efficiency losses from the three individual components do not add up to the total loss from shutting down the composite wedge.

There is negligible change in these results when we quality-adjust wages (see Table 6) suggesting that the efficiency losses from input price dispersion reflect features of the labor market other than differentials in the quality of labor that systematically lead high productivity plants to face higher wages. While establishing what those features of the labor market are is beyond the scope of our paper, the fact that high-productivity plants are at the same time the largest and the ones that pay high wages is not consistent with a simple story of monopsony power by large producers in the labor market. Importantly, the impact of the sales components other than wages (i.e., markups and residual wedges) is robust to quality-adjusting wages.

The important contribution of idiosyncratic markups for aggregate inefficiencies is distinct from their relatively minor role in accounting for the unweighted size distribution. The fact that a few plants do exhibit significant markups, and that these tend to be high-productivity and large size plants and thus carry a large weight in aggregate productivity (both efficient and actual), explains the important role played by markups in aggregate productivity and efficiency losses. It also resonates with the finding that markups do carry important weight to explain plant size differentials on an activity-weighted basis (Table 3).

Table 5: Allocative Efficiency: The Role of Distortions

$$AE_t = \left(\frac{1}{N_t} \sum_{I_t} \left[\left(\frac{\Delta_{ft}^{\frac{1}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\bar{\Delta}_t} \right) \left(\frac{tfpr_{ft}^{\frac{\gamma}{\sigma\left(1-\gamma\left(1-\frac{1}{\sigma}\right)\right)}}}{\overline{tfpr_t}} \right)^{-1} \right]^{\sigma-1} \right)^{\frac{1}{\sigma-1}}$$

	Sector type					
	All	Low $\gamma\left(1-\frac{1}{\sigma}\right)$	Intermediate $\gamma\left(1-\frac{1}{\sigma}\right)$	High $\gamma\left(1-\frac{1}{\sigma}\right)$	Low $sd(\mu_{ft})$	High $sd(\mu_{ft})$
	(1)	(2)	(3)	(4)	(5)	(6)
AE	0.624	0.727	0.651	0.401	0.671	0.405
Shutting down markups and wedges (only input price disp. remain)	0.679	0.807	0.715	0.392	0.732	0.430
Shutting down input prices and wedges (only markup disp. remain)	0.895	0.968	0.898	0.807	0.941	0.679
Shutting down input prices and markups (only wedges remain)	0.840	0.854	0.861	0.733	0.795	1.054
Shutting down wedges (only input price and markup disp. remain)	0.619	0.776	0.650	0.321	0.688	0.292
Shutting down markups (only input price disp. and wedges remain)	0.704	0.737	0.752	0.464	0.685	0.797
Shutting down input prices (only markup disp. and wedges remain)	0.762	0.855	0.761	0.667	0.787	0.640
Shutting down all (no TFPR dispersion)	1.000	1.000	1.000	1.000	1.000	1.000
Number of sectors	23	4	16	3	19	4
Range of parameter		$[0.18, 0.31]$	$[0.37 - 0.6]$	$[0.65, 0.69]$	< 0.1	> 0.1

Note: In this table we consider input prices combined, which include material prices and wages.

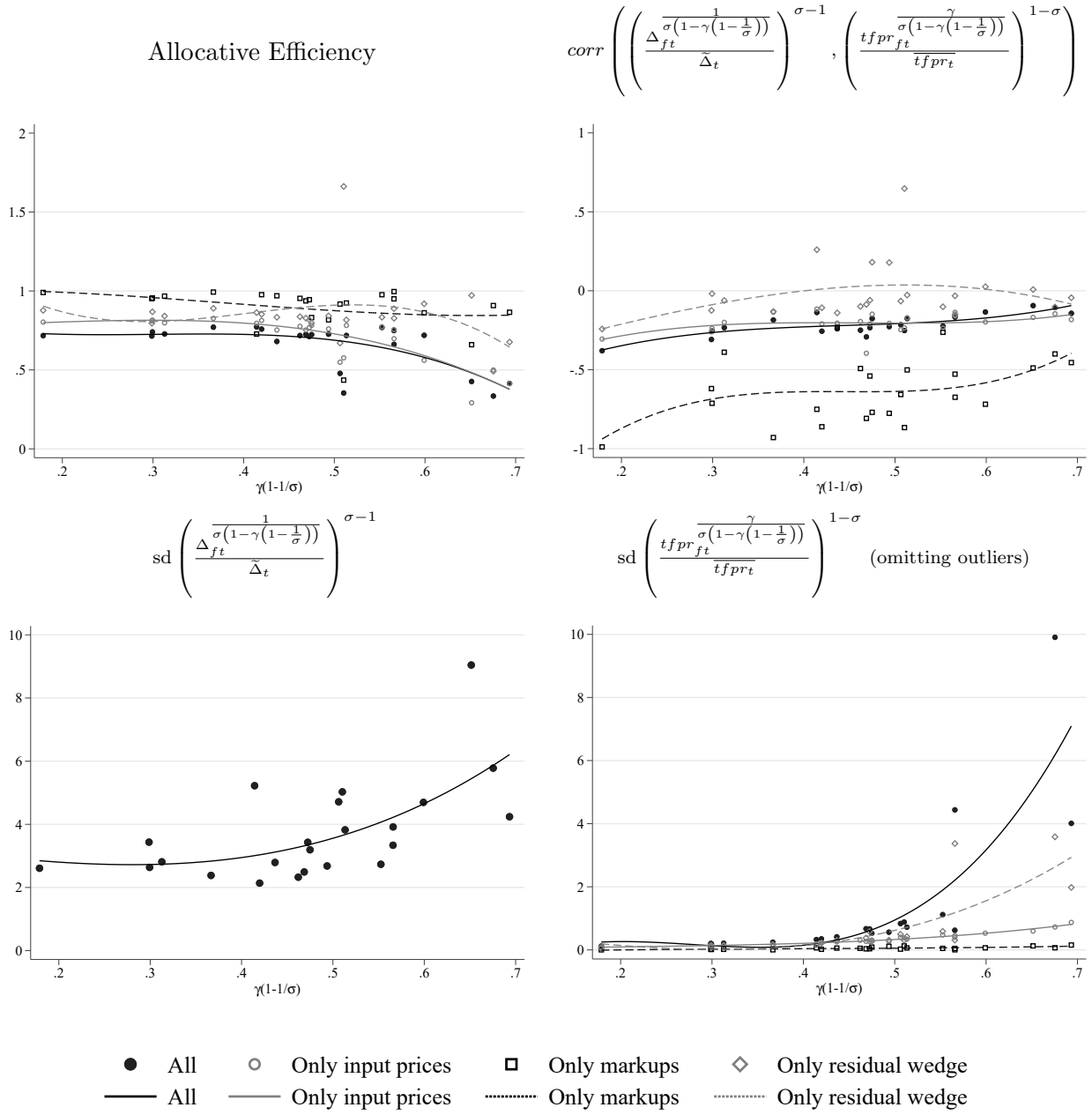
Table 6: Allocative Efficiency With Quality Adjusted Wages

	Unadjusted wage		Q-adj. Wage
	1982-2012	2000-2012	2000-2012
AE	0.624	0.579	0.586
Shutting down markups and wedges (only input price disp. remain)	0.679	0.658	0.687
Shutting down input prices and wedges (only markup disp. remain)	0.895	0.887	0.890
Shutting down input prices and markups (only wedges remain)	0.840	0.833	0.815
Shutting down wedges (only input price and markup disp. remain)	0.619	0.598	0.622
Shutting down markups (only input price disp. and wedges remain)	0.704	0.680	0.681
Shutting down input prices (only markup disp. and wedges remain)	0.762	0.730	0.722

Note: In this table we consider input prices combined, which include material prices and wages.

There is significant heterogeneity across sectors in terms of both the size of efficiency losses from distortions and the role that each type of distortion plays. Efficiency losses are in general larger in sectors where the revenue function displays little curvature (high returns to scale in production, γ, or high elasticity of substitution σ), so that efficiency would require much larger concentration of activity than seen in the data. This is seen in Figure 3 and in Columns (2) to (6) of Table 5. The top left panel of Figure 3 displays allocative efficiency by sector, as a function of the sector's $\gamma(1-\frac{1}{\sigma})$, showing the generally negative relationship between this parameter and allocative efficiency. Not only allocative efficiency tends to be lower in sectors with high $\gamma(1-\frac{1}{\sigma})$, but the relationship is nonlinear, with large drops in AE as one moves towards the highest values of $\gamma(1-\frac{1}{\sigma})$. The other panels of Figure 3 show that this is because the dispersion of both $TFPR$ and $TFPQ$ is higher at the right end of the $\gamma(1-\frac{1}{\sigma})$ spectrum and despite a less negative correlation between the composite wedges and the $TFPQ$ terms of AE. The relationship between revenue curvature and AE, however, is not strictly monotonic, since the relationship between $TFPR$ and $TFPQ$ also depends on the relative weights of different factors of production, and of σ (more on this in section 7).

Figure 3: Allocative Efficiency vs. $\gamma(1-\frac{1}{\sigma})$



Note: Lines correspond to a degree 3 polynomial fit. In panel 4 we omit outliers for the following three sectors: sector 355 with $\gamma(1-1/\sigma)=0.60$ and values of 34.0 for "All" and 37.3 for "only residual wedge"; sector 382 with $\gamma(1-1/\sigma)=0.45$ and values of 45.9 for "All" and 41.4 for "only residual wedge"; and sector 384 with $\gamma(1-1/\sigma)=0.65$ and values of 55.3 for "All" and 16.3 for "only residual wedge".

The three sectors for which $\gamma(1-\frac{1}{\sigma})>0.64$, reported in Column 3 of Table 5, present aggregate efficiency losses of 60% (AE=0.4). This is by contrast to a 37.6% overall aggregate loss (column 1) and a 27.3% loss in sectors where $\gamma(1-\frac{1}{\sigma})<0.31$[41]. It is both the case that, in

[41] On average $\gamma(1-\frac{1}{\sigma})=0.47$, see Table 1. The $\gamma(1-\frac{1}{\sigma})$ thresholds that separate groups in Table 5 reflect the fact that our estimate naturally cluster sectors in these three groups, with a large concentration of sectors between $\sigma=0.37$ and 0.6.

sectors with less curvature in the revenue function (i.e. to the right of the figure), composite wedges are more negatively correlated with $TFPQ\_HK$, and there is more dispersion in the terms of allocative efficiency involving $TFPR$ and $TFPQ\_HK$ (see the other panels of Figure 3).

Columns (5) and (6), in turn, show that idiosyncratic markups imply much larger efficiency losses in sectors with high markup variability. The overall efficiency loss is 26.6 percentage points larger in the four sectors with highest markup dispersion compared to the rest ($AE = 0.405$ in column (6) compared to $AE = 0.671$ in Column (5)). The efficiency loss from markup dispersion alone is close to 32% in those sectors ($AE = 0.679$ in column (6)), compared to less than 6% in the rest ($AE = 0.941$, column (5)).

Finally, we examine the value that heterogeneity in quality-adjusted productivity ($TFPQ\_HK$) has for aggregate productivity. We start from efficient productivity, TFP_t^{eff} (given by equation 19 in the absence of $tfpr$ dispersion) and evaluate its value against a counterfactual level that would arise if dispersion in $TFPQ\_HK$ were shut down. TFP^{eff} is a large 152% higher than what it would be in the absence of $TFPQ\_HK$ dispersion, most of this gain driven by dispersion in demand/quality/appeal, with a negligible impact of -7.7% from the dispersion of $TFPQ$. Quality/appeal/demand affects aggregate (quality-adjusted) productivity directly because consumers value it and indirectly because higher quality plants produce more to satisfy that demand. Meanwhile, higher technical efficiency at the plant-level has offsetting effects on plant revenue by increasing output directly, subsequently reducing prices. As a result, aggregate productivity is, in general, convex in dispersion in D but has a more ambiguous relationship with dispersion in technical efficiency ($TFPQ$).

7 Robustness to alternative estimates of production and demand function elasticities

One of our contributions is the design and implementation of a joint estimation method for the production function and the demand function, taking advantage of the availability of quantity and price data for both outputs and inputs. This use of the richness of the data makes the estimation consistent with the model, in that it takes fully into account the different dimensions of heterogeneity incorporated into the model and delivers estimates of both γ and σ. But, how important is this approach to identify the patterns that we find? Does it solve quantitatively significant biases that alternative estimation methods may be subject to? To answer the above question we implement usual alternative methods based on more limited sets of information to determine values for the parameter vector $\{\alpha, \beta, \phi, \gamma, \sigma\}$. The methods we implement are frequently used in the literature. Once we have estimated parameters under one of the alternative methods, we obtain estimates of $TFPQ\_HK$ and implement the decomposition of sales variability into the contributions of $TFPQ\_HK$ and $TFPR$, and our aggregate efficiency analysis, using the new estimates of parameters and plant attributes. Notice that, in absence of plant-level price information one would be unable to further decompose $TFPQ\_HK$ into its demand and technology components, and $TFPR$ into

input prices, markups, and residual wedges.[42] The comparison of results across alternative estimation methods sheds light on the value added of our joint estimation.

7.1 Alternative estimation methods

Databases used in the analysis of the role of distortions on the size distribution of businesses typically have information on a plant's revenue and input expenditures, but not on plants' output and input prices. Using this type of information, researchers estimate production elasticities from cost shares (e.g. HK), or through proxy methods using revenue deflated with an aggregate deflator as an –admittedly imperfect–measure of production. We attempt both approaches. As for the σ parameter, HK impose a value of 3, based on previous estimates, while others use alternative methods that we also try. We now briefly summarize the different alternatives we implement. Details are provided in Appendix M.

Cost Shares, CS: As in HK, we take factor elasticities equal to their respective cost shares, and impose $\sigma = 3$.

Proxy methods, ACF and DEU: We estimate the production function specified in equation (20) in the appendix, but using revenue as the dependent variable and materials costs rather than our internally-deflated materials. We implement a version that follows (Ackerberg, Caves and Frazer, 2015) and another following (De Loecker, Eeckhout and Unger, 2020), which we label, respectively, as ACF and DEU. The latter differs from the former because DEU explicitly recognize the potential price bias emerging from the use of revenue as dependent variable, which leads to a control function that includes market shares (see Appendix M). In both cases, we impose $\sigma = 3$.

Alternative joint estimation, KG: Blackwood et al. (2021) propose a way to use insights from Klette and Griliches (1996) to jointly estimate the production function parameters and σ. Using $P_{ft} = D_t d_{ft} Q_{ft}^{-\frac{1}{\sigma}}$ and its implication that $P_t = D_t Q_t^{-\frac{1}{\sigma}}$ we have that $\frac{P_{ft}}{P_t} = Q_t^{\frac{1}{\sigma}} Q_{ft}^{-\frac{1}{\sigma}} d_{ft}$. Thus, R_{ft} can be written:

$$R_{ft} = P_{ft}Q_{ft} = P_t Q_t^{\frac{1}{\sigma}} Q_{ft}^{1-\frac{1}{\sigma}} d_{ft} = P_t Q_t^{\frac{1}{\sigma}} \left(A_{ft} X_{ft}^{\gamma}\right)^{1-\frac{1}{\sigma}} d_{ft} \tag{29}$$

Based on this implication, we estimate the following version of the revenue function:

$$\ln R_{ft} = \overline{\alpha} \ln K_{ft} + \overline{\beta} \ln L_{ft} + \overline{\phi} \ln(Pm_{ft} * M_{ft}) + \frac{1}{\sigma} \ln E_t + \left(\left(1 - \frac{1}{\sigma}\right) (\ln A_{ft} + \ln P_t) + \ln d_{ft} \right) \tag{30}$$

where $\overline{\alpha} = \alpha\left(1 - \frac{1}{\sigma}\right)$, $\overline{\beta} = \beta\left(1 - \frac{1}{\sigma}\right)$, $\overline{\phi} = \phi\left(1 - \frac{1}{\sigma}\right)$, and $E_t = Q_t P_t$. The parameter that accompanies E_t allows us to estimate $\left(1 - \frac{1}{\sigma}\right)$ so that we can obtain the production elasticities by adjusting the estimated revenue elasticities correspondingly. Following Blackwood et al. (2021) we estimate 30 through proxy methods, using E_t or E_{t-1} to instrument ξ_{ft}^A.

Uniproduct: De Loecker et al. (2016) suggest the use of the sample of uniproduct plants as an alternative for the need to aggregate across products in multi-product units.

[42] By contrast, our framework inherits the characteristic of $HK's$ model that $TFPQ\_HK$ and $TFPR$ can be estimated using information on revenue and parameter estimates (equation 16).

We also estimate a version of our baseline estimation restricting the sample to uniproduct establishments.

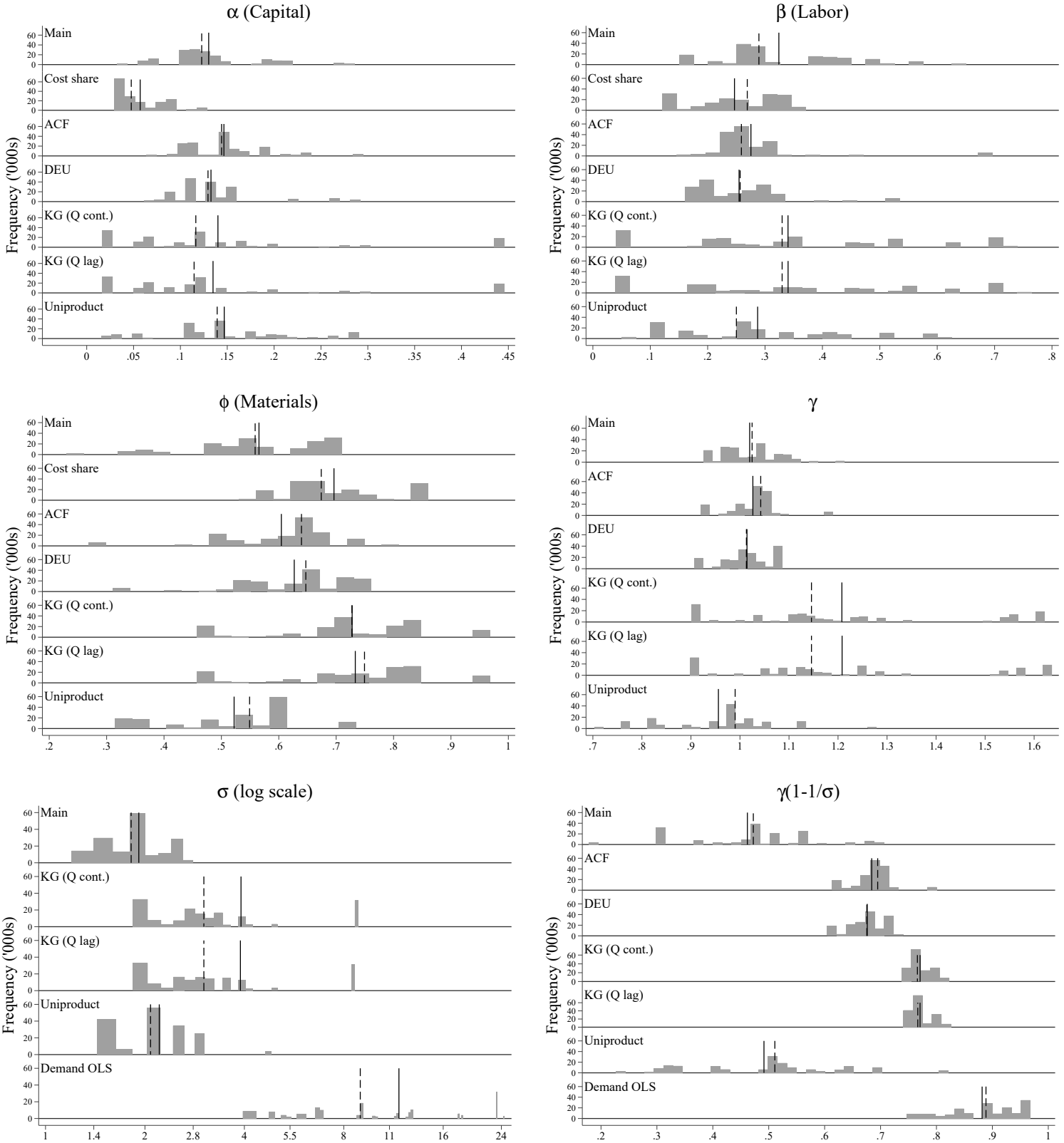
OLS demand estimation: To assess the importance of having access to production data to form our instrument for demand, we also carry an OLS estimation of demand function (26) to estimate σ. Such estimation takes advantage of the information on P_{ft} and Q_{ft} but ignores the information on input use that is taken advantage of in our baseline joint estimation to identify σ.

7.2 Parameter estimates

Figure 4 and the bottom panel of Table 7 describe the estimated parameters for our different alternative methods. Compared with our baseline estimation, cost share and proxy methods miss much of the variability in production technologies across sectors. In particular, beyond the fact that these methods do not estimate σ, and rather impose $\sigma = 3$ for all sectors, there is less dispersion in estimates of α, β and ϕ under these alternative methods vs. our baseline. Treating our baseline as truth for the purposes of this discussion, they also tend to overestimate (on average) the elasticity of production with respect to materials and underestimate that of labor. And, the practice of imputing $\sigma = 3$, used frequently in conjunction with these alternative production estimates, overestimates σ for all sectors in our sample (see also Table 1). As a result, alternative methods CS, ACF, and DEU significantly underestimate the concavity of the revenue function and its dispersion (last panel of Figure 4 and columns 2, 3 and 4 of Table 7, bottom panel, where $\gamma(1 - \frac{1}{\sigma})$ is much higher than in the baseline reported in column 1).

The KG method provides an attempt to estimate both production function and demand parameters using only information on revenue and input use, together with the structure of the model. Results display high dispersion, implausibly high in some cases. In fact, these estimations required externally imposing the restriction $\gamma(1-\frac{1}{\sigma}) < 1$ to ensure a non-explosive solution to the plant's problem. Even with that restriction, the mean estimated σ exceeds 3 in both versions of this approach, reaching numbers above $\sigma = 4$ for several sectors, with one sector even going above $\sigma = 8$. Production function parameters are overestimated, with returns to scale in production averaging well over 1, and in fact, exceeding 1 for most sectors. The degree of concavity of the revenue function is hugely underestimated in these approaches. $\gamma(1-\frac{1}{\sigma})$ exceeds 0.75 in KG, on average and for most sectors, compared with a value of 0.47 in the baseline (Figure 4 and columns 5 and 6 vs. 1 of Table 7).

Figure 4: Parameter Densities Under Alternative Estimation Methods



Note: Histogram weighted by the number of plants in each sector. Solid and dashed lines represent, respectively, the average and median of the distribution.

The estimation for uniproduct plants yields average parameter estimates not far from the average baseline results. But parameter estimates also exhibit more dispersion across sectors in the uniproduct estimation than the baseline scenario, likely a reflection of less precision derived from the loss of a large fraction of the sample.

Finally, the OLS estimation of the demand function yields biased estimates of the elasticity of substitution. By taking advantage of information on the use of inputs by suppliers, our baseline estimation method addresses the bias arising from supply-demand simultaneity in the price vs. quantity relationship.

7.3 Sales decomposition and aggregate productivity under alternative parameters

The curvature parameter of the revenue function $\gamma(1-\frac{1}{\sigma})$ is uniformly closer to one using alternative (traditional) estimation methods relative to our baseline. We know from our analysis above curvature is crucial in the measurement and quantification of the role of quality-adjusted productivity and wedges. Consistent with that perspective, the top panel of Table 7 shows that alternative methods tend to yield a higher (absolute value) contribution to sales variance of both quality-adjusted productivity and wedges. With curvature parameter closer to one, reconciling the observed dispersion in size in the data requires greater participation of both dimensions.[43] A related but distinct contributing factor is that alternative methods yield, as discussed above, less dispersion in factor elasticities.

From the analysis above, we know that greater dispersion in wedges holding other things equal dampens allocative efficiency. However, across columns of Table 7, other things are not held equal. For the alternative methods (excluding column 2), both the contribution of quality-adjusted productivity and wedges increase relative to the baseline as well as the loading parameters for these components into allocative efficiency (see equation (21)). Recall also that we find above that aggregate efficient productivity tends to be increasing in dispersion in quality-adjusted productivity. Given these offsetting forces, it is not surprising that the middle panel of Table 7 finds that some methods yield higher and some lower allocative efficiency relative to the baseline. More systematically, alternative methods yield a greater range of allocative efficiency across sectors (based on the gap between the max and min).

Our analysis has highlighted the tight connection between the determinants of the size distribution and the factors influencing allocative efficiency. Alternative estimation (traditional) methods yield greater dispersion in parameters across sectors, higher average curvature, a tendency for greater contribution to sales dispersion of both quality-adjusted productivity and wedges, and more dispersion in allocative efficiency across sectors. In addition, alternative methods based on revenue data by construction don't permit the decomposition of quality-adjusted productivity into its technical efficiency and quality/appeal components, as well as the decomposition of composite wedges into their input price, markup, and residual wedge components. Thus, the value-added of the price and quantity data for both outputs

[43] An exception is the second column that uses cost shares and $\sigma = 3$. This case stands out as having no dispersion in curvature across sectors as well as having an especially high materials output elasticity. The latter is important in this context since materials have volatility similar to sales dampening the dispersion of $TFPQ$ and thus the contribution of quality-adjusted productivity.

and inputs enables both internally consistent estimation and also a much richer decomposition of components driving the size distribution and allocative efficiency.

Table 7: Variance Decomposition of Sales and Allocative Efficiency Under Alternative Parameters

	Main	Cost shares, $\sigma = 3$	ACF , $\sigma = 3$	DEU, $\sigma = 3$	Klette-Griliches (Contemporary Q)	Klette-Griliches (Lagged Q)	Uniproduct	Cost shares, σ from OLS demand
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Variance Decomposition of Sales							
TFPQ-HK	1.139	1.121	1.242	1.219	1.243	1.241	1.172	1.639
Composite (HK) wedge	-0.139	-0.121	-0.242	-0.219	-0.243	-0.241	-0.172	-0.639
	Panel B: Aggregate Efficiency							
AE (aggregate)	0.624	0.692	0.647	0.648	0.534	0.539	0.612	0.679
AE (min)	0.335	0.312	0.135	0.333	0.015	0.015	0.279	0.256
AE (max)	0.772	0.909	0.900	0.891	0.892	0.945	0.813	1.149
AE (s.d.)	0.142	0.162	0.178	0.156	0.256	0.262	0.157	0.240
	Panel C: Parameter average and descriptive statistics (across sectors)							
$\gamma(1-1/\sigma)$ (min)	0.179	0.667	0.613	0.605	0.738	0.741	0.227	0.747
$\gamma(1-1/\sigma)$ (mean)	0.475	0.667	0.690	0.673	0.769	0.769	0.484	0.874
$\gamma(1-1/\sigma)$ (max)	0.693	0.667	0.787	0.724	0.813	0.813	0.819	0.959
σ	1.953	3.000	3.000	3.000	3.207	3.213	2.198	10.350
γ	1.028	1.000	1.036	1.010	1.220	1.220	0.960	1.000
ϕ (Materials)	0.520	0.676	0.599	0.612	0.702	0.709	0.512	0.676
β (Labor)	0.365	0.259	0.288	0.261	0.371	0.369	0.295	0.259
α (Capital)	0.143	0.065	0.148	0.137	0.148	0.142	0.154	0.065
sd(TFPQ-HK)	2.365	1.052	0.968	0.984	0.904	0.907	2.064	0.657
sd(TFPR)	0.512	0.407	0.404	0.401	0.419	0.422	0.540	0.407
Materials defator	CUPI	PPI	PPI	PPI	PPI	PPI	CUPI	PPI
σ_w	Main	Main	Main	Main	Main	Main	Uniproduct	Main

Note: Decomposition corresponds to the weighted average across ages up to and including age 50. We estimate a decomposition for each sector and then estimate weighted averages by sector revenue. $TFPQ\_HK$ values correspond to the sum of the contributions of D and $TFPQ$; Composite (HK) wedge wedge is the sum of the contributions of input prices, markups, and residual wedges. Column 2 estimates factor elasticities from cost shares and uses the often-used $\sigma = 3$. Column 3 implements an ACF estimation for the revenue function. Column 4 implements an estimation as in De Loecker, Eeckhout and Unger (2020) that uses an ACF estimation method for the revenue function with a control function including market shares. Columns 5 and 6 estimates a revenue function that combines the ACF and Klette and Griliches (1996) approach to obtain σ. Column 5 uses contemporary sectoral output while column 6 uses lagged sectoral output to identify the elasticity of substitution. Column 7 implements the baseline joint production and demand estimation restricting to plants that in all observed years produce the same single product. Column 8 estimates σ from an OLS demand equation.

8 Conclusion

Using a novel framework that integrates previous approaches taking advantage of rich data on both prices and quantities of outputs and inputs, we find evidence of a tight relationship between the determinants of the size distribution of activity and the determinants of aggregate productivity. The size distribution within industries is dominated by heterogeneity in quality-adjusted productivity with quality/appeal accounting for most of that variation but technical

efficiency playing an important supporting role. The size distribution of activity is compressed by wedges relative to that implied from the quality-adjusted productivity given the curvature of production and demand elasticities. A relatively mild compression of sales dispersion of 14% leads to a large loss in aggregation efficiency of 38%.

Much of the misallocation literature identifies only composite wedges but our framework and data permit decomposing them into input prices, markups, and residual wedges. Since the latter only account for about half of the composite wedges in our analysis, we have whittled down unexplained wedges considerably. Some of the literature has decomposed the variance of size across firms into demand and supply (marginal cost) components with the finding that demand accounts for almost of the variance. We have shown that the marginal cost component is effectively a composite of technical efficiency, input prices, and residual wedges, with the first working in the opposite direction of the latter two sub-components. Moreover, it remains unclear whether residual wedges are operating on the cost or demand side of variation. The implication is that there is more of a role to technical efficiency for variation in firm size than suggested by such demand vs. supply decompositions.

We show how determinants of the size distribution contribute to aggregate productivity. Dispersion in quality-adjusted productivity contributes positively mostly through the quality/appeal component, while the wedges are a drag on aggregate productivity. Idiosyncratic markups are a negligible factor in accounting for the size distribution but are quantitatively important as a drag on aggregate productivity. This distinct pattern emerges since markups are the highest for the most productive (and largest) plants. Our results provide guidance and perspective on the ongoing debate on the drag on productivity from markups.

Many open questions and areas for future research remain. Our findings on the role of input price heterogeneity (even adjusting for quality) point to important sources of such heterogeneity, including frictions in the markets for inputs as well as potentially monopsony power. Sorting this out should be an important topic for future research.

Our findings contribute to the policy discussion regarding interventions to address the limitations to business growth. Our results highlight that size-to-productivity wedges are important and especially prevalent for young businesses but also that dimensions internal to businesses are even more important than wedges to explain differential firm growth. On this internal side, the focus has frequently been on efforts conducive to improvements in technical efficiency. For instance, research on managerial practices that impact productivity has focused on production processes and employee management (e.g. Bloom and Reenen, 2007; Bloom et al., 2016) . Our approach highlights the multidimensional character of growth drivers that are internal to the business, including the appeal to customers and input prices potentially affected by its decisions. Our results align with those in Atkin, Khandelwal and Osman (2017, 2019) in pointing at quality as a crucial driver of business growth and at the fact that quality improvements may impose costs in terms of technical efficiency. Moreover, the results suggest that growth based on reducing barriers to quality differentiation is more conducive to welfare gains than that based on reducing dispersion in technical efficiency across businesses.

While we are able to attribute a large part of the role of HK composite wedges to input price and markup dispersion, our residual wedges are still a black box. Identifying the specific sources of wedges that dampen output and sales growth, especially for young plants, beyond input prices and markups that we analyze, is one area of research. One natural

candidate is adjustment costs that especially impact young businesses. These may include the costs of developing and accumulating organizational capital (such as the customer base). Our finding that between-plant differences in demand become more important in accounting for output growth volatility for more mature plants is consistent with this hypothesis. Also, the fact that we decompose quality-adjusted productivity into its technical efficiency and demand components yields guidance as to the potential source of wedges dampening growth.

Size-dependent policies and other characteristics of the regulatory environment are another set of candidate explanations behind our residual wedges, which we find to be highly negatively correlated with productivity, both in terms of efficiency and quality. Colombia is a country that underwent dramatic reforms over our sample period, some of them displaying cross-sectional variability (such as product-specific reductions to import tariffs in the early 1990s), and thus offers fruitful ground for investigating the impact of the regulatory environment on life-cycle dynamics. Future work that explored the relationship between regulatory and tariff reform and the evolution of the attributes and wedges we identify would be of interest.

Our findings provide insights into the relative importance of the variance in plant attributes valued by consumers (efficiency and quality) in explaining plant size, inviting further research into the ultimate sources of the variance in these attributes. While our framework allows for wedges that are correlated with current these attributes, and in fact we find that they are hugely negatively inversely correlated, we do not take explicit account of the endogenous response, to past performance and past wedges, of quality-adjusted quantity productivity and its components over the life cycle. Research that sheds light on the endogenous determinants of the variance in the supply side ($TFPQ$) and demand side (D) attributes of plants should have a high priority in future research. In an exploratory analysis shown in Appendix E we find evidence that $TFPQ$ and demand shocks are highly persistent and reflect indicators of endogenous innovation.

Another interesting area for future research is to link our findings regarding markups with other approaches where markups are estimated without resorting to assumptions regarding preferences. Recent analyses by De Loecker, Eeckhout and Unger (2020) present evidence of substantial dispersion in such markups across producers using an approach that is flexible on the structure of demand but that has the potential limitation of attributing to markups variation that may come from residual wedges or the structure of technology across producers. Our analysis using plant-level quality adjusted prices, while more restrictive in the sense of imposing a given demand structure, highlights challenges for pursuing this agenda. As we emphasize, even measuring plant-level output and inputs for multi-product plants that use a variety of inputs requires taking a stand on the demand structure. Tackling technology and markup heterogeneity in this multi-product, multi-input environment with ongoing quality change will be a challenge.

Data Availability Statement

The data used in this paper cannot be shared publicly, as the Colombian Annual Manufacturing Survey and the Technological Development and Innovation Survey are housed at the Colombian Departamento Administrativo Nacional de Estadística. Instructions on how to

request access, along with all replication programs and detailed explanations of data construction, are available at the following DOI: `https://dx.doi.org/10.5281/zenodo.7604117`.

References

Acemoglu, Daron, Ufuk Akcigit, Harun Alp, Nicholas Bloom, and William Kerr. 2018. "Innovation, Reallocation, and Growth." *American Economic Review*, 108(11): 3450–91.

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

Adamopoulos, Tasso, and Diego Restuccia. 2014. "The Size Distribution of Farms and International Productivity Differences." *American Economic Review*, 104(6): 1667–97.

Asker, John, Allan Collard-Wexler, and Jan De Loecker. 2014. "Dynamic Inputs and Resource (Mis)Allocation." *Journal of Political Economy*, 122(5): 1013–1063.

Atkeson, Andrew, and Ariel Tomás Burstein. 2010. "Innovation, Firm Dynamics, and International Trade." *Journal of Political Economy*, 118(3): 433–484.

Atkin, David, Amit K. Khandelwal, and Adam Osman. 2017. "Exporting and Firm Performance: Evidence from a Randomized Experiment*." *The Quarterly Journal of Economics*, 132(2): 551–615.

Atkin, David, Amit K. Khandelwal, and Adam Osman. 2019. "Measuring Productivity: Lessons from Tailored Surveys and Productivity Benchmarking." *AEA Papers and Proceedings*, 109: 444–49.

Aw, Bee Yan, Mark J. Roberts, and Daniel Yi Xu. 2011. "R&D Investment, Exporting, and Productivity Dynamics." *American Economic Review*, 101(4): 1312–44.

Baqaee, David Rezza, and Emmanuel Farhi. 2020. "Productivity and Misallocation in General Equilibrium*." *The Quarterly Journal of Economics*, 135(1): 105–163.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta. 2013. "Cross-Country Differences in Productivity: The Role of Allocation and Selection." *American Economic Review*, 103(1): 305–34.

Bento, Pedro, and Diego Restuccia. 2017. "Misallocation, Establishment Size, and Productivity." *American Economic Journal: Macroeconomics*, 9(3): 267–303.

Blackwood, G. Jacob, Lucia S. Foster, Cheryl A. Grim, John Haltiwanger, and Zoltan Wolf. 2021. "Macro and Micro Dynamics of Productivity: From Devilish Details to Insights." *American Economic Journal: Macroeconomics*, 13(3): 142–72.

Bloom, Nicholas, and John Van Reenen. 2007. "Measuring and Explaining Management Practices across Firms and Countries." *The Quarterly Journal of Economics*, 122(4): 1351–1408.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, Daniela Scur, and John Van Reenen. 2016. "International Data on Measuring Management Practices." *American Economic Review*, 106(5): 152–56.

Broda, Christian, and David E. Weinstein. 2006. "Globalization and the Gains From Variety*." *The Quarterly Journal of Economics*, 121(2): 541–585.

Broda, Christian, and David E. Weinstein. 2010. "Product Creation and Destruction: Evidence and Price Implications." *American Economic Review*, 100(3): 691–723.

Brooks, Eileen L. 2006. "Why don't firms export more? Product quality and Colombian plants." *Journal of Development Economics*, 80(1): 160–178.

Burdett, Kenneth, and Dale Mortensen. 1998. "Wage differentials, employer size, and unemployment." *International Economic Review*, 39(2): 257–273.

Caballero, Ricardo J., Eduardo M. R. A. Engel, and John C. Haltiwanger. 1995. "Plant-Level Adjustment and Aggregate Investment Dynamics." *Brookings Papers on Economic Activity*, 26(2): 1–54.

Caballero, Ricardo J, Eduardo M R A Engel, and John Haltiwanger. 1997. "Aggregate Employment Dynamics: Building from Microeconomic Evidence." *American Economic Review*, 87(1): 115–137.

DANE. 1982-2013. "Encuesta Anual Manufacturera, EAM (AMS)." Departamento Administrativo Nacional de Estadística, DANE.

David, Joel M., and Venky Venkateswaran. 2019. "The Sources of Capital Misallocation." *American Economic Review*, 109(7): 2531–67.

Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda. 2020. "Changing Business Dynamism and Productivity: Shocks versus Responsiveness." *American Economic Review*, 110(12): 3952–90.

De Loecker, Jan, and Frederic Warzynski. 2012. "Markups and Firm-Level Export Status." *American Economic Review*, 102(6): 2437–71.

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

De Loecker, Jan, Jan Eeckhout, and Simon Mongey. 2021. "Quantifying Market Power and Business Dynamism in the Macroeconomy." National Bureau of Economic Research Working Paper 28761.

De Loecker, Jan, Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik. 2016. "Prices, Markups, and Trade Reform." *Econometrica*, 84(2): 445–510.

de Roux, Nicolás, Marcela Eslava, Santiago Franco, and Eric Verhoogen. 2021. "Estimating Production Functions in Differentiated-Product Industries with Quantity Information and External Instruments." National Bureau of Economic Research Working Paper 28323.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu. 2018. "How Costly Are Markups?" National Bureau of Economic Research Working Paper 24800.

Eslava, Marcela, John Haltiwanger, Adriana Kugler, and Maurice Kugler. 2004. "The effects of structural reforms on productivity and profitability enhancing reallocation: evidence from Colombia." *Journal of Development Economics*, 75(2): 333–371. 15th Inter American Seminar on Economics.

Eslava, Marcela, John Haltiwanger, Adriana Kugler, and Maurice Kugler. 2010. "Factor Adjustments After Deregulation: Panel Evidence From Colombian Plants." *The Review of Economics and Statistics*, 92(2): 378–391.

Eslava, Marcela, John Haltiwanger, Adriana Kugler, and Maurice Kugler. 2013. "Trade and market selection: Evidence from manufacturing plants in Colombia." *Review of Economic Dynamics*, 16(1): 135–158. Special issue: Misallocation and Productivity.

Eslava, Marcela, John Haltiwanger, and Álvaro Pinzón. 2022. "Job Creation in Colombia Versus the USA: 'Up-or-out Dynamics' Meet 'The Life Cycle of Plants'." *Economica*, 89(355): 511–539.

Feenstra, Robert C. 1994. "New Product Varieties and the Measurement of International Prices." *The American Economic Review*, 84(1): 157–177.

Fieler, Ana Cecília, Marcela Eslava, and Daniel Yi Xu. 2018. "Trade, Quality Upgrading, and Input Linkages: Theory and Evidence from Colombia." *American Economic Review*, 108(1): 109–46.

Forlani, Emanuele, Ralf Martin, Giordano Mion, and Mirabelle Muûls. 2021. "Unraveling firms: Demand, productivity and markups heterogeneity." Unpublished manuscript.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review*, 98(1): 394–425.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2016. "The Slow Growth of New Plants: Learning about Demand?" *Economica*, 83(329): 91–129.

García-Santana, Manuel, and Josep Pijoan-Mas. 2014. "The reservation laws in India and the misallocation of production factors." *Journal of Monetary Economics*, 66: 193–209.

Garicano, Luis, Claire Lelarge, and John Van Reenen. 2016. "Firm Size Distortions and the Productivity Distribution: Evidence from France." *American Economic Review*, 106(11): 3439–79.

Gopinath, Gita, Şebnem Kalemli-Özcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez. 2017. "Capital Allocation and Productivity in South Europe*." *The Quarterly Journal of Economics*, 132(4): 1915–1967.

Guner, Nezih, Gustavo Ventura, and Yi Xu. 2008. "Macroeconomic implications of size-dependent policies." *Review of Economic Dynamics*, 11(4): 721–744.

Hallak, Juan Carlos, and Peter K. Schott. 2011. "Estimating Cross-Country Differences in Product Quality*." *The Quarterly Journal of Economics*, 126(1): 417–474.

Haltiwanger, John, Robert Kulick, and Chad Syverson. 2018. "Misallocation Measures: The Distortion That Ate the Residual." National Bureau of Economic Research, Inc NBER Working Papers 24199.

Hopenhayn, Hugo. 2016. "Firm Size and Development." *Economia Journal*, 0(Fall 2016): 27–49.

Hottman, Colin J., Stephen J. Redding, and David E. Weinstein. 2016. "Quantifying the Sources of Firm Heterogeneity." *The Quarterly Journal of Economics*, 131(3): 1291–1364.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India*." *The Quarterly Journal of Economics*, 124(4): 1403–1448.

Hsieh, Chang-Tai, and Peter J. Klenow. 2014. " The Life Cycle of Plants in India and Mexico *." *The Quarterly Journal of Economics*, 129(3): 1035–1084.

Jaumandreu, Jordi, and Jacques Mairesse. 2010. "Innovation and Welfare: Results from Joint Estimation of Production and Demand Functions." National Bureau of Economic Research Working Paper 16221.

Khandelwal, Amit. 2010. "The Long and Short (of) Quality Ladders." *The Review of Economic Studies*, 77(4): 1450–1476.

Klette, Tor Jakob, and Zvi Griliches. 1996. "The Inconsistency of Common Scale Estimators When Output Prices are Unobserved and Endogenous." *Journal of Applied Econometrics*, 11(4): 343–361.

Kugler, Maurice, and Eric Verhoogen. 2011. "Prices, Plant Size, and Product Quality." *The Review of Economic Studies*, 79(1): 307–339.

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

Manova, Kalina, and Zhiwei Zhang. 2012. "Export Prices Across Firms and Destinations." *The Quarterly Journal of Economics*, 127(1): 379–436.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica*, 71(6): 1695–1725.

Midrigan, Virgiliu, and Daniel Yi Xu. 2014. "Finance and Misallocation: Evidence from Plant-Level Data." *American Economic Review*, 104(2): 422–58.

Mortensen, Dale, and Christopher Pissarides. 1994. "Job Creation and Destruction in the Theory of Unemployment." *Review of Economic Studies*, 61(3): 397–415.

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

Redding, Stephen J, and David E Weinstein. 2020. "Measuring Aggregate Price Indices with Taste Shocks: Theory and Evidence for CES Preferences*." *The Quarterly Journal of Economics*, 135(1): 503–560.

Restuccia, Diego, and Richard Rogerson. 2008. "Policy distortions and aggregate productivity with heterogeneous establishments." *Review of Economic Dynamics*, 11(4): 707–720.

Restuccia, Diego, and Richard Rogerson. 2017. "The Causes and Costs of Misallocation." *Journal of Economic Perspectives*, 31(3): 151–74.

Sato, Kazuo. 1976. "The Ideal Log-Change Index Number." *The Review of Economics and Statistics*, 58(2): 223–228.

Vartia, Yrjö O. 1976. "Ideal Log-Change Index Numbers." *Scandinavian Journal of Statistics*, 3(3): 121–126.