NBER WORKING PAPER SERIES

AN INDUSTRIAL ORGANIZATION PERSPECTIVE ON PRODUCTIVITY

Jan De Loecker
Chad Syverson

Working Paper 29229
http://www.nber.org/papers/w29229

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2021

Chapter prepared for the Handbook of Industrial Organization (Vol IV). We thank the editors and three referees for their comments and suggestions, as well as Dan Ackerberg, Tim Bresnahan, Penny Goldberg, John Haltiwanger, Marc Melitz, Ariel Pakes, Amil Petrin, Michael Rubens, and Jo Van Biesebroeck. We also thank our respective co-authors, from whom we have learned a great deal and continue to do so. Finally, while this is our first formal collaboration in terms of written work, it has been our pleasure to have been able to call on our many research and teaching interactions over the years to build this chapter together. We thank Filippo Biondi, Nadine Hahn, and Luis Moyano Garcia for excellent research assistance. De Loecker acknowledges support from the ERC Consolidator Grant 816638. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

An Industrial Organization Perspective on Productivity
Jan De Loecker and Chad Syverson
NBER Working Paper No. 29229
September 2021
JEL No. D2,L1,L2,L6

ABSTRACT

This chapter overviews productivity research from an industrial organization perspective. We focus on what is known and what still needs to be learned about the productivity levels and dynamics of individual producers, but also how these interact through markets and industries to influence productivity aggregates. We overview productivity concepts, facts, data, measurement, analysis, and open questions.

Jan De Loecker
Economics Department
KU Leuven
Naamsestraat 68
3000 Leuven
Belgium
and CEPR
and also NBER
jan.deloecker@kuleuven.be

Chad Syverson
University of Chicago
Booth School of Business
5807 S. Woodlawn Ave.
Chicago, IL 60637
and NBER
chad.syverson@chicagobooth.edu

Contents

1 A Productivity Primer

Productivity is a central concept in economics. Its application spans a wide range of fields and is not housed within a particular discipline or approach. A telling observation is that the National Bureau of Economic Research (NBER) lists *Productivity, Innovation, and Entrepreneurship* as a research program separate from, among others, its *Industrial Organization* program. This highlights that while it is not treated as a separate field in graduate education, it stands distinct in substantive focus.

The scope of productivity analysis is broad. It spans from micro to macro, ranging from the analysis of individual production lines in a factory to the study of economy-wide aggregates. It treats these different scopes not just as separate objects but minds their connections as well, studying the efficiency of industry output allocation across heterogeneous producers and its relationship to changes in the operating environment (such as technical change, antitrust, or trade policy). Productivity analysis also includes the study of both the efficiency of production and technological change.

This subject matter breadth is accompanied by a methodological flexibility as well, with the literature applying a variety of approaches and methods to address these research questions.

This chapter focuses on the implications and applications of productivity analysis within Industrial Organization. There is a long tradition in IO of studying productivity-related topics like allocative efficiency, technological change, regulatory effects, cost efficiencies in merger analysis, and returns to scale, to name a few. The term productivity is, however, more often than not used in a fairly loose sense, usually referring to a measure of performance. In this chapter we make an explicit distinction between productivity in a strict production efficiency sense, on the one hand, and performance on the other. The study of production efficiency, the rate at which a producer can convert a bundle of inputs into a unit of output, is in essence about the technical relationship between output and inputs. Performance captures a variety of measures, but as this chapter will highlight, it is intimately related to efficiency. However, the distinction can be crucial when analyzing the very topics listed above.

1.1 Background and Focus

While having a long history in economics, the past few decades have seen the productivity analysis of individual producers and the corresponding industry- and economy-wide ag-

gregates become a central topic in both academic and policy circles. This renewed interest has paralleled at least three main developments.

First, over the last two decades the access to micro data has exploded. At the turn of the century, only a few large-scale producer-level datasets existed, with limited access to researchers. In contrast, the current list of countries for which micro census data is available (in manufacturing, at least) contains a rather large share of the world. In addition, private data providers have emerged offering comprehensive accounting data capturing typical variables used in productivity analysis, through the reporting of balance sheets and income and loss statements.

Second, accompanying the increased access to micro data has been a renewed interest in the estimation and identification of production functions. These are of course key objects of interest for most productivity analyses, both for their own sake as well as supply-side inputs into equilibrium analysis. This research has focused on obtaining reliable productivity measures for sets of producers when, as is the case, the researcher cannot directly observe productivity but producers can. This leads to two well-known biases, the simultaneity and selection biases, that researchers must face.

Third is the prominent role of productivity analysis in forming and executing economic policy. While policymakers still mostly focus on industry- or economy-wide aggregates, there is increasing recognition that this analysis is often most informative when built from the ground up using micro data. The melding of data, methods, and economically oriented policy analysis has spurred informative interactions among microeconomists, macroeconomists, and policymakers that have created many insights into productivity.

Our intent in this chapter is to organize and review the intellectual underlayment of this burgeoning literature. There are many facets. Our coverage includes key conceptual issues, facts about micro-level productivity, models of markets with heterogeneous-productivity producers, measurement and data, productivity estimation, the positive and normative implications of the static and dynamic allocation of activity across heterogeneous producers, and an overview of what we expect to be active areas of work in the near future. We do this while taking stock of several decades of empirical work on productivity using micro data. We will unavoidably miss certain dimensions, and simple space constraints mean we cannot do justice to many contributions and insights from this extensive literature. We do hope, however, to offer a structured view on the field of productivity and how Industrial Organization scholars have contributed.

1.2 Productivity Conceptualized

Productivity is conceptualized in a number of related ways. All productivity metrics in one form or another measure how much output producers obtain from a given set of inputs. As such they are measures of the efficacy of the supply side of the economy (though "efficacy" need not always be synonymous with social welfare).

An interpretation of productivity as an economic primitive is as a factor-neutral (aka Hicks-neutral) shifter of the production function.[1] Consider the general production function $Q = \Omega F(.)$, where Q is output and $F(\cdot)$ is a function of observable inputs capital such as capital, labor and intermediate inputs. Ω is productivity, the factor-neutral shifter.[2] It reflects variations in output not explained by shifts in the observable inputs that act through $F(\cdot)$. A higher value of Ω implies the producer will obtain more output from a given set of inputs. That is, it denotes a shift in the production function's isoquants down and to the left.

A second conceptualization is empirical: productivity as a ratio of output to inputs. This is tightly related to the production-function-shifter interpretation above. This can be seen by isolating productivity from the production function: $\Omega = \frac{Q}{F(.)}$. A is clearly an output-to-input ratio. Here, where output is divided by a combination of observable inputs, the productivity concept is named total factor productivity (TFP) (it is also sometimes called multifactor productivity, MFP). There are also single-factor productivity measures, where output is divided by the amount of a single input, most commonly labor; i.e., labor productivity. Because single-factor productivity measures can be affected not just by shifts in TFP but factor intensity decisions as well, TFP measures are often conceptually preferable. On the other hand, labor productivity is often easier to measure than TFP.

A third conceptualization is of productivity as a shifter of the producer's cost curve. Higher productivity shifts down the cost curve; that is, at the producer's cost-minimizing combination of inputs, its total cost of producing a given quantity is lower, the higher is its TFP level. This productivity conceptualization is related to the other two because the cost function is the value function of the producer's cost minimization problem, which takes the production function as its constraint. This cost function shifts down when

[1] Throughout we consider this predominant setup, of Hicks-neutral productivity, unless we explicitly discuss departures.

[2] We define the specific factors of production in more detail later on in the Chapter.

productivity rises.

Because it plays such an important role in the producer's production technology, measuring productivity and its influence on outcomes is the subject of an enormous literature. This work has found in many disparate settings that, as an empirical matter, productivity is hugely important in explaining the fortunes of producers, their workers, their suppliers, and their customers. We survey work on both the measurement and effects of productivity below.

2 Empirical Facts about Productivity at the Producer Level

A tremendous amount of empirical research during the past three decades has documented many empirical regularities about the levels and growth rates of productivity among producers. In this section, we summarize some of the best-established findings of that literature.

The typical unit of analysis (the "producer") depends on the particulars of the available data. Usually, it is a firm or establishment. An establishment is a geographically unique location of economic activity, be it a plant, warehouse, office, store, mine, and so on. One firm can own many establishments. A smaller body of work has explored production at an even more disaggregate level, such as the production line or work team. Regardless, the empirical patterns discussed below have been found to hold at any of these micro-level scopes of analysis.

2.1 Dispersion

One of the most ubiquitous findings in the literature is the enormous dispersion in productivity levels across producers, even within narrowly defined markets. For example, the typical 6-digit NAICS manufacturing industry in the U.S. or Canada has a 90-10 percentile TFP ratio of roughly 2:1. That is, the typical manufacturing industry (at this level of disaggregation, an industry is something like Metal Can Manufacturing) has a producer that can obtain twice the output from the same set of inputs of another operating in the same narrowly-defined industry. Note that this measure of dispersion ignores the extreme deciles of the distribution.

Nothing is particularly unusual about North American manufacturers when it comes

to productivity dispersion. Indeed, if anything the productivity dispersion among them is relatively small compared to other settings. Researchers have found 90-10 TFP ratios of 3:1 or more in manufacturing industries in developing and emerging economies. Others have documented higher ratios than this for various industries in other sectors in developed economies, ranging from business services to healthcare. Simply put, businesses differ greatly in the proficiency with which they convert inputs to outputs.

A reasonable first reaction to this enormous productivity dispersion is that it reflects the exclusion from the production function of inputs other than the standard labor, capital, and intermediate inputs. In some sense, this is right. Setting aside measurement error, observed output is made from and by *something*, and in a world of infinite data, one should be able to include those inputs in the production function. However, explaining the observed productivity dispersion is not just a matter of gathering a bit more data on some key non-standard inputs. Researchers in many empirical settings have added additional inputs and suspected productivity drivers to their production functions without substantially reducing measured productivity dispersion. For example, the early work of Griliches and its subsequent extensions included R&D expenditures or constructed R&D capital stocks without affecting measured dispersion. In the context of labor inputs, Fox and Smeets (2011) use highly detailed matched employer-employee data to build a multidimensional Griliches-type human capital measure of firms' labor inputs. This flexibly accounted for heterogeneity in worker skills and experience. Nevertheless, they too found no effect of including this more detailed input measure on estimated productivity dispersion.[3]

2.2 Persistence within Producers

Another documented empirical reality is that producers' productivity levels are persistent. Producers near the top end of their industry distribution in this period are likely to be near the top in the next period. Low-productivity producers are similarly likely to stay that way. This persistence rules out classical measurement error or true-but-white-noise

[3]We discuss extensively below an additional complication that arises in most producer-level data. Namely, output and input measures may be constructed from revenue and expenditures rather than quantities. This introduces price variation into the measurement of productivity. This variation can be driven by influences beyond just the production technology, such as market power or variations in local market conditions. The literature recognizes this issue and, as we detail below, treats it with combinations of theory and measurement. The facts described in this section have been found to apply generally, both to measures of productivity obtained using "pure" quantity data and more expenditure-based metrics.

productivity process as sources of the documented productivity dispersion. Whatever factors influence producers' productivity levels, they have staying power.

The literature has quantified the persistence in multiple ways. Some studies construct quantile-to-quantile transition matrices. These find that elements along the diagonal are consistently the largest, and they fall in size with distance from the diagonal. Other work estimates AR(1) specifications on producers' productivity levels in panel data. Typical coefficients might be 0.7-0.8 in annual panel data without producer fixed effects and 0.3-0.4 with fixed effects.

2.3 Correlations

Other variables have been found to be consistently related to productivity. To list some of the more common, multiple studies have found higher-productivity producers are:

- more profitable,
- larger,
- faster-growing,
- more likely to survive (this is perhaps the most ubiquitously documented pattern of those listed here),
- low price-setters (when the industry product is relatively homogeneous),
- higher-wage payers.

Other correlations have been measured in individual studies, but those listed here have the broadest empirical support.

3 A Simple Model of Equilibrium Productivity Dispersion

Canonical free entry models do not predict the persistent productivity differences observed in the data. In absence of heterogeneous fixed production factors, only firms that can operate at the lowest possible cost can operate in long-run equilibrium (that is, the long-run industry supply curve is horizontal). Given the empirical ubiquity of sustained productivity differences, researchers have built a class of commonly applied models to account for these and other associated empirical patterns. We sketch out here an example

of such a model, a simplified version of Melitz and Ottaviano (2008). We do so for two primary reasons. One, even this simple model offers several predictions that are consistent with, and can help explain, the basic mechanisms underlying the empirical regularities described in the previous section. Straightforward extensions of the model to more dynamic versions within its class imply still others. Two, the model's framework will be useful in facilitating discussions of various issues throughout the remainder of the chapter.

3.1 Demand

An industry comprises a continuum of producers, each of whom makes a single differentiated variety of the industry good. Both producers and varieties are indexed by i, and I is the set of industry producers/varieties. Demand for the industry's varieties is given by the representative industry consumer's preferences:

$$U = y + \alpha \int_{i \in I} Q_i \, di \; - \frac{1}{2}\eta \left(\int_{i \in I} Q_i \, di \right)^2 - \frac{1}{2}\gamma \int_{i \in I} Q_i^2 \, di \tag{1}$$

where y is the quantity of a numeraire good, Q_i is the quantity of variety i consumed, and $\alpha > 0$, $\eta > 0$, and $\gamma \geq 0$.

Utility is quadratic in total consumption of the industry's output, minus a term that increases in the variance of consumption across varieties. This last term builds curvature into the utility function by introducing an incentive to equate consumption levels of different varieties. The parameter γ summarizes substitutability. If γ is large, substitutability is low. Consumers want to limit idiosyncratic variation in their quantities of particular varieties, so consumption is relatively insensitive to any price differences. As $\gamma \to 0$, only the total quantity of industry varieties determines utility and substitutability becomes perfect. One can interpret γ as either a direct preference parameter or a reduced-form stand-in for other substitution barriers (unmodeled here but in other studies made more explicit) like trade costs, transport costs, or search costs.

3.2 Supply

Producers have a simple linear production technology $Q_i = \Omega_i X_i$, where Ω_i is productivity and X_i is a composite input with price P^X. As a result, producers have constant marginal costs $c_i = \frac{P^X}{\Omega_i}$. Thus cost variation reflects productivity variation.

No company will produce if its cost is too high to make a profit. Call the cost at which a firm earns zero profits c_D and the corresponding zero-profit productivity level $c_D = \frac{P^X}{\Omega_D}$.

Given this and demand as described above, Melitz and Ottaviano (2008) show a firm's revenues and operating profits are given by

$$r(c_i) = \frac{1}{4\gamma}\left(c_D^2 - c_i^2\right) = \frac{P^{X\,2}}{4\gamma}\left(\frac{1}{\Omega_D^2} - \frac{1}{\Omega_i^2}\right) \tag{2}$$

$$\pi(c_i) = \frac{1}{4\gamma}\left(c_D - c_i\right)^2 = \frac{P^{X\,2}}{4\gamma}\left(\frac{1}{\Omega_D} - \frac{1}{\Omega_i}\right)^2 \tag{3}$$

Entry into the industry is determined as follows. A pool of ex-ante identical potential entrants decides whether to pay a sunk entry cost f_E to take a productivity draw from a distribution $G(\Omega) = 1 - \frac{\Omega_M}{\Omega}, \quad \Omega \in [\Omega_M, \infty)$. This distribution is chosen for analytical convenience, as we will work with marginal costs and their distribution below. This productivity distribution implies that marginal costs are distributed uniformly between 0 and an upper bound $c_M = \frac{P^X}{\Omega_M}$.

If an entrant chooses to receive a draw Ω_i, it observes the draw and then determines whether to begin production and earn the corresponding operating profits as above. Because only potential entrants receiving draws $\Omega_i > \Omega_D$ will choose to produce in equilibrium, the gross expected value of paying f_E for a productivity draw is the expected operating profit conditional on $\Omega_i > \Omega_D$.

3.3 Equilibrium

The free-entry condition imposes that the gross expected value of paying for a productivity draw equals f_E. Changing variables to work with analytically more convenient marginal cost distribution, we can write this condition as follows:

$$\int_0^{c_D} \frac{1}{4\gamma}(c_D - c)^2 \frac{1}{c_M}\, dc = f_E \tag{4}$$

This condition pins down the equilibrium zero-profit cost draw c_D. Solving gives

$$c_D = (12\gamma c_M f_E)^{\frac{1}{3}} \tag{5}$$

which relates c_D (and hence cutoff productivity level Ω_D) to the model's primitives.

3.4 Empirical Implications

We summarize some empirical implications of the simple model above that have been documented in the literature. Other empirical findings can be linked to other, more complex models in its class, though we will not discuss them in detail here.

Productivity Dispersion There is equilibrium productivity dispersion. As long as $\gamma > 0$ and $f_E > 0$, the marginal cost and productivity distributions are nondegenerate. That is, if substitutability across product varieties is not perfect and there are fixed costs of setting up a business (or more strictly interpreted, learning what one's operating cost will be), industry producers will exhibit a range of productivity levels in equilibrium.

The productivity distribution is a truncation of the underlying distribution $G(\Omega)$. How much dispersion exists in equilibrium depends on the magnitudes of γ and f_E. As substitutability falls (γ rises), consumers are less willing or able to shift their purchases from one variety to another. This makes it easier for higher-cost producers to profitably operate. As seen in the revenue and profit functions above, the negative effect on firms' sales and profits is smaller when γ is large, raising Ω_D, the threshold productivity draw required for profitable operations.

A higher sunk entry cost f_E also supports greater productivity dispersion, though through a different mechanism than substitutability. Entry costs "protect" producers with low productivity draws from competition by limiting the mass of producers that take entry draws, reducing Ω_D.[4]

Productivity Persistence Producers' productivity levels are persistent. Here, this is trivially so, as the model is static and all producers have fixed draws. However, models with similar structures that allow more dynamic productivity process (e.g., Asplund and Nocke (2006)) nevertheless imply positive correlation in a given producer's productivity levels over time, while also preserving equilibrium productivity dispersion (and in some cases, even a stationary industry-level productivity distribution).

Profitability Higher-productivity producers are more profitable, as the profit expression above indicates.

Size and Growth Size is correlated with productivity. Revenues rise as marginal costs fall. Equilibrium quantities sold also decline in the producer's marginal cost and are given by

$$q(c_i) = \frac{1}{2\gamma}(c_D - c_i) \tag{6}$$

[4]If there is also a fixed *operating* cost in the model, this acts in a direction opposite of a fixed entry cost. The intuition is that this fixed cost is imposed after the firm knows its productivity draw. Higher fixed operating costs make it more difficult for low-productivity producers to be profitable, thereby raising Ω_D.

Producer size as measured through input use is somewhat more ambiguous because higher productivity reduces required input purchases for any given quantity. Here, however, the elasticity of quantity with respect to productivity is greater than one, and the production technology implies a negative unit elasticity of input to productivity. The combination of these implies that an increase in productivity implies net growth in input use. Thus higher-productivity producers are also larger even when measured by input use.

In this static model, producers' sizes are fixed, just like their productivity levels. However, it is clear from the size-productivity comparative static that dynamic versions of such models imply a correlation between the *growth rates* of size and productivity. Furthermore, if there are adjustment costs in factor or output markets, growth be correlated with productivity levels as well, as it can take time for firms experiencing a productivity innovation to adjust to the equilibrium size implied by their new productivity level. More explicit treatments of the dynamics of productivity can be found in the seminal articles of Jovanovic (1982) and Ericson and Pakes (1995).

Survival Productivity is correlated with survival. Firms receiving a productivity draw too low to be profitable in equilibrium do not produce. In the static model above, this "exit" decision occurs in the pre-production interregnum. If we imagine that this stands in for the reality of an early period of operation for young but not fully established producers, this is consistent with the empirical pattern of high exit rates among inefficient, young producers. That said, it is worth noting that the negative correlation between productivity and exit also holds even conditional on age in the data. This conditional relationship is captured by more explicitly dynamic versions of the model above.

The correlations between productivity, size, growth, and survival play an important role in motivating the burgeoning productivity research areas on allocation, reallocation, and misallocation. Allocation regards the relationship between productivity variation and the spread of economic activity across industry producers. For example, it answers questions like, what is the size of the productivity-output covariance? Reallocation explores how the productivity-activity relationship changes in response to shifts in industry primitives. Do changes in substitutability arising from reductions in transport costs increase the productivity-activity correlation, for instance? Misallocation asks the normative question of whether the observed productivity-activity relationship is welfare maximizing. We will discuss these topics in greater detail below.

Prices Producers with higher productivity charge lower prices. This is apparent in the expression for equilibrium prices:

$$P(c_i) = \frac{1}{2\gamma}(c_D + c_i) \tag{7}$$

Lower marginal costs imply lower prices. Here, the pass-through rate is 0.5 because of the linear demand system, but the directional result holds as long as the demand system implies positive pass through (an empirical regularity and a feature of commonly used demand systems). Things can become more complicated in reality, when products are vertically differentiated and productivity regards being able to make a higher quality product at a given cost. In this case, if equilibrium price is correlated with quality, price and productivity can be positively correlated. See Kugler and Verhoogen (2012) for example.

Wages The positive empirical correlation between wages and productivity is not embodied in this model nor its dynamic extensions. However, it could be accommodated with some effort. Empirical work has established two likely mechanisms for this correlation in the data (with opposite directions of implied causation, interestingly). One is based on the idea that unit labor inputs, however measured, differ in quality. These quality differences are reflected in their marginal products. In competitive labor markets, the equilibrium wage rises in labor quality. This allows firms to raise their productivity level (though perhaps not their profitability level, or at least not as much) by hiring higher quality labor at a wage premium. The other mechanism involves rent sharing between firms and their workers. Firms that experience positive innovations to productivity and profitability pass along some of the resulting profit gains to their workers in the form of higher pay.

4 Measurement of Output and Inputs

Given the empirical conceptualization above, all approaches build productivity metrics that are some form of output-to-input ratio. A number of measurement issues arise in this process. We overview these here. Measurement problems are especially important to keep in mind when working with productivity values. Because productivity is the output variation that cannot be explained by observable inputs, it is essentially a residual. As with all residuals, and as Abramovitz (1956) famously puts it, productivity is a measure of

our ignorance. Productivity metrics can end up embodying (to varying degrees, depending on the setting and the quality of the data) sources of output variation, like market power and measurement error, that are conceptually distinct from a production function.

4.1 Output Measurement

Producing output is what firms are built to do, but measuring that output for the sake of productivity measurement involves several potential concerns.

Sometimes the difficulty is quite fundamental: Defining what a producer's output actually is. What is the output of a bank, insurance company, or other firms in the financial sector? What do they "make"? Measuring a bank's output as its volume of loans would capture only a tiny part of the services that banks create, for example. Firms in other sectors are likely to have similarly difficult-to-define outputs; these include education, government, and nonprofits.

Another difficulty arises when one can conceptually define the output, but it is basically unmeasurable. These cases involve economic goods produced by the application of scarce resources but that, for one reason or another, are not captured in the current standard data collection apparatus. Healthcare is an example. What consumers seek to buy from the healthcare industry is health, but there is no straightforward way to measure this. Plus there is often a long and noisy lag between the purchase of healthcare services and any change in health. Producers of "free" digital goods like media sites also create outputs that can go unmeasured. Consumers' use of web searches, posts, views, etc., are not readily counted if they are not tied to a monetary transaction. While this is an old issue (free media has been present for a long time), it may be becoming more prominent.[5] Another unmeasured output exists when firms produce intangible capital goods for their own use in future production. In this case, the firm is producing an output that is neither counted in revenues nor incorporated into future capital input measures. We will return to this below in our discussion of capital measurement issues in productivity.

Measurement difficulties remain when output is more readily conceptualized and activity tied to the sale of those outputs is measured. Outputs in economic micro data

[5]It is important to recognize, however, that there are measured transactions associated with these goods that enter output. Free media and social network sites are often supported by (measured) advertising revenues. Consumers also need to purchase complementary goods like smartphones, tablets, broadband access, and mobile telephony to consume these free goods. Sellers of those complements should be pricing consumers' values of free goods into their own prices, which again are counted as outputs.

are not usually true quantities. They are instead more likely to be measured as buyers' expenditures on products, or equivalently, producers' revenues from selling them. While revenues may not be measured perfectly, they are probably one of the most accurately tracked economic values and are clearly correlated with output. However, a drawback of this approach—as necessary as it might be in the absence of quantity data—is that without product- or producer-specific price deflators (very rarely available in micro data), any price differences will be labeled as output differences. This poses obvious troubles in interpreting variations in measured productivity levels and growth across producers. If prices reflect at least in part idiosyncratic demand shifts or market power variation rather than quality or production efficiency differences, producers with high measured productivity may not be particularly technologically efficient.

In specific cases, some output quantity metric may be available. This may or may not be a close relative to the conceptual notion of output for the producer (e.g., counting cars, even if measurable, may not reflect the true output of automakers because of the heterogeneity of cars), but even if not it may still be a useful ingredient in a broader output index.

In absence of direct quantity metrics and when revenues measured as-is are not acceptable, the literature has proposed theory-based approaches to decompose revenue into price and quantity. These vary considerably in structure, from assuming (and estimating) a demand system to working off firms' first order conditions for inputs. These methods of course require assumptions, and the extent to which the necessary assumptions hold varies across particular empirical settings. Our sense is that the literature has not yet settled on "One Best Way" to do such decompositions. The assumptions of any given method do not hold exactly in any particular setting, and do not hold inexactly across settings equally. For now, researchers are left to determine the best option among a set of imperfect solutions.

Over time, the literature has become increasingly careful about denoting whether it is using revenue-based output and productivity measures or instead using output-based measures (whether those rare cases where quantities are directly measured, or instead when they have been backed out using the aforementioned methods). TFP is often labeled TFPR when measured using revenue-based output and TFPQ when output-based. By the definition of revenue, there is a simple identity linking the two metrics. If, as common, TFP is measured in logs (the log output minus the log of inputs), then $TFPR = lnP + TFPQ$,

where P is price.[6]

4.2 Input Measurement

Labor Variations in worker quality pose one of the trickiest issues in accurately measuring labor inputs. Workers and worker-hours are heterogeneous in reality. Treating them as homogeneous will attribute to productivity what are instead differences in (properly measured in a world with enough data) quality-adjusted labor inputs.

An approach researchers commonly employ to adjust for labor quality differences in micro data is to measure labor inputs using the wage bill (i.e., the producer's entire expenditure on labor). The notion is that market wages reflect workers' marginal products, so the wage bill serves as a quality-weighted total labor input measure. This correction has the advantage that wage bills, while not universally available, are often reported in production micro data. This methodology is not foolproof; wage variation might reflect the competitive structure of local labor markets (the labor-market-side analog to the output price variation issue discussed above), or causation could be in the other direction if more productive firms share rents with employees. Definitively pinning down labor quality's productivity contribution would require more direct labor quality measures, which are unfortunately quite rare in production data. That said, there is some evidence (Fox and Smeets, 2011) that the wage bill is a reasonable proxy for what is captured in more detailed labor quality measures.

Intermediate Inputs Intermediate inputs are almost inevitably measured as expenditures, the product of input quantities and prices, rather than separate quantity and price components. This raises the same price-variation-interpreted-as-quantity-variation issues discussed above. (The most common exception to this is electricity use, whose measurement in physical units—typically kWh—is made easier by its homogeneity.)

Capital Capital might be the most mis-measurement-prone component of any productivity measure. First, capital can exhibit considerable unmeasured quality variation. Capital vintages might vary in how much they embody the latest in technological progress,

[6] As we discuss below, despite some confusion in the literature, TFPR is not the residual of a production function estimated with revenue as the output measure. Such a residual generally includes parameters of the demand system that are not summarized by the (log) price, creating a gap between the residual and TFPR as defined here.

for example. Capital quality poses conceptually similar problems to the labor quality measurement issues discussed above. In practice, capital measurement can be even more troublesome. Measurable auxiliary correlates of labor quality like education and experience are sometimes available to allow quality adjustment of measured labor. Few such proxies for quality exist for measured capital.

Second, practitioners must typically use a capital stock to measure capital input, even though the true input into production is the flow of services that capital provides. Differences in capital utilization rates therefore create measurement error. A given capital stock can provide a high flow of capital inputs when used intensively or a low flow when not, but a typical stock-based capital input measure treats both the same.

Third, practitioners often build up capital stocks using the perpetual inventory method. Both the depreciation rate and current investment, which are necessary components of this process, are potential conduits for mismeasurement. Depreciation depends on the producers' capital mix and utilization intensity, both of which are typically unobserved. And measuring the correct level of investment requires a price deflator that equates effectively equivalent dollars of investment today to that in another period. Such investment deflators—even when available, which is not always, are probably more noise laden than output deflators.

A fourth problem is heterogeneity. The kind of capital that matters most for production can vary across industries. Some industries' production activities are extremely intangible-heavy, in others, it could be physical equipment and structures. Physical space, both size and geographic location, matters a lot in retail and wholesale.

Fifth, capital may not just be measured with error; it may not be measured at all. Intangible capital (brand value among a customer base, operational know-how, organizational culture, relationships with suppliers and customers, and so on) plays an important role in many markets, yet by nature is typically unmeasured. The contributions of these intangible capital inputs are therefore misattributed to productivity in standard measures, overstating true productivity.

Intangible capital's inherent unobservability can be managed in part by reinterpreting productivity as including the effects of intangibles. However, doing so can cloud the understanding of the mechanisms through which intangibles operate, how they vary across producers or over time, and how much measured productivity is affected by factors other than intangible capital.

Furthermore, as briefly mentioned in the output measurement discussion above, intan-

gibles have a second effect on productivity measurement. When first produced, investment goods, whether tangible or intangible, are output. A firm that makes intangible capital for its own future use as an input (e.g., builds a reputation for quality through product improvement efforts, manages customer relationships, reorganizes its internal structure to improve production or use new technologies) is making an output that is not counted as such. This initial production of intangible capital leads to understatement of productivity. The net effect of intangibles on productivity measurement therefore depends on the relative size of the productivity-overstating effect of not counting intangible capital inputs versus the productivity-understating effect of not counting intangible capital outputs when they are made. If intangible capital production ramps up faster than its placement as a capital input, then standard productivity measures will first understate a firm's productivity level. Later, as its intangibles come online as inputs, its productivity will be overstated. In steady state, these two measurement effects approach equality. However, in the meantime measured productivity growth and levels exhibit a J-curve effect: first falling, then rising substantially, only to eventually return toward baseline.

There has recently been a surge in attention paid to measuring intangible capital and more explicitly accounting for it in productivity measurement. The various approaches include building metrics from reported values in companies' balance sheets and income and loss statements. For example, one can assume some portion of the reported item Selling and General Administrative Expenditures is actually intangible investment and build a stock through the perpetual inventory method.[7] Just like with physical capital, this depends on appropriate deflators and depreciation rates, which are themselves not directly observable. Moreover, the ability to construct such proxies is dependent on data and industry context. It is therefore not surprising that no real consensus yet exists on the best approach. More research is definitely needed here.

4.3 Data Sources

Most IO-related work on productivity uses micro data on producers' outputs and inputs, and it structures measurement around some notion of a production function.

As noted above, productivity also has a cost-based interpretation, and some macroeconomic work on productivity uses this approach. At the micro level, however, cost data

[7]The item reported on the income and loss statement (abbreviated by SG&A). This poses a few challenges, chief among them the appropriate economic rate at which these assets need to be amortized over time and across assets, in the likely absence of underlying transactions in the economy – e.g., brand value of a company.

are rarely available. They are typically held by firms as proprietary and are hard to rely on due to mismatches between accounting standards and economic concepts. Plus, they rely on accurate measurement of often difficult to obtain micro-level prices.[8] As a result, we focus in this chapter primarily on methods developed to estimate production functions. However, remember that the approaches are connected. We can move from production functions to costs by adding cost minimization; the limit is the data.

The modal micro productivity study uses some form of micro data from countries' national statistical agencies. These data are collected with the intent of aggregating up to compute national income accounts, so they naturally have measures of outputs and inputs. The data are often richest for countries' manufacturing sectors, both because of the sector's historical importance and the fact that outputs and inputs are (at least somewhat) more straightforward to measure than in other sectors. Data outside manufacturing has become increasingly available, though posing a few more measurement issues for researchers to tackle. Depending on the particular national setting, additional production information might be available such as output prices, input prices, matched worker files, and financial data, though these are the exception rather than the rule. The major limitation to the use of such data is the (sensible and important, albeit taxing) confidentiality constraints that limit access to and reporting of such data.

Other commonly used data include the collected and collated balance sheets of publicly listed firms, like Compustat. These data are based on accounting conventions and, unlike the national statistical agencies' data, are not collected with the intent of facilitating aggregate economic analysis. Hence the mapping from data to production function is more strained. Labor inputs are not always measured well or cleanly separated from intermediates. There is lots of attrition due to mergers and acquisition, making survival analysis difficult. There is of course selection into being a publicly listed firm. And firms are inevitably multiproduct, with inputs seldom broken down by product, complicating the concept and measurement of the production function. Still, in some settings these data can be useful.

A middle ground between statistical agency and publicly listed firm data includes

[8]This same cost data limitation was an impetus behind the development, in IO, of methods to measure market power and conduct using demand estimation and output-based approaches. When micro-level cost data *has* been employed, it has traditionally come from firms in regulated industries, like electric and water utilities. (Readers of previous versions of this handbook will know this.) The unique regulatory and institutional environment of these industries lends itself to cost measurement. See Wolak (2003) for a discussion.

datasets like Orbis, which provides firm-level (or plant-level) data in long list of countries with annual balance sheet and ownership information. There are also micro data collected for particular industries or countries by international organizations such as the World Bank Enterprise Surveys.[9]

5 Recovering Productivity from the Data

Total factor productivity measures require labor, capital, and intermediate inputs to be combined into a single, composite input that forms the denominator of the TFP measure. The precise way to combine the inputs depends on the production function. The most common method is to weight each input by the elasticity of output with respect to that input, so the (logged) composite input is equal to the output-elasticity weighted sum of the logged individual inputs. This summation is exactly correct for a Cobb-Douglas production function. It is also, conveniently, the correct first-order approximation for any general production function.

Output elasticities are not directly observable; they are properties of the production function. Hence one must somehow estimate them, a process which itself introduces measurement error. There are two primary approaches used in the literature: factor shares and production function estimation.[10]

Regardless of which method one deploys and prefers in a particular application, we argue it is necessary to explicitly consider the market structure facing producers (i.e., the operating environment in both output and input markets) and the underlying model of production (single- or multi-product producers, for example). In addition, one needs to model two empirical ubiquities mentioned above: heterogeneity of productivity across producers and serial correlation in these differences (i.e., productivity persistence). Economic theory can guide us to tackle this problem in a general way, but it also implies that different settings, datasets, and questions often require different approaches and solutions.

[9] See `https://www.enterprisesurveys.org/en/enterprisesurveys`.

[10] A somewhat distinct literature with branches extending beyond economics into operations and engineering applies a third, entirely nonparametric, approach to productivity measurement: data envelopment analysis (DEA). We are not adept practitioners of DEA and as such do not cover it in detail here. An overview of DEA can be found in Cooper, Seiford, and Zhu (2011).

5.1 Operating Environment and Unit of Analysis

We want to focus on the economic environment under which the two main approaches are valid and informative about producers' abilities to supply the market. We focus here on behavioral assumptions and features of the operating environment, omitting some details of the econometric procedures. We refer interested readers to the original research.

It is useful to organize productivity measurement around three important dimensions: 1) Market structure: perfect or imperfect competition, 2) Unit of analysis: producer or product, and 3) Product: homogeneous or differentiated products.

Table 1 lists the various cases and how they map to the organization of our discussion. The market structure is listed by row (labeled A and B) and unit of analysis (producer or product) by column (1 and 2). *The Market Structure* distinctions should be readily apparent, although we will make an important distinction within imperfect competition between homogeneous and differentiated products, and the presence of strategic interaction between producers (i.e., oligopoly versus monopolistic competition).[11] As for the *Unit of Analysis* distinction, cases in the "producer level" column cover analysis of producers that are either explicitly single-product, situations where any underlying multi-product aspect of the producer are ignored, or where the researcher aggregates the data to the producer level. "Product level" refers to environments where multi-product firms are explicitly modelled and product-level data are used.

While the literature has not always explicitly categorized productivity measurement approaches in this way, we believe the taxonomy is useful for highlighting the main assumptions at work in the methods. It also facilitates discussions of where current practices lie and, just as importantly, where more research might be most useful. Speaking of that, the bulk of applications so far have implicitly worked with producer-level data under the assumption of perfect competition ($A.1$). This may require interpretation of the literature's findings accordingly. However, there is a growing set of work that is explicit about product differentiation under a variety of market structures (monopolistically competitive and oligopoly, both captured in $B.1.2$).

[11] Perfect competition by definition only considers homogeneous products, though in some settings it pools producers across different (geographic) markets when estimating the production function. In those cases, the resulting price variation needs to be considered.

Table 1: The operating environment and unit of analysis

	UNIT OF ANALYSIS	
	1. Producer Level	**2. Product level**
MARKET STRUCTURE		
A. Perfect competition	A.1	A.2
B. Imperfect competition	B.1	B.2
Homogeneous Product	B.1.1	B.2.1
Differentiated Products	B.1.2	B.2.2

Note: All cases consider heterogeneous firms (in terms of productivity).

5.1.1 Market Structure

The fact that we still distinguish between perfect and imperfect competition in the fourth volume of the *Handbook of Industrial Organization* may surprise some of those new to the productivity literature. In fact, much of the early work in identification and estimation of production functions assumed perfect competition. We do not see this as an inherent problem; rather, it serves as both a benchmark and a stepping stone toward more recent (and still unsettled as to their relative benefits and costs) methods that handle imperfect competition. We make an important distinction between models with and without strategic interactions.[12] We do this in the context of differentiated products, which covers a large share of economic activity and industries.

When discussing production function estimation under imperfect competition, we need to distinguish between product and factor market imperfections. This distinction is important when evaluating empirical approaches designed to estimate production functions.[13] Furthermore, we make an important distinction between imperfect competition through demand-side heterogeneity (like monopolistic competition with product differentiation)

[12]Unless noted otherwise, we only consider non-collusive models of oligopoly.

[13]For example, whether wage variation reflects worker quality heterogeneity or labor market power (i.e., monopsony), or the presence of capital market frictions (e.g., financial constraints), affects the validity of certain approaches and may require additional assumptions or structure.

versus purely through competition (e.g., homogeneous-good Cournot competition). This distinction is important when contrasting the factor-share and production function estimation approaches. It is also a crucial difference when comparing the *control function*, or *proxy*, approach and *dynamic panel data* estimation approaches.

5.1.2 Unit of Analysis

We distinguish between productivity analysis at the producer level (usually this refers to firm- or plant-level) from product-level analysis. In some cases product-level measures of output are recorded in the data and then aggregated to the producer level (we come back to this later in this section), without further attention given to the product-specific values, or even the fact that producers deliver multiple products at all. In those cases the distinction is moot.[14] The "Producer level" column contains the majority of empirical applications in the literature. The "Product level" column thus captures settings where researchers both observe product-level measures of output and explicitly consider multi-product production (or costs). While there was an older research tradition of estimating multi-product cost functions (see Chapter 1 in Volume I of this *Handbook* series), the focus shifted to estimating production functions for a host of reasons we already discussed. Recently, insights from the production function estimation literature have been used to analyze multi-product production output and input data.

5.1.3 Output and Input Data

The classification in Table 1 does not indicate whether researchers observe quantities of output and inputs, as opposed to revenues or expenditures (i.e., the products of quantities and prices). We do not include this important measurement dimension in Table 1, because we work under the premise that applied work relies on deflated monetary variables (output and input data, except of course employment), albeit often at different levels of aggregation for inputs compared to output.[15] This implies that we can safely ignore the difference between say revenue and output quantity (or expenditures and input quantities) for the cases captured in rows A.1 and B.1.1. In other words, either through deflating or directly

[14]When we consider the issue of observing revenue instead of physical output, the presence of multi-product producers is sometimes treated explicitly as it enters through the demand side, through product-level demand shifters.

[15]While it is not unusual for production data to include detailed industry-level output price indices (e.g., 4-digit NACE in the EU, or 6-digit or even more disaggregated in the US), this is less often the case for inputs.

observing physical units of output and input, the researcher assumes all producers within the industry under consideration make identical, homogeneous goods. Cases in the last row (capturing B.1.2 and B.2.2.), however, have to confront head-on the presence of differences in product attributes and therefore quality. The literature has only very recently started to consider this case, let alone in a multi-product setting.

5.1.4 Trade-Offs across Approaches

It is useful to make explicit which of these cases the two distinct approaches, production function estimation (PFE) and the factor-share (FS) approach, can accommodate (at least at this point in the literature). We also state where we see the most favorable prospects for researchers to make progress in advancing through the various columns and rows of our admittedly simplified taxonomy. Below, we will discuss each approach in detail, but at this point we want to emphasize how certain assumptions impact the validity of the approach.

The FS approach is valid under all market structure scenarios, provided cost shares can be computed with the available data. Effectively, if one is willing to assume that each producer statically minimizes cost each period, that all inputs are flexibly adjustable, and technology is characterized by constant returns to scale, an input's cost share will be directly informative about its output elasticity. The PFE approach relies on other restrictive assumptions, though perhaps these have not always been spelled out explicitly. We argue that both PFE approaches (control function and dynamic panel techniques) definitely capture case A.1 – i.e., perfect competition.

Deploying the control function approach in an imperfectly competitive setting requires certain modifications (including additional information) as well as restrictions on the specific form of imperfect competition, especially when considering models of oligopoly.[16] Finally, the dynamic panel data approach can in principle be applied to case B.1.1 without restricting the model of imperfect competition.

The cases in the multi-product column add distinct challenges, chief among them the unobserved input allocation across products, making a restricted version of PFE feasible.

[16]This is topic of recent research, see Ackerberg and De Loecker (2021) for a more detailed discussion.

5.1.5 Notation and Setup

Here and throughout this chapter, unless otherwise noted, we imagine some producer i (be it a firm, establishment, or some other production unit; we will only distinguish among these when it is important) producing a product j. For the most part we will consider single-product producers, so $i = j$. Output is Q. Inputs are X^H. We consider two types of inputs $H = \{V, F\}$, where V contains more flexible inputs like labor, intermediate inputs, and energy, and F contains the (sometimes quasi-fixed) capital stock. The corresponding factor prices are given by P^{X^H}. Lower cases denote logs. Productivity is Ω; its log is ω. Production function parameters will be denoted with β.

When working with production functions, we usually use the Cobb-Douglas form. It conveys most of the necessary intuition and makes notation easier. For most of what we do in the chapter, what matters is Hicks-neutrality (i.e., that productivity is multiplicatively separable from factor inputs), not the specific form of the production function.

Consider the following logarithmic specification of the production function. Our interest lies obtaining the output elasticities so that we can recover an estimate of productivity:

$$q_{it} = \beta_V x_{it}^V + \beta_F x_{it}^F + \omega_{it} + \epsilon_{it}, \tag{8}$$

We consider two types of inputs, variable (V) and fixed (or quasi-fixed) in the period (F). Anticipating the most common empirical approaches, we distinguish between errors in recording output (ϵ_{it}, assumed to be classical measurement error) and the unobserved productivity term ω_{it}.[17] We start our discussion with producer-level analyses from Table 1 (the single-product setting). The unit of observation (i) is a firm or plant. Both the factor share and production function estimation approaches introduce a set of distinct potential measurement issues, capturing both model misspecification and sampling error.

5.2 Factor Shares

The factor share approach for measuring factor elasticities (sometimes referred to as the first-order-condition approach) relies on the condition for static cost minimization: an input's output elasticity equals the product of that input's cost share and the scale elasticity. This implied optimization is in contrast to the production function approach, which

[17]The literature has predominantly considered a Cobb-Douglas specification. Departures often include translog (see De Loecker and Warzynski (2012) and De Loecker et al. (2016)), and increasingly the constant elasticity of substitution specification (see Grieco, Li, and Zhang (2016)), and departures from Hicks-neutrality (see Doraszelski and Jaumandreu (2018) and others).

requires only the existence of *a* production function, without the additional behavioral assumption.[18] The factor share approach has a long history in productivity measurement in macroeconomics (e.g., Basu and Fernald (1997) and indeed Solow (1957)), though it has seen steady and even increasing use in micro-level productivity analysis.

Of course the approach requires measuring inputs' cost shares. This can be less measurement prone than measuring input quantities, because, as discussed above, costs are based on expenditures, which tend to be directly measured, and do not rely on separately accurate measurement of price indexes. Capital costs are the most difficult. Direct measures of expenditures on capital inputs are rare, especially if that capital is owned (and hence rental rates are implicit) rather than rented.

While the theory behind the cost-minimization approach is straightforward, and the method avoids needing to estimate a production function regression, it is not assumption-free. If there are factor adjustment costs, the link between observed cost shares and the needed output elasticities will not hold at any given moment, because firms will be operating at input levels away from their long-run desired level. This misspecification error may be partially mitigated by using cost shares that have been averaged over time (or in micro settings, across producers as well). Such averaging smooths out idiosyncratic adjustment-cost-driven misalignments between actual and optimal input levels. The notion is that input levels are optimal on average, even if individual producers may be operating with idiosyncratically high or low inputs. That said, this is not a complete fix. It is only under very specific conditions that the average would exactly equal the true elasticity. Collard-Wexler and De Loecker (2016) suggest a modification in light of measurement error in capital, and propose to take the median over the producer-level ratio.

In practice, the scale elasticity is often assumed to be one, allowing for a straightforward calculation of the output elasticity for input H using:

$$\frac{P_{it}^{X^H} X_{it}^H}{TC_{it}}, \tag{9}$$

for $H = \{V, F\}$ and $TC_{it} = \sum_H P_{it}^{X^H} X_{it}^H$. In light of the production function we had in mind, a summary statistic of this ratio is used to compute the industry-specific elasticity. Typically, researchers take the average across all producers and a particular period in the

[18]However, as we will note, the estimation techniques that were developed to combat the econometric challenges in the production function approach (i.e., the endogeneity of inputs) bring in behavioral assumptions through the back door. These assumptions are arguably stronger than static cost minimization in many settings.

data:

$$\beta_X^H = \frac{1}{N_t} \sum_i \frac{P_{it}^{X^H} X_{it}^H}{TC_{it}} \tag{10}$$

A strength of the factor share approach is that it accommodates all cases listed in Column 1 without any modification. Product differentiation poses no challenge as long as one is willing to assume the production function is the same across all products. It does not rely on comparable output and input quantities to compute the elasticity.

In typical practice, researchers using the method apply it assuming constant returns to scale. In this case the output elasticity equals the cost share. Relaxing this assumption requires the estimation of the scale parameter via a regression of output on a composite input constructed as the cost-share-weighted sum of logged inputs. This introduces the same challenges laid out below regarding production function estimation. See Syverson (2004a) and De Loecker, Eeckhout, and Unger (2020) for illustrations of this approach.

In the perfect competition case of A.1, an input's cost share equals its revenue share, $\frac{P^X X}{PQ}$. The logic is simple. Under constant returns to scale and with all inputs variable, average cost equals marginal cost. Under perfect competition, marginal cost must equal the price a producer receives. In this case, one can avoid computing the cost of capital, which can be tricky in practice (rental payments to capital service flows are rarely directly reported). Capital's revenue share can instead just be one minus the directly measured input cost shares. This approach has a long tradition in the field of agriculture economics.[19]

While in theory the factor-share approach can equally handle the multiproduct production case (column 2), the researcher must observe factor expenditures separately for each product. As we will discuss below, these are not in the typical dataset. It may be difficult to even conceptualize product-level inputs under a multi-product production setup.

Given the simplicity of its implementation, it may seem that the factor share approach would be destined to be the main technique for estimating output elasticities. This is not the case; most studies estimate production functions. If we were to speculate as to the reasons why, one possibility is researchers' discomfort about some of the assumptions of the framework. Imposing constant returns to scale is a strong assumption. At the very

[19]It does pose a bit of a logical inconsistency. Under perfect competition and constant returns to scale, nothing definitively pins down a producer's empirical size. Optimal size is zero if marginal costs are greater than price, infinite if marginal costs less than price, or indeterminate if marginal costs equal price.

least, we recommend testing for this, even though this is not currently standard practice. Perhaps more importantly, there is substantial evidence, and we will come back to this later, of adjustment costs in factors of production. While traditionally thought to matter for the capital accumulation process, micro-level evidence suggests adjustment costs are important for labor demand as well (see Cooper, Haltiwanger, and Willis (2015) and Bloom (2009)). Whether the quick fix of averaging cost shares across producers and time sufficiently reduces this bias in practice remains a topic for future research.

5.3 Production Function Estimation (Producer level)

A second approach to obtain output elasticities is to estimate the production function by regressing output on inputs. The estimated coefficients are, depending on the functional form assumptions, either direct estimates of the elasticities or functions thereof. Total factor productivity is the estimated residual from the production function.

There are econometric challenges involved in this approach, with a considerable literature attempting to address these concerns.

There is the *transmission* or *simultaneity* bias: producers' input choices are likely to be correlated with their productivity levels. In almost any model of producer behavior, factor demands are a function of productivity. This creates a correlation between the explanatory variables (factor inputs) and the error term (productivity) in the production function regression, in violation of the assumptions of OLS.

In settings where productivity is being measured at the firm level, an additional potential selection bias is present. Because lower-productivity producers are more likely to exit from the sample, the econometrician will only observe a selected set of productivity draws.

It is important to underscore that the standard reaction, *go look for an instrument*, might be an option in specific settings, but it does not lend itself towards general use in production function estimation using micro data. This is for at least two reasons. First, it is a well-documented fact that ideal instruments to deal with the simultaneity bias, input prices, are often either not available to the econometrician, or their variation across producers is unlikely to satisfy the exclusion restriction (i.e., be orthogonal to the productivity shock).[20] Second, and related, the internal consistency of an instrument is severely restricted once it is understood that, at least implicitly, restrictions are imposed

[20]See Griliches and Mairesse (1995) and Ackerberg et al. (2007).

on market structure, the degree of product differentiation, product scope, and how output and inputs are recorded. As will become clear below, demand-side shifters may be good instruments, but this in turn affects the validity of the approach taken more generally.

5.3.1 Perfect Competition (A.1)

Until the past decade or so, perfect competition had been the main environment considered in the production function estimation literature.[21] There are good reasons for this. The productivity literature grew in considerable part out of work on productivity in the agriculture sector. Assuming perfectly competitive output and factor markets was quite reasonable in many such settings. As the literature increasingly turned its focus to manufacturing, this assumption became harder to maintain. Increased product differentiation and technological change affected the distribution of market shares across producers and the amount of producer market power. Nevertheless, even now an approximate majority of applications assume this environment (implicitly, for the most part), and even many recently developed econometric techniques remain rooted in this paradigm.

There are two broad approaches under the setup of perfect competition, the control function and the dynamic panel data approaches. We refer the reader to the overviews of Ackerberg et al. (2007) and Ackerberg, Caves, and Frazer (2015) for more details on the relationship among the two. However, both approaches can be extended and adjusted to accommodate departures from perfect competition. It is also instructive to highlight that the (revenue) factor share approach remains a viable alternative under this situation (A.1), albeit assuming constant returns to scale and no adjustment costs in *any* input (at least along the time dimension of the dataset, typically a year).

Control Function Approach The control function approach was pioneered by Olley and Pakes (1996), modified by Levinsohn and Petrin (2003), and consolidated by Ackerberg, Caves, and Frazer (2015). It connects the production function to an underlying economic model that describes the behavior of producers and the operating environment in which they compete. This permits, in the spirit of Griliches (1967), an interpretation of the estimated production function coefficients and an evaluation of the potential biases that plague least squares estimation through the lens of economic theory.[22] This approach

[21] Unless noted explicitly, we refer to both perfectly competitive product and factor markets.

[22] See also Reiss and Wolak (2007), Section 4, for a discussion in the context of structurally estimating production and cost functions.

includes two fundamental pieces: 1) construction of optimal input demand through modeling either static input choices or forward-looking investments, and 2) a specific time-series for the productivity process.

The input demand assumption stipulates that the first order condition for the producer's demand of some factor input (in logs, d) is an unknown function of productivity ω_{it} and other producer-specific observable inputs:

$$d_{it} = d_t(\omega_{it}, x_{it}^V, x_{it}^F). \tag{11}$$

In the classic setup, factor prices and the output price are assumed to be common across firms. These and any other non-producer-specific influences on factor demand are accommodated by the time subscript on the input demand equation $d_t(.)$.

The key assumption is that this optimal input demand is monotonic in (scalar) productivity, conditional on the other observables in the input demand. Monotonicity implies a unique mapping between observable input demand and unobserved productivity, thus allowing the input demand function to be inverted to recover productivity as a function of data and parameters:

$$\omega_{it} = d_t^{-1}(d_{it}, x_{it}^V, x_{it}^F). \tag{12}$$

Under these assumptions, the unobserved productivity shock can be controlled for (exactly) by this function, obviating the simultaneity problem. The precise form of the control function is generally unknown, so in practice the researcher approximates $d_t^{-1}(.)$ using a flexible function of d_{it}, x_{it}^V, and x_{it}^F, like a high-degree polynomial or other non-parametric techniques.

The control function in Olley and Pakes (1996) is an investment policy function (i.e., a dynamic control) derived from a special case of Ericson and Pakes (1995): $d_{it} = i_{it}$, where i_{it} denotes logged investment.[23] Levinsohn and Petrin (2003) propose using a static control like intermediate inputs or electricity use instead of investment to deal with two major concerns. First, producer investment is lumpy and is often observed to be zero. Because the input demand (investment) function is not invertible at this level of investment (e.g., with fixed costs of investment, a range of productivity levels are consistent with zero investment), the researcher cannot use these zero-investment observations in estimating

[23]This restricts the underlying model of market competition to one where only producer i's productivity determines its investment choices, conditional on capital and the aggregate state in the market. In addition, the law of motion of productivity, introduced below, is restricted to exogenous Markov processes, instead of allowing investment to affect future productivity draws.

the production function. Second, when considering slight deviations from the original Olley and Pakes setup, deriving the investment equation that allows inversion proves to be challenging. The trade-off to using a static input proxy is that investment decisions are highly informative about a producer's expectations of future productivity, and this information is largely lost in the static case. While the Olley and Pakes (OP) approach has to keep the function $d_t(.)$ non-parametric, in the Levinsohn and Petrin (LP) approach the functional form directly relates to the production function. In the Cobb-Douglas case, for example, the inverted input demand equation is log linear.[24]

This explains how the control function can address the simultaneity bias. However, one cannot simply include this control function in a production function regression and be done. The reason why is that the inputs of the production function are also in the control function. The derivative of output with respect to any input in such a regression is not simply the output elasticity. Separating the production function from the control function is where the timing assumptions come in.

A second key assumption of the control function approach is that productivity follows a first-order Markov process:

$$\omega_{it} = g(\omega_{it-1}) + \xi_{it}. \tag{13}$$

This is the core timing assumption, and its application raises the issue of dealing with selection (survivor) bias.

Selection Bias The original Olley and Pakes (1996) approach deals with the non-random exit of producers by turning the (expected) productivity distribution into a truncated distribution based on a producer-specific exit threshold that depends on the producer's state variables, in particular capital. Selection bias is likely to be especially pronounced in periods of drastic change in the operating environment. This has been discussed prominently in the literature, but we do want to call one common wisdom into question. The OP methodology highlights the importance of correcting for this selection bias. It materializes most in the output elasticity on fixed factors of production, in their case capital, since it is precisely this variable that generates the option value of remaining in the market, given unfavorable productivity shocks. OP show that the use of an unbalanced panel gets one most of the way (practically speaking, raises the capital coefficient to a more realistic value). This is for good reason; not selecting the sample on survival

[24]It is this case that highlights the identification challenges we discuss below.

through the entire sample period insulates from selection on the unobservable productivity shock.

This result has become an argument for *no* further treatment of the selection bias once an unbalanced panel is used. However, we note that this strategy does not insulate from selection on productivity *and* inputs. This would be the case if, say, a producer's decision to exit the market depended not only on its productivity forecast but its size (measured by capital) as well. Or if firms with a large number of employees had different propensities to stay in the market due to unobserved government actions (e.g., bailouts).

To address selection bias, control function methods add a step that recovers the producer's survival probability (from $t-1$ to t, say) as predicted by the relevant state variables of the problem. This survival probability is then included in the forecast of productivity (the $g(.)$ function). In the standard OP setup, the relevant state variables end up being investment and capital, as productivity has again been solved out for. See Ackerberg et al. (2007) for more discussion of the selection bias correction.

Procedure We now have the necessary pieces to estimate the output elasticities. We follow the more general treatment of Ackerberg, Caves, and Frazer (2015, henceforth ACF), also a two-step procedure. The first step replaces the unobserved productivity term with the inverse of the optimal input (investment or intermediate input) demand equation (equation 12). Collecting terms yields an equation that predicts producer output using inputs and the relevant control variable (d):

$$q_{it} = \phi_t(d_{it}, x_{it}^V, x_{it}^F) + \epsilon_{it}, \tag{14}$$

where $\phi_t(.)$ captures predicted output, or $\omega_{it} + \beta_V x_{it}^V + \beta_F x_{it}^F$. The first stage has only one purpose, to strip predicted output from measurement error.[25] The subscript on the control function is important, both theoretically and empirically. It is typically described as capturing variation in the function over time, thus capturing market-wide movements in demand and supply conditions. However, it can be interpreted a bit more broadly. For example, even under perfect competition, researchers may want to take into account that producers (facing the same production function) are sometimes active across different

[25]This can be quite important in its own right; see Collard-Wexler (2013) for an application. Another interpretation of ϵ_{it} besides measurement error is that it is actual *surprise* output that was not predictable even to the producer, given its chosen inputs and known productivity level. The distinction between pure measurement error and realizations of output affects the specific econometric treatment of this error term.

markets, e.g., regionally segmented markets. Thus they face different factor and output price conditions. These could be subsumed in a market-specific term $\phi_t(.)$, where now the subscript denotes a market-time combination.

The remainder of the approach is centered on timing assumptions regarding input choices after the productivity shock ξ_{it} is recovered using the first step estimates and the assumed productivity law of motion. We know that $\omega_{it} = \phi_{it} - \beta_V x_{it}^V - \beta_F x_{it}^F$. For a guess of the parameter vector β, we can compute productivity and use the assumed law of motion on productivity (equation (13)) to recover the productivity shock ξ_{it}. The final, second step of the procedure then relies on the moment conditions:

$$\mathbb{E}\left(\xi_{it}(\beta)\begin{pmatrix} x_{it-s}^V \\ x_{it}^F \end{pmatrix}\right) = 0 \tag{15}$$

This approach distinguishes between inputs that react to the productivity shock within the period, x^V, and those that do not x^F. The specific law of motion assumed on this input x^F (in levels) injects another timing assumption into how this dynamic input's accumulation process takes place. In almost all work (and in the OP, LP, and ACF papers), x^F represents the logged capital stock (or $X^F = K$). The standard law of motion is given by

$$K_{it} = (1-\delta)K_{it-1} + I_{it-1}, \tag{16}$$

where I_{it-1} is investment at time $t-1$. This timing assumption is what permits identification using the moments above. Conditional on the persistent part of productivity, lagged investment is assumed to be orthogonal to the productivity shock. In other words, the capital stock at time t cannot react to the contemporaneous shock to productivity. This is *not* the case for the input x^V (labor in the case of OP-LP-ACF), which is precisely why different moment conditions are suggested.

The control function approach was used in a variety of settings using the approaches laid out in the original articles of OP and LP, and not much attention or care had initially been given to the precise moment condition for these variable input coefficients. Ackerberg, Caves, and Frazer (2015) and Bond and Söderbom (2005) independently arrived at the conclusion that in the context of this setup (case A.1), the identification of a perfectly flexible variable input is challenging and even infeasible unless additional assumptions are made.

The easiest way to see the challenge is to consider a gross output Cobb-Douglas production function in labor, capital, and an intermediate input. The latter is the control

variable, d in our notation. Solving for the optimal intermediate input demand and plugging its inverse in the production function (thereby eliminating unobserved productivity) leaves the researcher with an equation that no longer features *any* of the production function coefficients. This demonstrates the non-identification result in the first stage of the control function approach.[26] At the same time, a variety of other data generating processes for the labor choice eliminate this concern.[27]

The literature paused for some time between the seminal works of Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg, Caves, and Frazer (2015). But the conclusion is that the user has a great deal of flexibility to decide what moments she is willing to use (essentially, letting s be either 1 or 0), a choice that ultimately reflects timing assumptions. This presents a researcher with many choices to tailor the method to the specific institutional and market details of the application.

The focus on production function identification in this body of work led to a more careful evaluation of the underlying economic models and econometric evaluation of the moment conditions used in estimation. A recent literature further studies the timing assumptions; see Ackerberg et al. (2020b) and Ackerberg (2020). A promising avenue may be to supplement the production data with survey questions related to the timing of firms' decision making. This adds empirical content to researchers' notions of firm information sets.[28] Another challenge in the standard ACF setup is global identification. Various approaches have been forwarded to eliminate the spurious local minima presence in finite samples (Kim, Luo, and Su (2019) and Ackerberg et al. (2020a)).

Dynamic Panel On the other side of the Atlantic, an alternative approach was developed to estimate dynamic panel data models. This was not done primarily for production function estimation. Rather, Arellano and Bond (1991) considered a much more generic

[26]Inverting the profit-maximizing material demand equation under perfect competition gives: $\omega = -\beta_V x^V - \beta_F x^F + d + c$, where c is a constant consisting of the common output and materials input prices and the material output elasticity. Plugging this into equation (8) yields $q = c + d$.

[27]One can achieve identification of one flexible input, say labor, by considering optimization error in labor choices, while relying on another input used in fixed proportion to output to proxy for productivity. A dynamic control, like investment in OP, provides another identification scheme. This essentially relies on an additional timing assumption where within a period investment decisions are made after labor choices, and importantly, sub-period productivity shocks hit the firm, breaking the non-identification. See Ackerberg, Caves, and Frazer (2015) for more discussion.

[28]It also connects to a separate literature in IO that studies producer decisions under a variety of assumptions on their information sets, like entry models (see Berry and Reiss (2007)).

setup of identifying parameters in (what now have become) standard panel data models. These were later modified and extended by Blundell and Bond (1998) and have been used since to estimate production functions. This approach has seen somewhat less uptake compared to the control function approach. While appealing in theory, in practice users estimated what were often thought to be unrealistically low or even negative capital coefficients. This can occur when (quasi) first-differencing data, which throws out variation across producers in the capital measure and raises its noise-to-signal ratio. Indeed, one of the motivations for the original OP approach was preserving the typically enormous cross-sectional variation for the sake of identification.

At a methodological level, there are two main differences between the dynamic panel and control function approaches. First, the dynamic panel method relies explicitly on a linear productivity process (an AR(1)). Second, it allows for a fixed effect in productivity. As observed by Ackerberg, Caves, and Frazer (2015), this setup no longer requires the control function approach's first stage. Further, as will become clear, it bypasses the need to invert the input demand equation and the complications that come along with that.[29] Instead, it relies on a statistical process for productivity and can proceed by directly forming moments on the joint error term, productivity, and the measurement error.

Let us illustrate the approach by considering a special case of the Markov process (equation (13), $\omega_{it} = \rho\omega_{it-1} + \xi_{it}$ – i.e., an AR(1) process.[30] Starting from the production function, substituting in the specific linear productivity process and replacing lagged unobserved productivity using the production function gives rise to:

$$q_{it} = \beta_V x_{it}^V + \beta_F x_{it}^F + \rho\omega_{it-1} + \xi_{it} + \epsilon_{it} \tag{17}$$

$$q_{it} = \beta_V x_{it}^V + \beta_F x_{it}^F + \rho(q_{it-1} - \beta_V x_{it-1}^V - \beta_F x_{it-1}^F - \epsilon_{it-1}) + \xi_{it} + \epsilon_{it} \tag{18}$$

The moments conditions used to obtain the production function coefficients are similar to the ones of the control function:

$$\mathbb{E}\left(\tilde{\xi}_{it}(\beta)\begin{bmatrix} x_{it-1}^v \\ x_{it}^f \end{bmatrix}\right) = 0, \tag{19}$$

where $\tilde{\xi}_{it} = \xi + \epsilon_{it} - \rho\epsilon_{it-1}$. Note that this is not exactly what the dynamic panel approach suggests. It typically allows for a fixed effect, requiring more work to identify

[29]Of course, if one were to drop the measurement error term ϵ_{it} from the control function setup, the first step could be dropped there as well. Moments would be formed directly on the productivity shock, given that $\omega(\beta)_{it} = q - \beta_V x_{it}^V - \beta_F x_{it}^F$. This would no longer require invoking optimal input demand behavior and the concomitant restrictions guaranteeing identification.

[30]The dynamic panel data literature considers the presence of a fixed effect, ω_i.

the persistence parameter (as now q_{it-1} would be correlated with ω_i). This often leads to a (quasi) first differencing strategy.[31] The dynamic panel approach, by leaning entirely on the linearity of the productivity process, does not treat the selection bias. Note that it is precisely the presence of measurement error in output, while taking into account the non-random exit of firms and not committing to a particular process for the productivity shock, that requires the OP approach to consider an optimal input demand equation.[32]

Here, we employ the fact that measurement error is independent from the productivity shock. However, the moment condition on x^V will not identify the coefficient when this input is fully flexible and input prices do not vary across firms. What breaks the non-identification of the variable input coefficient is the presence of either serially correlated demand shifters or factor prices. (Though note that the former is formally outside the environment we assumed under case (A.1).) The presence of either of these links input use over time, turning the lagged variable input into a valid instrument. Of course if one believes all inputs face adjustment costs, effectively all inputs are like x^f_{it} in the model, that would also restore identification. This strategy relies, however, on assuming that adjustment costs are a function of the level of inputs, creating variation across producers independent from the arrival of new productivity shocks.[33]

The applied researcher thus faces a trade-off between including observable factors influencing input demand explicitly, or instead relying on a particular statistical process of the productivity shock, and in addition on the *presence* of factors that turn lagged inputs valid instruments. From an applied point of view, the latter may be attractive in that it does not require to observe such a factor. At the same time, however, there is no way to check whether these factors are relevant and satisfy the necessary conditions to support identification.

Discussion Olley and Pakes (1996) created their approach to study productivity growth that accompanied deregulation of the telecom equipment industry. The non-random exit and reallocation of market shares among industry producers called for an approach that provided estimates of plant-level efficiency using standard production data, while ad-

[31] We refer the reader to Arellano and Bond (1991) and section 4.3.3. in ACF for a detailed comparison of both approaches.

[32] Below we discuss another important distinction, namely, the presence of endogenous productivity processes that are heterogeneous across producers.

[33] There is a literature that supports these assumptions in the micro data, for labor and capital use (see Cooper and Haltiwanger (2006)).

dressing the simultaneity and selection bias. The focus was thus on obtaining important economic objects while relying on a theoretically sound framework within the market of interest to interpret the data. The approach was subsequently taken out of the market context of their application and became the go-to routine for production function estimation. That is a tall order for any method, and can (and did) lead to some instances of pounding square pegs in round holes.

These difficulties spurred a series of papers that further deepened the literature's understanding of both its potential extensions and limiting caveats. Levinsohn and Petrin (2003) built on OP by demonstrating that instead of relying on an inverted investment policy function, a perhaps simpler static optimization problem (the hiring decision for fully flexible factors, like intermediates) can control for the unobserved productivity shock. A motivation for this modification is that especially in producer-level data from developing countries, reported investment is often zero or missing, preventing inversion and making the observation worthless.

Ackerberg, Caves, and Frazer (2015) revisited the precise identification conditions under which these approaches are valid. It offered a more general way to deal with potential non-identification concerns. More recently, Gandhi, Navarro, and Rivers (2020) used a perfectly competitive framework to study the identification properties under a non-parametric production function (conditional on a scalar Hicks-neutral productivity shock). These two papers also made clear the non-identification of the variable input coefficient, here β_V, under the canonical setup, a point previously made by Bond and Söderbom (2005).

An interesting element of the approach in Gandhi, Navarro, and Rivers (2020), given the discussion above, is that it essentially incorporates a (non-parametric) factor share approach. The shape of the production function and the first-order condition for a flexible input give rise to a non-parametric share equation that features the derivative of the production function. Adding a Markov productivity process yields a non-parametric estimator of the production function parameters through integration over the flexible input.

While the identification arguments in these papers remain largely theoretical, they have proven extremely useful in pushing the literature to a more careful treatment of the problem. The literature has noted that under perfect competition and absent input price variation or demand shifters, the output elasticities of perfectly variable inputs are challenging to identify, and may require very strict and implausible assumptions. This may not be too much of problem, given that one can naturally appeal to factor shares to

compute the output elasticity (albeit at the expense of having to assume constant returns to scale).[34] We refer the reader to Ackerberg et al. (2007) for a more detailed discussion of the econometric challenges and the commonalities and contrasts between the control function and dynamic panel approaches.

These developments also showed directions for relaxing the assumption of perfect competition. However, the richer market environment forced researchers to deal with the elephant in the room: output and input price heterogeneity.

5.3.2 Imperfect Competition (B.1)

While the assumption of perfect competition has been the norm in much of the productivity estimation literature to this point, few markets strictly adhere to its assumptions. Most producers have some degree of market power, whether in product or factor markets. Explicitly accounting for the consequences of imperfect competition when estimating production functions requires the accommodation of several new elements.

First, departures from perfect competition can capture both non-strategic producer-level differences (reflecting demand or product quality differences) as well as strategic interactions among producers. The latter is a topic of active and recent research, while the former has received more attention in the literature relying on a monopolistically competitive setup.[35] Second, the potential presence of horizontal or vertical product differentiation complicates the comparison of quantities produced. Additionally, in most datasets output and input are recorded as sales and expenditures, respectively. This introduces price errors into quantity measurements.

An important issue is that the popular control function methods implicitly restrict the form of competition. The dynamic panel approach, on the other hand, is not affected by strategic interaction across producers or demand-side differences across producers, as long as one can compare output and inputs (more on this below) and is willing to stick with the exogenous linear productivity process, thereby ignoring selection bias.

In general, the key ingredients of production function estimation approaches discussed above are preserved under imperfect competition. This includes the law of motion for

[34] One can construct a hybrid factor-share/control-function approach, where first the variable output elasticity is directly computed and the remaining elasticities (on the fixed factors) are estimated to free up the returns to scale. See Collard-Wexler and De Loecker (2016) for an illustration.

[35] Most applications to date have been in international trade, e.g., De Loecker (2011a) and Bilir and Morales (2020)

productivity and timing assumptions on inputs. However, an important distinction is that the input demand equation needs to allow for differences in demand due to differentiated products or strategic interactions. That is, the optimal input demand equation (equation 11) needs to reflect the operating environment and market structure. To see this, it is useful to go back to the input demand equation:

$$d_{it} = d_t(\omega_{it}, x_{it}^V, x_{it}^F, \boldsymbol{z}_{it}), \tag{20}$$

where the additional variables in $\boldsymbol{z}_{it}$ reflect input price differences, demand-side differences (e.g., different demand curves or shifters), and marginal revenue shifters from strategic interaction among market firms. For instance, De Loecker and Warzynski (2012) includes wages and export status in z, while De Loecker et al. (2016) inject price, market share, product dummies, and tariffs. Note that any omission of such relevant factors would invalidate the first stage by effectively not controlling fully for unobserved productivity shocks. Put differently, there is no one-to-one mapping between productivity and input demand if we fail to include all relevant shifters (coming from either cost, demand, or market structure), fouling the necessary inversion that would allow inputs to proxy for productivity.

In the most general oligopoly settings, this has stark implications for what should be included in the control function. It should include not just the index producer's demand shifters but also all competing producers' productivity levels and relevant state variables.

In this section, we discuss these challenges. We organize our discussion along the lines of Table 1. The first case, homogeneous good oligopoly case, can in principle rely on the approaches discussed under A.1. We devote most of our attention to settings where producers face different conditions in the operating environment, and where the researcher has to confront the additional price errors coming from observing revenues and expenditures. The details of the implementation depend on whether the researcher wants to allow for strategic interactions among the producers.

Homogeneous Good The imperfectly competitive, homogeneous good setting (e.g., Cournot competition) has not been explicitly analyzed in the productivity estimation literature. However, most of the main insights from the perfectly competitive setting apply.

Under product homogeneity, even though the output's price is above marginal cost, the

uniformity of price means producers' outputs are perfectly comparable.[36] Researchers can deploy the factor share method or dynamic panel data approach with little modification.

Control function methods work as well, though the control function (based on input demand, whether investment or intermediates as the case may be) should be modified to reflect the effect of competitive interactions among market producers. To see why, recall the discussion above regarding the presence of additional input demand shifters $\boldsymbol{z}$. These must be included in the input demand function so that any two firms with the same productivity and capital stock will select the same amount of the input d. In imperfectly competitive, homogeneous good settings, it is important to correctly index this function to reflect the relevant market (one might label it "the state of competition in the market"), capturing output and factor prices. This underscores the importance of correctly indexing the control function by either time or time-market, depending on the industry's structure and level at which the good can be considered homogenous. In many ways, one can treat such differences in competition across (time or geographic) markets as analogous to the aforementioned differences in aggregate market conditions across perfectly competitive markets.[37]

It is important to underscore, however, that the challenge of identifying the variable input coefficient (β_V in our notation), in either the control function or dynamic panel approach, is significantly reduced under models of oligopoly. This is because productivity shocks of competitors now move around a producer's individual residual demand curve, and therefore its optimal input demand. The exact conditions and details of this setting are the subject of ongoing research.[38]

Product Differentiation Product differentiation affects all production function estimation approaches, as it is fundamentally about how to compare output and input data across producers when products have different attributes and thus vary in (perceived) quality by consumers. The fact that most users have output and input data recorded in

[36]In practice output can be measured either through directly observing quantities or through use of a deflator based on the observed market-clearing price. The same procedure can be adopted for inputs, though typically deflators are only available at more aggregate levels.

[37]There is, however, a tension in deploying non-parametric estimation techniques in the first stage, where implicitly the number of firms in the market is assumed to go to infinity. This topic requires further research. Apart from that complication, one could interpret the existing work deploying the control function estimator as at least theoretically consistent with a symmetric oligopoly setting like a homogenous good Cournot model.

[38]Ongoing work of Ackerberg and De Loecker (2021) discuss this in more detail.

the form of revenue and expenditure can present additional complications, though even when physical output and inputs are recorded, the presence of product differentiation still poses a challenge, as physical units may not be particularly comparable.[39]

To keep notation light, let us consider a simple one-factor production function consisting of a variable input and productivity:

$$q_{it} = \beta_V x_{it}^V + \omega_{it}. \tag{21}$$

Suppose the researcher has data on (deflated across time) sales and the input expenditure for the set of active producers in the market. The practice of deflating stems from the tradition of using industry-wide time series data on output and inputs and industry deflators (typically constructed by statistical agencies). Deflating producer-level output and input data by industry-wide deflators does remove aggregate movements in prices over time, but leaves the deviation around industry average price and input prices in the error term of the production function.[40]

While theory tells us we should relate physical output to physical inputs, the data implies we now relate (deflated) sales to (deflated) material expenditures:

$$r_{it} = q_{it} + p_{it} = \beta_V(p_{it}^V + x_{it}^V) + \omega_{it} + p_{it} - \beta_V p_{it}^V \tag{22}$$

where p denotes the log price, p_{it}^V the input price of material, and r log revenue. Of course this will only be informative about the technical relationship (and thus measure the output elasticity) if all producers in the industry face the same output and input prices, thereby turning this regression of sales on material expenditures into a regression of output on materials. If either output or input prices vary across producers, the estimation of the production function will be generally biased both through a correlation between measured input and the error term, yielding a biased estimate of the production function coefficient (β_V), as well as through the productivity residual capturing both output and input price variation.

While this issue had long been recognized by practitioners, reflected in the use of the most disaggregated output and input deflators available, it came to the fore after the

[39] An example of this from the literature regards units of measured employment, typically total employees or hours worked. Because of concerns that there may be systematic differences in the quality of these units across producers, some applications use the wage bill (i.e., expenditures on labor) to measure labor instead, to hopefully have better measured quality-adjusted labor input.

[40] In what follows, our notation implies that the output and input price are expressed relative to the industry price index.

seminal contribution of Klette and Griliches (1996). They discuss the implications of output price heterogeneity and introduce an approach that provides separate estimates of the production function and the assumed demand system. Klette and Griliches (1996) introduced the *omitted price variable bias* that occurs when relating revenues to inputs. Its treatment was integrated into the control function approach in Levinsohn and Melitz (2002) and De Loecker (2011a). At the same time, Foster, Haltiwanger, and Syverson (2008) introduced the distinction between TFPR and TFPQ and how these concepts relate to firm survival.

Since this earlier work, researchers have adopted a variety of practices to mitigate the presence of unobserved prices in the production function. These practices have centered on using output price indexes to convert output in monetary units (e.g., USD) to physical units, adding more structure to the problem through assuming and estimating a demand function jointly with the production function, or exploiting the predicted correlation between output and input prices. This work has mostly focused on dealing with output price variation, for at least two reasons. First, there seems to have been a notion, albeit not backed by much formal testing, that input price variation is somehow less prevalent than output price variation within industries. Second, researchers have increasingly found production data with output prices but not input prices available, leading to a bit of a "looking under the lamppost because that's where the light is" phenomenon.

To summarize, in applied work dealing with productivity estimation when products are differentiated, one or a combination of the following is adopted:

1. Deflate revenue data using:
   (a) Industry-wide price index,
   (b) Product-level prices,
   (c) Product-producer level prices;
2. Add a demand system and demand shifters,
3. Exploit the output-input price wedge.

We discuss each of these in more detail.

Deflating Revenue The revenue deflation strategy converts observed revenue to output data using additional price data. The default is the use of an industry-wide price index. This eliminates common price trends from revenue data. In the specific case of

homogeneous products without any other sources of price heterogeneity (be it a spatial or regulatory dimension), this exactly converts revenue into quantities. Over the past decade or so, researchers have made increasing efforts to find more disaggregated price data, whether at the producer-, product-, or producer-product level. Researchers have used these data to construct producer-level (often firm-level) price indexes. This approach, however, brings in multi-product production through the back door, which can require special treatment (see below).[41]

Adding Demand-Side Information The demand system approach does not rely on price data. Instead, it explicitly models the revenue generating function from the production function and demand primitives. This approach requires observed demand shifters to disentangle demand and supply parameters. There have been a variety of different demand functions imposed across applications, chosen for a combination of reasons including tractability, generality of application, and empirical fit.

For purposes of illustration, we outline the approach of Klette and Griliches (1996) and De Loecker (2011a) in a simplified case. First, let all producers face a common input price. Second, assume demand can be described by a horizontal differentiated (conditional) demand system:

$$Q_{it} = Q_t \left(\frac{P_{it}}{P_t} \right)^{\psi} \exp\left(\nu_{it}\right), \tag{23}$$

with Q_t, P_t, ν_{it}, and ψ be industry-wide demand, the average price (i.e., the observed price index), an idiosyncratic demand error, and the elasticity of demand, respectively. The main insight of Klette and Griliches (1996) is to express observed revenue as a function of inputs, productivity, and demand-side shifters and parameters. Inverting and taking logs of equation (23) yields an expression for the (log) inverse demand system. Consider logged revenue, $r_{it} = q_{it} + p_{it}$ and replace the logged price by the inverse demand system.

[41] There are an array of examples of the use of producer-level prices in production function estimation. See Collard-Wexler and De Loecker (2015), Smeets and Warzynski (2013), Dhyne et al. (2020), Foster, Haltiwanger, and Syverson (2008), Rubens (2020), Allcott, Collard-Wexler, and O'Connell (2016), Slavtchev, Bräuer, and Mertens (2020), Morlacco (2017), Eslava and Haltiwanger (2020), Forlani et al. (2016), Valmari (2016), Orr (2019), Itoga (2019), Stiebale and Vencappa (2018), Atalay (2014), Backus (2020), Doraszelski and Jaumandreu (2018), Pozzi and Schivardi (2016), Ornaghi (2006), and Mairesse and Jaumandreu (2005).

Deflated revenue is then given by:

$$r_{it} - p_t = \left(\frac{\psi+1}{\psi}\right) q_{it} + \frac{1}{|\psi|} q_t + \tilde{\nu}_{it} \tag{24}$$

$$= \alpha_V x_{it}^V + \frac{1}{|\psi|} q_t + \tilde{\omega}_{it} + \tilde{\epsilon}_{it}, \tag{25}$$

where the second line plugs in the production function, $\alpha_V = \left(\frac{\psi+1}{\psi}\right)\beta_V$, $\tilde{\omega}_{it} = \left(\frac{\psi+1}{\psi}\right)\omega_{it}$, and $\tilde{\epsilon}_{it}$ captures idiosyncratic output and demand shocks. Explicitly writing the revenue function in terms of production and demand drives home two important points. First, the revenue function coefficients are reduced-form coefficients that combine the demand and production parameters (β, ψ). In other words, the "revenue elasticity" of a factor of production is not its output elasticity from the production function. Second, both observable (here q_t) and unobservable demand shifters enter the specification in addition to unobserved productivity. Thus, if one is to recover productivity or the output elasticities, these demand shifters have to be taken into account. And this is all in addition to treatment of the well-known simultaneity bias.

De Loecker (2011a) combines the control function approach with the Klette and Griliches (1996) structure. The application shows how to combine aggregate and product-level data to separately identify the demand and production function parameters, allowing the study of productivity while controlling for demand-side heterogeneity. The integration of both approaches rests on the important observation that whichever factor demand d is used in the productivity proxy (static (LP) or dynamic (OP)), at a very minimum one has to think through how demand-side heterogeneity enters the input demand equation and include controls for it. In this particular study, the additional shifters $\boldsymbol{z}$ include product and product-segment demand shifters as well as producer-level quota indicators. Including these serially correlated, observable demand shifters is not only important in the first stage (as in equation (20)), it also supports identification of the variable input coefficients by creating correlation across variable input choices across time periods for a given producer (using moments in equation (15)).

The results indicate that standard estimates that ignore the price error tend to underestimate returns to scale. Further, the imposed nested conditional demand system (nested CES) allowed recovery of product-nest-specific Lerner indices. These demand parameters allowed decomposition of the residual of the revenue function into the structural productivity term (ω_{it}), the implied firm-level price, and the bias induced through the biased

coefficients (reflecting a bias in the scale elasticity).[42] A bottom line of the study was a substantially different interpretation of the productivity effects of trade reforms. Because the relevant reforms acted simultaneously as both demand shifters and productivity drivers at the firm level, controlling for demand muted the productivity effects significantly. Revenue-based productivity measures confound demand and supply responses. This approach has since been used in the other international trade contexts. See, for example, Roberts et al. (2017).

Pass-Through A potential third approach has seen adoption only recently. It starts from the observation that demand and supply primitives predict the relationship between output and input prices. In many plausible conditions, they are positively correlated. Consider, for example, products where higher-quality versions (whose producers can sell for higher prices) require higher-quality inputs (whose producers must pay more to obtain). In this case, the biases tend to work against each other—one being in productivity's numerator and the other in its denominator—and in extreme cases can cancel out. While this is the topic of ongoing work, we want to highlight that there is value in integrating the insights from the control function with a formal treatment of the output-input price wedge error.

The presence of both output and input price variation in the revenue function error term introduces the pass-through from input to output prices (conditional on productivity). This setting is arguably the most relevant empirical setting, admitting both output and input price heterogeneity. This leads to a composite error term in the empirical specification of the production function:[43]

$$\tilde{\omega} = \omega + \epsilon + p - \beta_V p^V \tag{26}$$

This is the case De Loecker and Goldberg (2014) label the standard setup; it is also considered by De Loecker, Eeckhout, and Unger (2020). Even though output and input prices are not observed directly, one can rely on observables (e.g., market shares in De Loecker, Eeckhout, and Unger (2020)) that govern the pass-through to infer their relative

[42]The literature sometimes refers to the residual of the revenue function as "revenue based productivity." Note that, despite some commentary to the contrary, this is not generally equivalent to TFPR. TFPR is defined as the product of TFPQ (physical productivity) and price. While the revenue residual includes both of these elements, they need not enter multiplicatively, and the residual generally also depends on other parameters of the demand system.

[43]In the more general setup the same applies to input price differences for the fixed factor.

size. The key idea is that firms with identical productivity will pass on input shocks differently depending on their market share. This allows the price measurement error term to be backed out separately from the production function parameters.

Beyond Price Data: How to Compare Quantities? The issue of price heterogeneity highlights a more fundamental problem that plagues even analyses where physical output and input data are observed directly in the data. Namely, how does one compare units of differentiated outputs and inputs across producers on a common (productivity) scale?

An intuitive reaction might be to convert the physical data to revenue and cost data to let prices capture the quality variation.[44] While this makes intuitive economic sense, it poses a fundamental identification problem: what explains higher revenue per bundle of expenditure? Is it higher productivity, or differences in demand and competitive pressures, and how could these be separated? A recent set of work tackling this problem has emerged. Its solutions evoke those just discussed: introducing information from the demand side paired with a model of competition, or combining product-level price and quantity data and a model of product differentiation. The various solutions offered so far in the literature highlight that the specific context of the application is crucial. We will discuss, however, what we believe are some general insights that may prove to be useful for future work in this area. We conclude this section with a few suggestions.

This challenge resembles that faced when estimating a demand system under product differentiation. This is a common issue in IO; indeed, other chapters in this volume are dedicated to this. An early start on estimation of production functions in characteristics space was in Berry, Kortum, and Pakes (1996). This work used insights from the differentiated product demand models developed in Berry, Levinsohn, and Pakes (1995) and others.

More recently, De Loecker et al. (2016) propose another framework to handle differentiated products. They limit the dimensionality of the differentiation and derive a quality index that allows comparison of quantities across producers. The starting point is an empirically relevant case, where output quantity data and (deflated) input expenditures are observed.[45] The method rests on an important assumption that higher quality inputs are

[44]This solution is similar to using the wage bill to capture the differences in labor quality across producers (see e.g., (Fox and Smeets, 2011)).

[45]This framework also works when both output and inputs are recorded in physical units, and where the same

required to produce higher quality output. Denoting product quality by ν, the approach considers a production function of the form: $Q(\nu) = F(X^V(\nu))\Omega$. In other words, producing a unit of output of quality level ν requires inputs featuring the attributes necessary to achieve that level of quality.[46]

The main observation that De Loecker et al. (2016) make is that the mapping from input use to output use, conditional on productivity and product quality, allows identification of the production function's parameters. We illustrate the approach under a vertical model of quality. Let all consumers agree on a ranking of quality, and let the price be the sufficient statistic such that $\nu = v(p)$. The following regression

$$q_{it} = \beta_V x_{it}^V + \omega_{it} + v(p_{it}) + \epsilon_{it}, \tag{27}$$

can trace out the parameter of interest, β_V, provided that the unobserved productivity term can be controlled for. A control function approach can be used here, provided that the input demand equation is adjusted as discussed above. Let us entertain the thought experiment of having two products with identical prices, a control for productivity, variation in material use that maps into output use, identifies the production parameter (that is precisely the role of the measurement error in output, or unanticipated shocks to output). Conversely, holding fixed productivity, with identical material use, but different output quantities, the price variation will identify the function $v(.)$ capturing the quality variation component. The parameters are identified using:

$$\mathbb{E}\left(\tilde{\xi}_{it}(\beta,\gamma)\begin{bmatrix} x_{it-1}^v \\ p_{it-1} \end{bmatrix}\right) = 0. \tag{28}$$

In the absence of product differentiation, the last moment can be dropped and we identify the production parameters. Now, let firms produce different varieties reflecting different attributes. These attributes are costly to add (compare low-end with high-end cars for example), but they do not necessarily materialize in 1:1 output price differences due to market structure, demand and other environmental conditions. The second moment identifies the parameter that controls the degree of product quality variation through price variation (in this case). Put differently the function $v(p)$ allows to compare products across firms in terms of perceived quality. The moment that identifies this relies on a well-known instrument in empirical IO. Demand estimation has to confront the endogeneity

index can be used to control for unobserved quality.

[46] If product differentiation is solely tied to fixed expenses (e.g., advertising) of an otherwise homogeneous good, we are back to case B.1.1.

of prices, and the gold-standard to date is to find a cost shifter that moves around the price, but is orthogonal to any demand error. It is precisely the correlation between the productivity shock and the contemporaneous price that would violate the moment condition, and therefore lagged prices are used. What fundamentally solves the unobserved product differentiation variation, is the correlation between output and input price that informs us about the degree of product quality variation.

Finally, it is instructive to consider a special case where the log price enters linearly with a coefficient of one. The implication is that input and output prices are perfectly correlated across all firms. That puts us back in the standard case where we relate revenue to expenditures. De Loecker and Goldberg (2014) highlight this as the one knife-edge case for which we can identify the production function and recover productivity using revenue and expenditure data (note, not revenue and input quantity data). There a few other auxiliary assumptions needed to get there: constant returns to scale and a common proportional input price error (i.e., if a firm has 5 percent higher wages, it also pays a 5 percent higher material price).

5.3.3 Impact on the Coefficients of Interest

As made clear above, the literature has wrestled with the multiple conceptual problems that can arise in production function and productivity estimation. However, there really has been no systematic analysis of how the results of interest actually change when the *correct* parameters are used, as opposed to the biased ones obtained in more naive estimations. For example, economic theory predicts that in most situations, the coefficient on the flexible input X^V will be biased upwards because more productive producers, all things equal, will produce more and thus require more inputs. So, is that the typical finding in the literature? If so, are the biased and unbiased estimates significantly different from each other, and do they lead to different conclusions regarding the research question at hand?

We do not wish or attempt a meta-analysis of the vast literature. However, the common wisdom is that *they change according to what economic theory predicts*, and we believe this is a fair assessment of the literature's results. But what is lacking, from our perspective, is a more systematic analysis of the uncertainty around these correct parameter estimates and their differences from their naive counterparts. This comparison is of course complicated by the fact that the production function is inherently multivariate in nature, making it

harder to see an immediate impact on the results of interest. An important piece of information in this regard is the correlation between the various inputs, as this ends up having a big impact on how the potentially corrected parameters values play into the results.

An illustration of an analysis where the effects of the econometric correction are explicitly mentioned is Collard-Wexler and De Loecker (2016). The authors construct confidence intervals around the parameter of interest (in their case the productivity premium of minimill steel producers, and the various decompositions terms) based on the standard errors obtained on each of the approaches. They find that correcting for unobserved prices and productivity shocks yields both statistically and economically different results. This is important given that the methods deployed to undo the endogeneity concerns have a tendency to lead to somewhat higher standard errors given that they tend to be more data-hungry.

5.4 Multi-Product Production

Product differentiation highlights the presence of a product space. In contrast to the literature on market power using demand analysis (see XX and XX in this Handbook), the product space has traditionally not received much attention in productivity analysis. Most studies have instead focused on the producers (leaving the underlying product space for what it is) or sometimes single-product producers.[47]

It is well-known from older, theoretical literatures on multi-product production (see Diewert (1973)) and multi-product cost function analysis (see McFadden (1978)) that these settings do not yield simple mappings from inputs to a single output due the presence of joint inputs and economies of scope. Formally, there does not exist a "production function," but rather a "production possibility frontier" or "transformation function." While applied researchers were of course aware of the presence of the product space and variations in product mix across producers, the focus was on obtaining productivity measures at the producer level. A major reason for this is the lack of data on products produced or, more often, the allocation of inputs across various products. Production data are typically reported at the establishment or firm level; even if product-level breakdowns are reported for outputs (which they are often not), rarely are reported inputs similarly apportioned.

[47] See e.g., Syverson (2004a) for an application to a single product industry (ready-mixed concrete).

This single-product focus, or the aggregation to the producer level as discussed above, eliminates the presence of economies of scope, joint inputs, and the like.[48]

The impetus for a renewed focus on products came for a large part from the empirical trade literature and its use of product-level customs administrative data. An early treatment of multi-product firms, Eckel and Neary (2010), describes theoretical features of multi-output production. Bernard, Redding, and Schott (2010) further demonstrate the importance of multi-product firms in the U.S. manufacturing sector by showing the importance of this dimension in determining industry-wide patterns of production, efficiency, and trade. Using a fairly aggregate notion of a product (5-digit U.S. SIC code), they show that multi-product firms account for 40 percent of firms but 87 percent of production. Goldberg et al. (2010) report similar numbers for Indian manufacturing, albeit using a more disaggregated definition of products (roughly 1,800 unique product codes (CMIE)). This fact has been further confirmed throughout the literature in a wide range of economies with differing average income levels. Ignoring multi-product firms would leave out the majority of economic activity, and not treating their multi-product nature explicitly in productivity analysis risks drawing incorrect conclusions about the nature of production within and across industries.

Productivity analysis of multi-product firms poses substantial measurement and methodological challenges. As mentioned above, even if product-level output is observed, inputs are seldom broken down by products. Now, as shown by Diewert (1973), allocating inputs across products is not required when a production possibility set and the associated transformation function are considered. However, the standard problem of product differentiation now also plagues this problem, making the early theoretical literature on transformation functions not directly applicable. Furthermore, just like firm entry and exit, product adding and dropping are endogenous responses to firm- and market-specific shocks. This makes the set of products produced by a firm an endogenous object, consisting of potentially differentiated products produced using a bundle of inputs, and a particular level of efficiency. The latter can in theory vary across products within a producer, although this has not been the predominant theoretical starting point in the IO theory literature on multi-product producers.

These challenges aside, researchers have made strides in treatment of multi-product

[48]Formally, this does not rule out economies scope in the form of spreading fixed costs (absent from the production function) across products, or applying superior managerial capabilities across product lines. It does, however, require the production function to be common across products.

producers. One set of work starts from a product-specific production function and aggregates explicitly to the firm level, where output and input data are observed. This benefits from using product-level information explicitly, but also rules out important aspects of multi-product production that could generate productivity benefits.

Some micro data, often customs-related, offers extra product-level information within firms. Because of their inherent link to tradability, such data is typically limited to the manufacturing sector. We will discuss how the existing literature has used such data along with economic theory to apportion firms' inputs across their multiple products. The presence of imperfect competition and product differentiation interact with this input allocation problem. Our discussion is brief and selective and makes clear this is very much a literature in progress. We organize the discussion as follows. First, we present what can be thought of as a standard approach of aggregating (certain) product-level data to the firm (or plant) level.[49] Second, we point the reader to ongoing work that considers the empirical analogue of the multi-product transformation function in the presence of homogeneous products under perfect competition. Third, we contrast the multi-product production and cost function and discuss a special case that allows for economies of scope and multi-product production while permitting product quality variation and imperfect competition.

5.4.1 Allocation of Inputs to Products

We briefly describe the popular practice that turns the multi-product data into producer-level production, hereby eliminating economies of scope and joint inputs. One starts from observing product-level output data (typically revenue data) and producer-level input data. The literature has considered a variety of product-level shares to aggregate product-level inputs to producer-level input data. The shares range from simply using the number of products, to using shares directly proportional to product revenue shares (De Loecker (2011a), Foster, Haltiwanger, and Syverson (2008) and Collard-Wexler and De Loecker (2015)). Implicit in these aggregation strategies is that the (unobserved) productivity term only varies at the producer level, and not across products within the firm. This is an economic rather than an econometric assumption, and implies that the researcher can only make statements about productivity differences across producers, excluding the

[49] In some applications plants may be less likely to produce multiple products, which can sometimes aid in eliminating the multi-product challenge. However, this is not generally the case, so we do not distinguish between firm and plant aggregation here.

perhaps relevant product dimension.

Explicit Aggregation from Product to Producer level In light of our notation, this means that we start from a product-level production function, where j denotes a product, for a given producer in a given time period:

$$Q_j = (X_j^V)^{\beta_V} (X_j^F)^{\beta_F} \Omega \tag{29}$$

The assumption that productivity is common across products, and that inputs can be allocated to inputs in a neutral way (i.e., $X_j^H = a_j X$ and $\forall H$), allows to aggregate to the producer level using

$$Q = \sum_j Q_j = \sum_j a_j^{\beta_V + \beta_F} X^V X^F \Omega \tag{30}$$

Depending on the application, the choice of the share a_j has been either the number of products or the output share ($a_j = \frac{Q_j}{Q}$, where it is understood that this has been done using product-level revenues in most settings). Most researchers have either imposed constant returns to scale ($\beta_V + \beta_F = 1$), which eliminates the extra term, and gives rise to the well-known producer-level production function equation (e.g., Foster, Haltiwanger, and Syverson (2008)). In the case where this assumption is not imposed, additional care needs to be given to this term. For example, in the case where the share is simply related to the number of products ($a_j = J^{-1}$), the production function specification includes an additional term: the (log) of the number products. In both cases, this effectively puts us back in the producer-level analysis discussed above, with the slight wrinkle that when returns to scale are not imposed, that the aggregation approach requires to think through the potential endogeneity of the number of products.[50]

5.4.2 Estimate Transformation Function (A.2)

In terms of our classification table, this approach resides in $A.2$, a perfectly competitive environment.[51] We illustrate this with the application of Dhyne, Petrin, and Warzynski (2014) studying production in Belgian bakeries producing bread and pastries. A crucial assumption is, again, that these two products are homogeneous and are both produced

[50] See De Loecker (2011a) for an application of this approach.

[51] In theory the approach could allow for imperfect competition in the output market, but it would have to confront the additional sources of input demand variation across firms when controlling for productivity. So we keep it under perfect competition for now, to insure validity of the entire approach including implementation.

by every bakery in the data – denoted (in logs) by q_{ijt} with $j = \{1, 2\}$. The data report quantities produced for each product and total input use, distinguishing between variable and fixed factors of production.

The main estimating system of equations is obtained using the empirical analogue to Diewert (1973) and Lau (1976). It is given by, where we drop the producer and time subscript:

$$q_1 = \gamma_1 q_2 + \beta_{V1} x^V + \beta_{F1} x^F + \omega_1 \tag{31}$$

$$q_2 = \gamma_2 q_1 + \beta_{V2} x^V + \beta_{F2} x^F + \omega_2 \tag{32}$$

where now $J + J \times 2$ parameters must be estimated (6 in the case of the two products): the coefficients (β_V, β_F) for each product, γ_1, and γ_2. These last two coefficients capture the nature of multi-product production, and the approach rests on the assumption that $\gamma_j < 0$ and $\forall j$.

This a very appealing and quite general setup, though an additional econometric challenge emerges aside from the assumptions stated above. In every model discussed in this chapter so far, there is only one unobservable productivity term. Now, there are J of them: all the productivity shocks $(\omega_{i1t}, \omega_{i2t}, ..., \omega_{iJt})$. Another endogenous regressor, $q_{i,-j,t}$, shows up in the regression. Again this is of dimension $(J-1)$. Additional data and assumptions are needed to identify this system. Not only does one need to instrument for the additional endogenous regressors, one must revisit the conditions allowing inversion of the input demand equations in the presence of multiple unobservables. The literature is still very much grappling with these issues. Dhyne, Petrin, and Warzynski (2014) take a first step by considering a two-product case (bread and pastries) and combining the OP and LP proxies of investment *and* materials to control for the two unobserved productivity terms. Thus

$$\begin{pmatrix} \omega_{i1t} \\ \omega_{i2t} \end{pmatrix} = h_t(k_{it}, m_{it}, i_{it}), \tag{33}$$

where $h_t(.)$ is now a bijection. This further highlights the additional restrictions required of the control function approach in this scenario. Note that the dynamic panel data approach discussed above has the potential to free up the invertibility problem while allowing imperfectly competitive environments, provided reliance on a linear productivity process and a homogeneous product setup. This seems an attractive approach going forward.

5.4.3 Product Differentiation and Imperfect Competition (B.2)

In Section 5.3.2 we discussed the approach suggested by De Loecker et al. (2016) under the single product assumption. The main insight is that projecting quantity produced on deflated inputs does not provide information about the technology parameters if products are differentiated (if product quality varies across producers, in their wording). The unobserved input price variation introduces a bias. The proposed control for unobserved input prices is based on the notion that output quality is related to input quality, conditional on controlling for productivity. The paper goes further by using the product-level production function estimation step as an input to recovering input allocations across products.

The main assumption is that there exists a product-specific production function but not a *product-firm-specific* production function.[52] This rules out physical synergies across products as a source of economies of scope. It does, however, not rule economies scope in the cost space. A multi-product firm can, say, spread fixed costs, or superior managerial capabilities across products or leverage bargaining power to obtain discounts on purchased inputs. Formally, these sources can generate economies of scope – in a two-product example, while $Q_1 = F_1(X_1)\Omega$ and $Q_2 = F_2(X_2)\Omega$, the cost function $C(Q_1, Q_2) \leq C(Q_1) + C(Q_2)$.[53]

Let us run with the bread and pastries case of Dhyne, Petrin, and Warzynski (2014) to illustrate the approach of De Loecker et al. (2016) (they consider a more general case), and let there be a product-specific production function, each under constant returns to scale. The approach starts with the presumption that there exists a sample (albeit selected) of single product producers allowing to estimate the production function parameters for each product (giving estimates of β_{Vj}, β_{Fj}).[54] A second important assumption, and a

[52]Formally, the firm's production function is separable across products.

[53]See De Loecker et al. (2016) for more discussion. Consider the following analogy. Most of us have assembled furniture from Ikea. The production function $F_j(.)$ is the manual (blueprint) for assembling a particular piece. Someone who is mechanically inclined can probably put any particular piece together at a lower (time) cost; that is, they will have a productivity advantage that spreads across different pieces. This across-product efficiency advantage is reflected in Ω.

[54]Note that this framework does allow for the endogeneity of multi-product status that can arise if more productive firms add products over time. This selection bias in the single-product estimation step can be addressed in a similar way to the exit-based selection in Olley and Pakes (1996). There is a sense that the time series dimension of the data will matter, but of course, it can easily rule out certain production processes to begin with. For example, the production process of gasoline and diesel – see e.g., Burkhardt (2019). This is a sharp contrast with the approach of Diewert (1973) where all producers need to produce both products in order to characterize the transformation function.

difference with Dhyne, Petrin, and Warzynski (2014), is that De Loecker et al. (2016) follow the theory literature on multi-product firms and model productivity as a firm-level characteristic, and not product-firm specific.[55]

We can now revisit the two product-level regressions (where we drop the firm and time subscripts):

$$q_1 = \beta_{V1} x_1^V + \beta_{F1} x_1^F + \omega \tag{34}$$

$$q_2 = \beta_{V2} x_2^V + \beta_{F2} x_2^F + \omega \tag{35}$$

The main difference here is that inputs are attributable to products, implying that there exists a production function per product.[56] The input allocation problem implies that x_j is not observed, however, given this assumption, we can express it in terms of the share of product j in the input's total input expenditure:

$$\exp(\rho_j) = \frac{P_j^H X_j^H}{\sum_j P_j^H X_j^H}, \tag{36}$$

for $H = \{V, F\}$. This states that the allocation is done in cost space, and it is common across all inputs.[57]

Collecting all observables on the left hand side we get the following system of 2 equations with what look like 3 unknowns (ρ_1, ρ_2, ω):

$$q_1 - \beta_{V1} x^V - \beta_{F1} x^F = \rho_1 + \omega \tag{37}$$

$$q_2 - \beta_{V2} x^V - \beta_{F2} x^F = \rho_2 + \omega \tag{38}$$

It looks like we are one equation short to identify the objects of interest. However, note that $\sum_j \exp \rho_j \equiv 1$; i.e., all costs are attributable to products. This procedure then delivers, for each firm and time period, the input allocation shares, ρ, as well as firm-specific productivity.

[55] See e.g., Eckel and Neary (2010).

[56] Note that in De Loecker et al. (2016) this is actually more general: total expenditures on factor are attributable to products.

[57] De Loecker et al. (2016) acknowledge that inputs are measured in dollars (potentially reflecting quality differences across products and firms), as reflected by the presence of (log) expenditures e^V, e^F rather than physical inputs as in this illustration. They do, however, restrict the input allocation to be factor-neutral.

Illustration To keep the algebra simple, imagine that both production functions are characterized by the same constant returns to scale (CRS) production technology ($\beta_{H1} = \beta_{H2}\ \forall H$). Productivity (in levels) Ω is obtained using

$$\sum_j \frac{Q_j}{X}, \tag{39}$$

where $\ln X = \beta_V x^V + \beta_F x^F$. This simply implies that $\exp(\rho_j) = \frac{Q_j}{\sum_j Q_j}$. In other words, in this illustration with a multi-product homogeneous good producer, under a common technology with CRS across all products, facing common input prices, the inputs are allocated across products depending on the quantity shares. De Loecker et al. (2016) nests this as a special case, and this illustration highlights the important steps and requirements. However, it is more general in that it can accommodate 1) product-specific technologies with unspecified returns to scale, 2) input price variation, and 3) product differentiation of the type that can be described by an index. The key assumptions are that single-product producers can deliver the output elasticities, costs can be assigned to each input by product, and that the input allocation shares are product specific (and not factor specific).

Recent work has adopted an alternative approach using first order conditions across all inputs, paired with an explicit restriction on the demand system and an underlying model of competition to recover the input allocation across products – see. Valmari (2016) and Orr (2019) for such an approach.

5.5 Cost versus Production Functions

In this section, we briefly remind the reader how production and cost function estimation are related. We also consider the differences and how those may translate into deciding between them. There is a long tradition of using cost functions to estimate economies of scale and scope (for multi-product production) as well as to evaluate the productive efficiency of individual producers and their aggregates. For an early treatment of cost and production functions see Walters (1963). Nerlove (1961) provides a clear discussion on the challenges facing production function analysis and advantages of using the cost function instead. The opening chapter of an earlier volume of this Handbook, Panzar (1989), is devoted to the analysis of cost functions and implications for the theory of industry structure, scale and efficiency. Reiss and Wolak (2007) cover some further issues regarding empirical implementation of cost function estimation.

Deriving the Cost Function Starting from our production function and considering a static cost minimization problem (equalizing the marginal rate of technical substitution to the ratio on input prices), we obtain the following cost function in logs:

$$\ln C_{it} = c_0 + \frac{1}{\beta_V + \beta_F} q_{it} + \frac{\beta_V}{\beta_V + \beta_F} p_{it}^V + \frac{\beta_F}{\beta_V + \beta_F} p_{it}^F + \frac{1}{\beta_V + \beta_F} \omega_{it} + \epsilon_{it}^* \tag{40}$$

The same simultaneity and selection biases are present in cost function estimation. The former comes through the correlation of output and unobserved productivity shocks. The same techniques used to estimate production functions can in principle be used to deal with this, including the control function and dynamic panel approaches. The cost function form also highlights the necessity in this approach of observing factor prices in the data (the derivation here assumes competitive input markets).

Another well known challenge of cost function estimation arises in the requirement that the (log of) total economic cost of production (C_{it}) is observed. This has been a recurring topic throughout this chapter, and in the IO literature more broadly: do we think we can credibly read this item from the data, or not? One's willingness to consider the cost function approach depends on the answer to this question, as well as the willingness to accept the additional assumptions required to derive a cost function, like static cost minimization across all factors of production.[58]

An early cost function estimation literature leveraged institutional details of the industry under study to deal with econometric challenges. For example, Nerlove (1961) studies returns to scale in U.S. electricity supply. Regardless of the regulated price, output and productivity are correlated across production units though the price mechanism of regulation that aggregates individual supply curves and clears the market. A subsequent literature relied on cost functions in this way to study allocative efficiency and market power.

Rightly or wrongly, the literature has drifted away from cost functions and toward production function estimation. Cost functions hold some strong advantages. Accounting data are typically expressed in expenditure terms rather than quantities; cost itself is an expenditure, and estimating a cost function does not require input quantities to be observed. Economies of scale and scope are more easily handled conceptually with cost functions: see our discussion on the challenges of multi-product production analysis using

[58]Some users have adopted short-run cost function estimation by conditioning on fixed or quasi-fixed factors of production. This imposes the static cost minimization assumption only for the variable input, x^V in our notation, and implies that fixed input quantities be included on the right hand side.

the output-input space. However, a cost function approach still requires assumptions on economic behavior, essentially cost minimization and an assumption on how factor markets clear. Further, one must still observe output quantities to estimate returns to scale. Perhaps the greatest hurdle is that producer-level factor prices are required to estimate a cost function at the micro level, unless this variation is assumed away, of course, or can be captured by relevant controls.

Perhaps another "sociological" factor at work behind this shift from estimating cost to production functions were arguments accompanying the New Empirical IO wave that called the use of *any* cost data into question.[59] The biggest concerns centered on payments to capital services, but the points applied more broadly. These misgivings were based on solid arguments and concerns. However, it is fair to say that a major thrust of the literature's responses to such concerns was in some ways the most extreme possible: to estimate production costs without cost data. The practice involved obtaining marginal costs by inverting observed product prices through a first-order pricing condition based on assumed firm pricing conduct and estimated demand elasticities. This solution didn't seem to fully grapple with the fact that the revenues and factor expenditures data of concern for cost function estimation were the same source data for the alternative, production function estimation. Complaints about one applied just as well to the other. In addition, the very same discredited cost data was often used to support the structural model used to recover the marginal cost and markup estimates from this approach.

The (renewed) interest in the productivity residual, in part fueled by the prominence of models of firm heterogeneity in international trade and macro models (an important example in the context of trade is Melitz (2003)) and the arrival of the rich producer-level micro data, further pushed the development of econometric techniques to estimate production functions. The analysis of cost functions, and recovering objects such as economies of scale and scope (in the case of multi-product production), were relegated to the sideline.

We propose that researchers put the cost and production function approaches, and their required data, on an equal footing—namely, full of challenges (what empirical approach is not?) yet still useful.

[59] See Bresnahan (1989) for discussion.

5.6 Measurement and Specification Errors

Like in any applied field, productivity measurement is subject to a host of data and specification errors. To this point we have highlighted sources of errors that are derived from an underlying theoretical model (either supply, demand, or conduct related). The variation in the data thus reflects what economic theory suggests should move around output and input data, broadly defined. However, there are still other measurement and model misspecification concerns. While it is not practical to discuss all possible sources of measurement and specification error here, we highlight a few issues that we believe to be priorities: measurement error in capital and misspecification of the production function and the underlying productivity process.

5.6.1 Measurement Error

One type of measurement error that we explicitly allowed for in the theoretical frameworks above is error in recording output, ϵ. Recall that in the control function setup, this has an alternative interpretation of an unanticipated output shock that was not part of the producer's information set when input decisions were made.

Griliches and Mairesse (1995) gently warn productivity researchers of a host of other challenges, chief among them measurement error in inputs. Several studies outlined in the overview articles by Bartelsman and Doms (2000) and Syverson (2011) are clear testament to the issues that can arise. However, there has not been much attention to formal treatments of input measurement error. We focus here on the many potential forms of measurement error in capital, the input arguably the most prone to measurement error for all the reasons highlighted in Section 4.2. We keep sight throughout of also treating the simultaneity bias.

While the standard reaction in the presence of measurement error in a covariate is to look for an instrument, this turns out to be quite challenging given the presence of the unobserved productivity shock. A valid instrument for a mismeasured input must not just be correlated with the true underlying input, it must also be orthogonal to the productivity shock. Just as this can pose issues for the use of input prices to instrument for endogenous inputs to address the simultaneity bias (due to market power in factor markets, for instance), the same logic applies as an instrument for measurement error. The search for instruments has been further complicated by the introduction of the control function approach, which implies that a non-linear function in inputs is estimated in a

first stage, as highlighted in equation (14).[60] This limits the applicability of IV techniques to address input measurement error.

Collard-Wexler and De Loecker (2016) consider the standard setting (A.1) and study the impact of measurement error in capital. The fundamental distinction with other inputs is that capital's valuation is rarely linked directly to a transaction that would cleanly reveal it. The value of the capital stock therefore has to be computed by the researcher or taken from the accountant, allowing for a long list of potential sources of measurement error.[61] The heart of the approach is the observation that the source of the measurement error is likely to stem from computing the stock, therefore requiring appropriate depreciation factors, and prices. However, reported investment expenditures or alternative measures of the capital stock such as replacement capital are likely to be less riddled with measurement error, or at least different errors. They use this observation to suggest instrumenting the capital stock with lagged investment expenditures, or alternatively, replacement capital. (Their approach still builds a control for simultaneity bias.) Candidate instruments thus need to be orthogonal to the measurement error in capital, but they are allowed to be correlated with the persistent part of the productivity term. An interesting byproduct of this approach is that it has implications for the bias of the other factors of production, through the correlation across all factors of production.

After demonstrating the impact of measurement error in capital on standard approaches under A.1 using Monte Carlo analysis, Collard-Wexler and De Loecker (2016) apply their estimator to three distinct datasets covering the manufacturing industry of China, India, and Chile. Across all industries they find estimated capital coefficients that are about twice as high as those obtained using standard techniques (ACF in this case). This indicates that measurement error in capital is substantial under the maintained assumptions, and it will impact subsequent productivity analyses. More work is clearly needed that connects these findings to the specific sources of the error, e.g., heterogeneity in the assets across firms, depreciation heterogeneity, or through differences in the price of capital.

[60] In this setting, Kim, Petrin, and Song (2016) build on non-linear IV techniques to tackle measurement error in capital.

[61] Becker et al. (2006) compare the reported book value of capital (on the balance sheet) to the constructed one (built using the perpetual inventory method) in the U.S. Census of Manufacturers. They find quite a bit of disagreement across the two measures, further pointing to the difficulties in assessing the value of a producer's capital stock in any given point in time.

5.6.2 Model Misspecification

We briefly discuss three specification errors that pester empirical work: 1) omissions of productivity drivers in the underlying law of motion of productivity (introducing an internal inconsistency in the research design), 2) heterogeneity in the production function, and 3) incorrect functional form of the production function.

Productivity Process Suppose a researcher wishes to study the effect of innovation on productivity at the producer level. Can she first estimate the production function and recover productivity, then relate those recovered productivity estimates to measures of innovation? It turns out that this two-step approach is problematic in practice. It fails to recognize that the variable of interest, say R&D, should enter the first production function estimation step. To see why, imagine that one were to rely on OLS to estimate a production function. This would imply that input use does not respond to productivity shocks induced by the change in the operating environment. But that relationship is precisely the object of interest in the second step.

We can represent this lack of internal consistency by stating that any change in the operating environment, be it exogenous to the producer or a deliberate action denoted by $\mathcal{A}_{it}$, conceptually belongs in the productivity process:

$$\omega_{it} = g(\omega_{it-1}, \mathcal{A}_{it-s}) + \xi_{it}. \tag{41}$$

We use $t - s$, with $s = 0$ or $s = 1$, to denote that this productivity shifter can enter either at time t or in lagged fashion ($s = t - 1$). Omitting the action variable from the productivity process leads to biased coefficients and therefore invalidates the productivity analysis itself.[62]

This internal inconsistency plagues even sophisticated methods that deal with the endogeneity of inputs.[63]

[62]Note that the only way to claim that this omission does not bias the coefficients is to argue that input decisions are orthogonal to $\mathcal{A}$. But that seems at odds with the research question of if, and how, productivity is affected through changes in this variable.

[63]There is one approach that attempts to by-step this issue by considering the *augmented production function*, whereby the term $\mathcal{A}$ is directly added to the production function specification. This approach still faces the problem of endogeneity, however, not just for inputs but potentially for the action variable included as well (unless this is clearly an exogenous change in the operating environment). See De Loecker (2011a) for a detailed discussion of the various approaches.

This discussion also suggests that the dynamic panel data approach is much more restricted in dealing with endogenous productivity processes, and as highlighted above, it underscores the importance of linear productivity processes. The control function approach is on the other hand more flexible, but it requires researchers to think through carefully how the additional productivity driver enters the model. In other words, what time subscript does $\mathcal{A}$ enter the law of motion with?

Recent work has pointed out the severity of this inconsistency and offered ways to deal with it. Firms engage in a variety of activities to raise their efficiency over time through general investment, active R&D and innovation, and learning by doing. The omission of these variables in the productivity process can lead to drastically different conclusions. We revisit this in section 6 when discussing productivity drivers.

Technology Heterogeneity Throughout this chapter we have assumed there is an industry-wide production function under which all firms produce, and they only differ in terms of the efficiency whereby inputs are converted into units of output, as embodied in the Hicks-neutral TFP multiplier. That is of course a major simplification. There is relatively little work on the explicit technological differences across firms within a single industry aside from any productivity term. In practice, producers may have some scope to adjust their production function, selecting among a menu of technologies.

An example of a study that loosens this constraining assumption somewhat is Van Biesebroeck (2003). This work studies car production and models the technology adoption choice of producers (lean or mass production). The model is structurally estimated using micro-level data on production and technology indicators.

Functional Form The predominant functional form for production functions in applied works is Cobb-Douglas. While Cobb-Douglas is the first-order approximation to any production function, it does impose a strong assumption of a unitary elasticity of substitution across inputs. Further, in practice, researchers sometimes impose constant returns to scale upon it, which can be a severe restriction depending on the setting. More generally, holding to any set of parameters for Cobb-Douglas (or any functional form) over a sample can raise further issues. For example, for a range of questions related to secular changes in the aggregate economy, say changes in market power, failing to allow for time-varying production functions can lead to faulty inference.[64]

[64]See De Loecker, Eeckhout, and Unger (2020) and Syverson (2019) for a discussion.

A separate body of work has emerged that considers the Constant Elasticity of Substitution (CES) specification. This nests two popular specifications, Cobb-Douglas and the Leontief (fixed proportion) technologies. Using CES introduces an additional parameter to free up the allowed substitution across inputs. See Grieco and McDevitt (2017) for a recent study relying on a CES specification in the context of measuring the quality of care in the U.S. dialysis industry. The use of the Leontief form is attractive when modeling a particular production process at a reasonable disaggregated level; see Hendel and Spiegel (2014) on Israeli steel mills and De Loecker and Scott (2016) for U.S. brewing as examples.

Still more flexible is the translog production function, which is a second-order approximation to any general production function. The translog was introduced by Christensen, Jorgenson, and Lau (1973) and its use has continued in both production and cost function applications since then. See Feenstra (2003) for an example.

The introduction of the control function approach has at least partly paved the way for richer substitution patterns by considering production functions of the form $f(x^V, x^F; \boldsymbol{\beta}) + \omega$. For example, De Loecker and Warzynski (2012) and De Loecker et al. (2016) rely on industry-specific translog production functions, generating producer-time varying output elasticities. Gandhi, Navarro, and Rivers (2020) generalize the ACF approach under case A.1 and allow for a non-parametric production function using, essentially, a hybrid between the factor share approach and the timing assumptions introduced in the control function approach.[65]

The main restriction remains, however, that productivity is a factor-neutral production function shifter (i.e., Hicks-neutral), and that the coefficients of the production function are common across producers. Departures from Hicks-neutrality introduce substantial challenges for the identification of production functions; see, for example, Ackerberg and Hahn (2015), Balat, Brambilla, and Sasaki (2016), Doraszelski and Jaumandreu (2018), Li and Sasaki (2017), Fox et al. (2017), and Demirer (2020). That said, there is a vast theoretical and empirical literature documenting the importance of factor-augmenting technologies. The practice of outsourcing, the arrival of labor-changing technologies (through automation and robotization) has given rise to a significant body of work (on the edges of the field of IO) that relies on departures from the Hicks-neutral mantra.[66] Bresnahan, Brynjolfsson, and Hitt (2002) study the interplay of skill-biased technological change in

[65] Just like in the dynamic panel data approach, they forego the input demand inversion step by imposing restrictions on the nature of competition and demand-side heterogeneity.

[66] See, for example, the work of Acemoglu (2002) and Acemoglu (2003).

the U.S. and the impact on labor demand. In particular, they find that the combination of IT and organizational innovation impact labor demand. In the final section we discuss the importance of this line of work for the productivity literature going forward.

Finally, throughout this chapter we have explicitly treated the gross output production function. There is, however, a long tradition in the productivity field of relying on value added as an output measure. One argued advantage of value added is that intermediate inputs respond especially strongly to productivity shocks, therefore making the simultaneity bias very challenging to deal with. On the other hand, the conditions which a value added production function can straightforwardly be derived from a more primitive gross output production function are quite restrictive.

6 Productivity Analysis

A substantial literature studies productivity drivers arising from both producer-level actions and changes in the operating environment. The effects of these influences can be measured at the producer level or at a more aggregate market (or industry or economy) level. We find it helpful to organize the literature according to two dimensions: the source of the productivity driver and the unit of analysis.

Productivity drivers can be internally or externally sourced. The distinction depends on whether producers can potentially initiate the productivity effect through their own actions (and as such are internal drivers), or rather rely on a passive, exogenous process (external drivers). The latter is understood to capture technological change not explicitly modelled by the researcher, but could well be a result of political or economic interactions outside the scope of analysis (for example, trade shocks that affect an industry).

The second dimension is the level at which productivity effects are estimated, analyzed, and reported. The tradition in the micro productivity literature is to report across-producer moments of estimated effects of certain actions or changes in the operating environment. This analysis is frequently extended to compute implications for aggregate outcomes. In most studies done by industrial organization researchers, the aggregate has typically been a market or industry. In other fields, however, and indeed even within industrial organization more recently, studies of multi-sector and economy-wide aggregate outcomes are common.[67]

[67]We interchange the "market" and "industry" when referring to aggregate impacts in this discussion. The difference may be relevant in empirical analysis, however. Consider, as an example, an industry whose producers

There are connections between the different cells in Table 1 and the dimensions introduced above. For example, the assumptions used to measure productivity limit the scope of the productivity analysis and restrict the interpretation of the estimated effects. Under the workhorse model of single-product perfect competition (case A.1 in Table 1), for instance, policy changes or producer-level actions cannot differentially impact demand (and thus prices) across producers.

6.1 Producer-Level Productivity Analysis

We discuss the general approach to identify productivity drivers at the producer level and distinguish between passive and active drivers. We then detail a set of specific drivers that the literature has identified across a wide range of settings and datasets.

6.1.1 Identifying Producer-Level Drivers

A standard approach would first measure productivity using one of the approaches discussed in Section 5, relying on a panel of producers across one or more industries (typically in manufacturing), and consider a regression of the general form:

$$\omega_{it} = \delta \mathcal{A}_{it} + \text{controls} + \epsilon_{it}, \tag{42}$$

Depending on whether the driver of interest $\mathcal{A}_{it}$ is an exogenous or an endogenous factor, additional structure may be required to identify the coefficient of interest. We briefly describe these two cases.

Exogenous Drivers Perhaps the largest users of this approach are the trade and development literatures, where considerable interest lies in studying the impact of policy reforms on firm-level productivity.[68] A by now classic paper in this genre is Pavcnik (2003). The study uses the OP approach to estimate productivity of Chilean manufacturing firms and then relates these to changes in (output) tariffs. (Pavcnik (2003) also considers the aggregate effect by following the Olley and Pakes (1996) decomposition, which we discuss in Section 6.2.2.) A large body of empirical work has followed this approach. Extensions and modifications have been made by incorporating spillover effects through forward and backward linkages in the production or FDI chain. Other settings consider changes in

operate across a number of quasi-independent geographic markets.

[68] See Melitz and Trefler (2012) and De Loecker and Goldberg (2014) for more detail.

technology, regulation, or environmental policy. Note that it can often be the case in such settings that the exogenous driver of interest may be common across sets of producers.

While we do not attempt to list the specific conclusions of such a large body of work here, it is important to underscore that for these productivity regressions to be internally consistent with the use of certain productivity estimation approaches, they should explicitly state the conditions under which 1) productivity is measured, and 2) the assumption the two-stage approach using a linear specification is a valid approach. There are multiple modelling decisions to make. Does $\mathcal{A}$ enter with a t or $t-1$ subscript? Is the effect common and fixed across producers and time? And so on. These are all important dimensions to consider.

As discussed in the previous section, this raises the issue that even if $\mathcal{A}$ is exogenous, this two-stage approach of first estimating productivity, and then relating it to potential drivers cannot in general deliver the correctly estimated impact of the change in the operating environment – here, the parameter δ. An example is in De Loecker (2011a), where the interest lies in identifying the impact or reduced quota restrictions in the European textile market. The quotas are allowed to non-parametrically affect productivity following a generalized law of motion: $g(\omega_{it}, \mathcal{A}_{it-1}) + \xi_{it}$. Following an ACF-style approach, $g(.)$ is estimated alongside the production (and in this case, demand) parameters. This achieves two things. First, it allows for an internally consistent approach where productivity can move with quota protection variation. Second, it allows for heterogeneous effects of quotas across producers and time. The bottom line is that if the changes in the environment are part of the information set of a producer, they should be included in the model for forecasting productivity, $g(.)$.

Endogenous Drivers When the hypothesized drivers are endogenous, the same arguments in the previous section hold. Again, starting from equation (42) calls into question the two-stage approach. Now the extra complication is the fact that the factor $\mathcal{A}_{it}$ lies in the hands of the producer. As with exogenous drivers, a large empirical literature investigates endogenous drivers using a version of equation (42), though in this case additional assumptions or IV strategies are used to address the endogeneity of the choice variable. Still, most of this work does not incorporate ($\mathcal{A}$) into the productivity measurement framework. A separate approach is to allow the endogenous productivity driver to enter the production function directly (this is sometimes labeled the augmented production function approach). This raises the concern that, depending on the application, ($\mathcal{A}$) may not

always constitute an obvious factor of production that is potentially substitutable with other factors.

This is not merely a matter of econometric sophistication or style. It can substantially alter inference and the conclusions made about the importance of the productivity driver, both quantitatively and qualitatively. Only a handful of papers to date rely on such an approach (i.e., including the driver of interest directly in the law of motion) to study the impact of endogenous decisions of producers. De Loecker (2013) identifies learning-by-exporting effects by allowing past export experience to impact the productivity process. Including lagged export status in the law of motion of productivity then permits identification of heterogeneous effects of exporting on future productivity. The results indicate that these effects are substantially heterogeneous across producers of different productivity levels (coming from the interaction of ω_{it-1} and $\mathcal{A}_{it-1}$), but also that standard approaches fail to identify these effects. Doraszelski and Jaumandreu (2018) rely on this approach to estimate the productivity effects from innovation using a rich panel of Spanish producers with detailed information on R&D expenditures. They again confirm the dispersed R&D effects across the productivity distribution. Braguinsky et al. (2015) allow for heterogeneous effects by including ownership information in the productivity process, in the context of a unique setting in the Japanese cotton-spinning industry.

6.1.2 Sources of Productivity Differences

Research has established a causal relationship between many factors that operate at the producer level and productivity. We review some of the more prominent of these in this section.

Managerial Practices While once the object of much more speculation than evidence, the effect of managers and managerial practices on productivity has been the object of a flourishing literature over the prior 10-15 years. The early part of this research literature established robust correlations between a host of practices, see Bloom and Van Reenen (2007). More recently this work has expanded to more convincingly establish causal effects through a variety of methods, from model-based application of plausibly exogenous variation to randomized control trials.

The practices found to matter have varied somewhat across studies, though the literature overall has tended to focus on operational management. Therefore the studied practices tend to focus on how the producer coordinates its production process, like its

goal setting and evaluation; inventory management; policies on training, evaluation, and allocation of human capital; and management of customer relations. How the producer arrives at its main thrust of business (what products to make, what markets to operate in, the broad strokes of its production technology, and so on) has been taken as given. Work exploring this latter set of decisions, sometimes called strategic management, is still nascent.

Another related set of research still at a relatively early stage involve efforts to separate the effects of managers from management practices. This is an important question. Practices can in principle be taught and transferred across settings, making them less rival than managerial inputs embodied in the manager herself. Which of these is the more empirically powerful has implications for policies that might improve productivity as well as the potential scope for management-based productivity growth.

Unobservable Input Quality Standard TFP metrics "remove" the effects of labor and capital inputs on output. As such, standard measures of labor and capital have no direct effect on productivity. However, if there are unmeasured quality differences in these factors, given productivity's role as an output residual, these will be reflected in productivity. Because an enormous labor literature has documented dispersion in labor quality, and other research has found vintage effects or other important sources of heterogeneity in capital inputs, it is highly likely that in many settings some productivity variation reflects differences in factor quality that are missed in, say, worker-hours or measured units of capital stock.

Depending on available data, the institutional setting, and the research question, it may be possible and advised to construct productivity measures that explicitly adjust for factor quality differences. For example, a common practice is to measure labor inputs as the producer's total wage bill rather than employees or employee-hours. This is based on the notion that market wages reflect variations in workers' contributions to production (as discussed above in the context of the wage-productivity correlation). Even more detailed data on worker characteristics, such as the increasingly available matched employer-employee data, might allow even more precision in accounting for labor's marginal product.

Similarly, in settings where considerable detail is available about the attributes of the capital stock, more comprehensive quality-adjusted measures of capital inputs are possible. A related element of measuring capital inputs is the stock-versus-flow issue described above, with variations in utilization rates being one source of "quality" variation

in capital (stock) inputs that normally are attributed to productivity. Specific empirical settings may allow capital input measures to be adjusted for utilization.

Intangible Capital Intangible capital comprises capital inputs that are difficult or impossible to measure. As inputs, producers use them to obtain their output, but by their nature the influence of intangibles on output will be reflected in productivity rather than as a marginal product of an observable factor.

Producers have many potential types of intangible capital at their disposal depending on the setting. Examples include a producer's reputation, brand value, camaraderie among its workers, production know-how, installed customer base, trust with its input providers, and relationships with lenders.

Recent research has built a conceptual foundation for characterizing the attributes of intangibles in production technologies and their likely equilibrium consequences (Westlake and Haskel, 2017). Related work has attempted to build proxies in micro data for intangible capital stocks and empirically tie them directly to productivity, profitability, and other producer-level and aggregate outcomes (e.g., Saunders and Brynjolfsson (2016), Peters and Taylor (2017), and Crouzet and Eberly (2020)). Just as with variations in factor quality, to the extent that researchers can construct measurable proxies for intangibles, they can account for intangibles' influence on output and their connection with productivity at least partially severed.

The particular type and role of intangible capital could affect different aspects of productivity. Some types, such as production know-how, are likely to raise productivity through technical efficiency, TFPQ. Types like brand capital on the other hand are more likely to have demand-side influence, raising the price the producer's output will sell at on the market. While unlikely to influence TFPQ, the price effect will raise TFPR.

Firm Structure Productivity can be influenced by the organizational structure of a firm's production units. Varied studies have established connections between productivity and structural elements such as vertical integration, centralization, the number of industries the firm operates in, and the relative sizes of its production units in those specific industries. The mechanisms vary across settings, ranging from the elimination of vertical inefficiencies like holdup, to optimizing the managerial span of control, to more efficient transference of intangible inputs across production units, and beyond.

Product-Side Differences Many product markets exhibit nontrivial demand variation across varieties. These could be in either vertical or horizontal attributes. These differences in demand are reflected in dispersion in the prices that industry products are sold at.[69] Producers able to sell at higher prices than their industry cohorts will have higher TFPR, all else equal.

Furthermore, many producers engage in product innovation. These efforts could well be reflected in productivity measures, and this connection has been verified in many empirical settings.

Product-based variation in productivity raises an important point of interpretation regarding revenue- and output-based productivity measures like TFPR and TFPQ. TFPR, which combines both the ability of a producer's technology to create units of output and the price at which the producer can sell that output, can vary for reasons unrelated to the producer's physical efficiency. However, that should not be taken to dismiss TFPR as a potentially informative measure of the producer's social efficiency. Producers who can make goods of greater quality at the same cost as others are in fact more productive as a conceptual issue; they deliver higher utility per unit cost. If quality and price are correlated in output markets, then TFPR productivity measures will reflect this. TFPQ may not. In such cases, TFPR is a better metric of the firm's capabilities than TFPQ.

6.2 Aggregate Analysis: Resource (Re/Mis)Allocation

The arrival of micro data generated a wealth of information about producer-level differences in inputs and output as summarized by measured productivity differences. Models of firm heterogeneity were built around these facts. These were designed in part to speak to the influence of micro-level variation in explaining aggregate outcomes. With producer-level productivity measures in hand, a number of approaches were developed to *aggregate* these measures from the producer level to the level of interest, ranging from a market or an industry to an entire sector or even economy. This aggregating up can be done in a variety of ways, and the literature started from empirically appealing aggregations and associated decompositions to theoretical and structural approaches. Regardless the ap-

[69] It is obvious how vertical product differentiation might be correlated with prices. Horizontal differentiation can also be tied to price variation if there are differences in demand for specific horizontal attributes that vary across industry varieties. Suppose for example that industry producers operate in overlapping but distinct geographic markets, and transport costs are nontrivial. If demand for the industry's product varies across space, those producers located in or near high-demand markets are likely to earn a price premium.

proach we refer to the aggregation of producer-level measured productivities as $A_t(\boldsymbol{\omega}_t, \boldsymbol{s}_t)$, where the vector $\boldsymbol{s}_t$ captures the weight of an individual producer.

6.2.1 What Does Theory Predict?

The model laid out in Section 3 can be used to discuss the allocation and reallocation of economic activity in an industry. As discussed there, the substitutability parameter γ can be a reduced form stand-in for a more explicitly modeled market structure. When γ is large, substitutability is low and there is extensive productivity dispersion. When γ falls, productivity dispersion falls with it.

We also know that producer size is positively correlated with productivity. We can quantify this relationship within the model. The covariance between productivity (cost) and revenues is:

$$\begin{aligned} Cov\left[r(c_i), c_i\right] &= \mathbb{E}\left[r(c_i)c_i\right] - \mathbb{E}\left[r(c_i)\right]\mathbb{E}c_i && (43)\\ &= -\tfrac{1}{4\gamma}\left(\mathbb{E}c_i^3 - \mathbb{E}c_i^2\mathbb{E}c_i\right) && (44)\\ &= -\tfrac{1}{48\gamma}c_M^3 && (45) \end{aligned}$$

This covariance is negative, of course; higher-cost producers are smaller. Importantly, its magnitude depends on γ. As substitutability falls, so does the covariance between size and cost (or size and productivity). This is a manifestation of the fact that limits to consumers' ability or willingness to substitute across producers also limit the response of market share to productivity differences. On the other hand, as substitutability becomes perfect ($\gamma \to 0$), the covariance skyrockets; the smallest productivity difference leads to enormous differences in market share.

These intuitions from the model (which again generalize to the many other models in this class) lend themselves to discussions of the effect of competition on the allocation and reallocation of economic activity across heterogeneous producers in a market. If competition is interpreted as the equilibrium amount of substitutability in the demand system (and this is a reasonable interpretation, as substitutability is embodied in the slopes of firms' residual demand curves, and those slopes are also a measure of the amount of market power firms have), then the above implies more competitive markets have more skewed market share distributions, all else equal. Furthermore, if we think of the comparison as being for a particular market in two different periods, the model implies that an increase in competition raises substitutability and shifts market share toward

higher-productivity producers. The implied dynamic reallocation reinforces the static allocation.

The specific mechanisms that tie competition to substitutability are manifold and have been the object of many studies. They include trade, transport, or search costs. Examples of studies include Syverson (2004b), Syverson (2004a), Goldmanis et al. (2010), Crouzet and Eberly (2020) and Autor et al. (2020).

Note that these insights about allocation and its associated implications for reallocation tie greater concentration to more competition. The more competitive is a market, the more skewed is its market share distribution. The largest, most efficient producers are the most dominant. While many might worry about concentration leading to greater market power, the opposite holds here. It can be shown in the example model above that profits fall and welfare increases when substitutability and concentration grow (Syverson, 2019). To be clear, there are other classes of models where concentration and market power are positively correlated. Here, however, the opposite holds, and there is evidence that—perhaps especially in settings where producers have heterogeneous costs—the predictions hold up in the data.

6.2.2 Empirical Work

The empirical work studying aggregate outcomes using micro-level productivity estimates can be broadly classified into two main approaches: 1) share-weighted sums of producer-level productivity, and 2) measures of productivity dispersion. Both approaches are very much related in terms of how they inform us about the sources of aggregate productivity gains and potential changes in losses of output due to inefficient production. The detection of the sources of aggregate output loss is a classic question in economics, with diverging traditions across multiple fields. In IO, this pertains to the presence of market power (in all shapes and forms, see the chapter in this Volume by Asker and Nocke), though typically "aggregate" refers to a single industry or market. In this section we focus specifically on how technology, competition policy, and market power affects the efficiency of productive resource allocation.

Decomposing Industry Aggregate Productivity Since the seminal work of Baily et al. (1992) and Olley and Pakes (1996)), researchers have constructed industry- or sector-level aggregates as a share-weighted average of producer-level productivity values – i.e., $A_t = \sum_i s_{it}\omega_{it}$. The weights s_{it} are typically the share of a producer's sales in total

sales of the industry or sector, though sometimes input (e.g., employment) shares are used instead.

This aggregate is in itself not the only object of interest, though in practice it often closely tracks observed aggregate series. What researchers are often most interested in is the decomposition of this aggregate into *within* and *between* components. While very simple transformations of the data, they can be quite powerful in terms of describing how the aggregate performance of a collection of producers evolved over time. They describe what amount of change in the aggregate arose due to a common within-producer evolution versus a reallocation of activity away from relatively unproductive units and toward relatively productive ones. Economic theory can explain how both mechanisms can operate in principle, though the between component is especially tied to the operations of markets. Such reallocation is precisely what economists think well-functioning markets should do. Some of the studies we present below indicate how deregulation and the process of creative destruction initiated by technological change drove favorable reallocations of this type. On the other hand, we also discuss how market power, say through the presence of cartels, can hamper this process. The resulting misallocations create welfare losses because the industry is less productive in aggregate than it could be under a more efficient allocation. (That is, its producers could jointly produce more total output from its current aggregated inputs, if those inputs were better allocated.) This is distinct from the traditional deadweight loss coming from quantity distortions.

Before our detailed exploration of decompositions of productivity aggregates used in the literature, it is important to note a key conceptual point made in Petrin and Levinsohn (2012). Under general conditions, the growth in a share-weighted average of producers' productivity levels in an industry is *not* the theoretically correct measure of industry productivity growth. In essence, the difference is sourced in the fact that across-producer gaps in measured productivity need not equal gaps in marginal products. This raises questions about whether the within-between decompositions used in the literature accurately quantify the relative contributions of producer-level changes to aggregate productivity growth. Petrin and Levinsohn (2012) present an empirical approach that conforms to a first-order approximation to theoretically defined aggregate productivity growth. While perhaps the most preferred metric in many situations, its implementation requires researchers to deal with what can be delicate empirical issues. Perhaps due to a combination of this factor as well as familiarity with and comparability to the "accounting-type" share decompositions, most researchers to this point have continued to use weighted-sum decompositions in em-

pirical applications, despite the conceptual mismatch. A further convergence in theory and practice along this dimension would be welcome.

We first consider one of the more popular decompositions in the literature, that proposed by Olley and Pakes (1996)). It decomposes aggregate productivity into an unweighted average productivity and the covariance of producer-level market share and productivity:[70]

$$A_t = \overline{\omega}_t + cov_t(s_{it}, \omega_{it}). \tag{46}$$

Inspecting the time-series patterns of these two objects during a period of interest, say during a trade reform, points to the relevance of reallocation of resources in the economy. These decompositions have been applied to a variety of settings, across a wide range of regions in the world.

A major stylized fact emerging from these studies is that movement in this covariance is important. It can explain anywhere from twenty to forty percent of the total change in share-weighted average industry productivity across a wide range of studies and settings. In the context of international trade, this suggests that a large part of the productivity gains from opening to trade arises through a market-based reallocation process where more productive producers take over market share of less productive ones. This process is a central feature in modern trade theory models, as introduced by Melitz (2003).

Of course this still implies that the within-producer component is even larger. Standard drivers of growth such as technology adoption, R&D and other innovative activities (both product and process), and improved managerial practices are still acting and important. There is a sense in which some of the current literature, by focusing on the reallocation component, has left the within-producer component under-explored. Producer-level productivity regressions are therefore complementary to aggregate decompositions. However, few studies closely integrate both aspects to jointly address reallocation and within-producer effects simultaneously.

The OP decomposition is inherently a cross-sectional and empirical decomposition of observed average performance within an industry. There are, however, a variety of other decompositions that focus specifically on the time series dimension. For example, Haltiwanger (1997) propose a decomposition of aggregate productivity growth into within,

[70]Formally, the second term is not exactly the covariance, as it does not scale by the number of producers in any given cross-section. Thus entry and exit of producers will complicate this computation if done over time. This is precisely the motivation of Melitz and Polanec (2015) to introduce a correction for entry and exit, and develop a dynamic Olley and Pakes decomposition.

between, and net-entry components, where the within component is obtained by holding market shares fixed at time $t-1$.[71]

We now turn to discussing exogenous and endogenous drivers of the allocation of resources. Depending on the setting, the relevant focus may be on *reallocation* or *misallocation*

6.2.3 Exogenous Drivers: Reallocation

We sketch out in more detail two archetypal studies that use plant-level production data to study the impact of a plausibly exogenous change to the market operating environment. The first, Olley and Pakes (1996), investigates deregulation in the U.S. telecom equipment manufacturing industry during the 1960-1990 period. The second, Collard-Wexler and De Loecker (2015), looks at the arrival and diffusion of a new technology for steel production over 1963-2007. Each industry witnessed substantial aggregate productivity growth during the respective periods, with reallocation of activity towards more productive producers playing a large role. Both articles obtain plant-level productivity measures by estimating production functions for the industries' populations of plants over the relevant time periods. Both settings feature significant entry and exit (in line with the facts first documented by Dunne, Roberts, and Samuelson (1989)), markedly rising aggregate labor productivity, and substantial reallocations of market share away from incumbents. This Schumpeterian process gave rise to increased industry performance in both markets. The studies highlight the importance of obtaining reliable productivity measures. Furthermore, the usefulness of additional data on output prices and plant-level technology indicators becomes apparent in the analyses in Collard-Wexler and De Loecker (2015).

Deregulation Olley and Pakes (1996) begin by estimating the productivity levels of U.S. telecommunication equipment manufacturers before and during a period of intensive industry deregulation.[72] This industry underwent major restructuring beginning in the

[71] We refer the reader to Nishida, Petrin, and Polanec (2014), Melitz and Polanec (2015), and De Loecker, Fuss, and Van Biesebroeck (2018) for more discussion on the various decompositions.

[72] The authors state clearly their interpretation of the productivity numbers as measures of sales per unit input, rather than quantity per unit input, due to the lack of producer-level deflators. However, we note that because industry-level deflators were available, the associated industry aggregate productivity numbers could be interpreted as "true" TFP (i.e., a pure supply-side measure of technological efficiency). It is the across-producer variation that cannot be interpreted as productivity dispersion. This requires a careful interpretation of the reallocation analysis.

late 1960s due to two related changes. First, the substantial technological change brought many new products for both delivering phone services (digital switches, fiber optics, etc.) and for using the phone lines (fax, modem, etc.). Second, and as a particular focus of the paper, the deregulation that occurred after the Carterfone decision in 1968 halted the industry's history as a near monopoly due to the procurement practices of AT&T from its equipment-producing subsidiary, Western Electric. Waves of new products (modems, fax machines, etc.) and new companies came into the industry. Later, further regulatory changes forced AT&T to lease out its lines to other long-distance carriers, having a further effect on equipment purchases.

The essence of the analysis is summarized in Tables 1-4 of the article. In 1963, the industry had 104 firms operating 133 plants making 5.86 billion (1982 USD) of output with 137,000 workers. By 1987, there were 481 firms operating 584 plants that made 22.41 billion (1982 USD) with 184,000 workers. This is a three-fold increase in sales per worker. The research question is quite simple: what were the sources behind this massive productivity increase? The increase in industry firms from 104 to 481 suggests a potential reallocation of market share away from incumbents and towards more productive entrants. This is exactly what the decomposition used in the paper is designed to measure: to separate and quantify reallocation from efficiency improvements common to all industry producers. To do this analysis, OP first estimate a Cobb-Douglas (value added) production function in labor and capital, as described in Section 5.3.1. In addition to addressing the standard simultaneity concern, the estimation procedure takes into account the substantial entry and exit seen in the industry during this time. The procedure yielded a much larger estimate of the output elasticity of capital than did the (commonly employed alternative at the time) producer-fixed-effect estimator using a balance panel: 0.35 versus 0.06. This greatly affected the implied dynamics of productivity; more on this below.

The authors find that periods of high aggregate productivity growth were characterized by substantial reallocation, as reflected in a large positive covariance term. That is, the more productive producers accounted for increasing market shares over time. The economic interpretation is that the deregulation facilitated this shift, leading to an increase in aggregate industry performance. The precise reallocation mechanism is not discussed or further analyzed, however. The authors did not evaluate the importance of relying on productivity estimates that take into account the simultaneity and selection bias, but we revisit this below.

Technology Collard-Wexler and De Loecker (2015) study the creative destruction process induced by the (exogenous) arrival of a new technology. The setting is the U.S. steel industry from 1962 to 2012. This new production technology, the minimill, made producers efficient at smaller scales and allowed them to operate independently of production in upstream input markets (coal, lime, etc.). Despite a reputation otherwise, the U.S. steel industry over the period had one of the fastest productivity growth rates among manufacturing industries, behind only high-tech sectors like semiconductors. Steel producers made roughly the same tonnage of output in 2012 as they did in 1965, but with just one-fifth of the workforce. This enormous increase in labor productivity coincided with substantial industry TFP growth as well. The decomposition of share-weighted productivity indicated dynamics showed similar patterns to those in OP, with a reallocation of market shares away from incumbents and to entrants.

The study's goal was to precisely identify the driver(s) of this massive productivity increase. A number of candidates present themselves, not just minimill technology but also import liberalization, unionization shifts, and improved management practices. As with OP, the authors first estimate a production function to recover plant-level productivity estimates, then perform within-between decompositions of their weighted average to help identify the sources of industry productivity growth. Importantly, the data make apparent whether the producer employed the old (integrated mill) or new (minimill) production technology, allowing decompositions both within and between plants using either technology as well.

The initial analysis confirms the relative importance of reallocation. About one-third of the gains in the industry's weighted average productivity could be directly traced back to the reallocation of market shares from the old to the new technology. Interestingly, the competing candidate explanations above have little explanatory power, as their trajectories were similar across the two technologies.

Of course, the decomposition also implies that roughly 70 percent of industry productivity growth was due to other factors. The authors rely on product-level information in the data to distinguish between high and low quality steel products. The results reveal that head-to-head competition in low-quality segments induced selection among old-technology producers that led to productivity growth. But it wasn't all selection; incumbents saw big productivity gains too. In fact, surviving integrated mills experienced higher productivity growth rates than the rest of the industry, eventually catching up with the minimills and creating a substantial within-producer growth that accounted for a considerable share of

overall industry productivity gains. The results highlight the importance of the interaction of market-wide, across producers, effects and within-producer improvements.

The study also indicates the importance of correcting for the output price bias when producer-level deflators are available. The results from an analysis using expenditures rather than quantities were significantly different from those taking advantage of producer price information.

The sensitivity of the results to various measurement decisions raises a broader point. The components of the decompositions discussed above are stochastic objects. We encourage the practice of reporting the confidence intervals of decompositions' components to properly evaluate the econometric techniques used to combat the various biases that plague the production function estimation.

And while the archetypal studies above focus on manufacturing industries (as much of the early productivity literature did), there is by now a considerable body of work looking at reallocation and productivity in other sectors. Examples include retail (Foster, Haltiwanger, and Krizan (2006)), healthcare (Chandra et al. (2016)), and wholesale (Ganapati (2021)).

6.2.4 Endogenous Drivers and Aggregation: Market Power

A multitude of producer-level decisions can impact the aggregate performance of a market. However, one that is probably of most interest to IO scholars involves firms' responses to market power. This will affect not just the volume of market transactions, but also the associated production costs. Both factors ultimately influence the total surplus generated in the market.

Market power's welfare losses are well appreciated conceptually, and they are the primary motivation behind antitrust policy. Much of the empirical attention, however, has been trained on the well-known deadweight loss from quantity restrictions. Market power's influence on production inefficiency (i.e., costs that are too high) has received considerably less empirical treatment. We surmise that this is in part a natural consequence of the once onerous data requirements for quantifying productive inefficiency. Measuring such losses requires observing reliable measures of producers' marginal costs and ex ante knowledge of the firms and methods responsible for (potential) market power abuse. However, the arrival of the producer-level micro data that drove the productivity literature frees up this constraint. Researchers are no longer tied to traditional analysis reliant on assuming

homogeneous firms (for instance, in the traditional macro style using aggregate industry time series, like in Harberger (1954), or a fair share of the theoretical literature as well). One can now look inside an industry's total cost of production and relate it to productivity dispersions that may well interact with market power.

Notable exceptions are the work of Borenstein, Bushnell, and Wolak (2002) on market power in the California electricity market, and Asker, Collard-Wexler, and De Loecker (2019) studying the effect of OPEC's market power on productive inefficiency. Both studies leverage detailed cost and production data to build a supply model that allows counterfactual production allocations across producers in the absence of market power. This focuses on the welfare rectangle that captures the loss due to misallocation of production, given the quantity produced. This contrasts with the traditional focus on the welfare triangle that reflects the quantity distortion induced by market power.

Asker, Collard-Wexler, and De Loecker (2019) use observed cost and production data for every oil well in the world between 1970 and 2014. Assuming a Leontief technology specification for crude oil production, they derive marginal costs from data on production and comprehensive itemized costs. The starting point of the analysis is the fact that marginal costs are highly dispersed across producers, disproportionally lower in OPEC countries, and that reserves (that is, capacities) are higher in OPEC countries. These facts give rise to a substantial welfare loss from the actions of the OPEC cartel. Productive inefficiency is computed by characterizing the minimal cost allocation of production across wells that would yield the observed market-level quantity. The only wrinkle in the analysis is the finite resource extraction character of oil production. This is dealt with by using a theoretically derived sorting algorithm that dictates the order and sequence of wells called to produce. They compute a NPV welfare loss of about 750 billion (USD) due to the misaligned production allocation driven by OPEC.

While this study is specific to the global oil market, it suggests a similar phenomenon could be at work much more broadly. Given the ubiquity of productivity dispersion and the fat-tailed distribution of producer outcomes, there is considerable scope for market power to create welfare loss through misallocated production. A complicating countervailing consideration is the existence of product differentiation, and investigations into the issue will need to address this, but this is precisely why these studies of homogenous product markets (electricity and crude oil) are instructive.

6.3 Misallocation

The investigation of inefficient production allocations due to market power offer a convenient segue into a discussion of a broader set of research looking at productivity and *misallocation*.

There is a long tradition in industrial organization of studying factors that hamper industry performance, broadly defined. Factors that limit the optimal deployment of resources in an economy can be classified as either *primitives*, capturing technology, preferences, and environmental factors (with the exception of course of man-made environmental outcomes), or *behavior* and *policy*, capturing the actions of firms, governments and institutions, affecting economic outcomes. In the context of IO, the primitives are typically the particular technologies firms use to produce and how consumers value products. These can of course also entail different kinds of adjustment costs among producers or consumers. The behavior/policy category involves departures from optimal resource use due to the actions of producers, governments, or other institutions more generally. Market power features prominently in this category. Moreover, agency and governmental interventions in the market for corporate control and antitrust rules more generally, can equally play an important role.

It is useful to note that productivity dispersion per se is not sufficient for losses due to misallocation. Consider an industry with heterogeneous-productivity firms. Absent any frictions, the market equilibrium leads to an optimal deployment of resources where the marginal revenue product of (any) input is equalized across firms. This observation, which extends back to Lucas (1978), is the starting point of the macro literature on misallocation (see Restuccia and Rogerson (2008), Hsieh and Klenow (2009), and an overview by Hopenhayn (2014)). While this literature is not yet central to IO, the setting of Olley and Pakes (1996) fits it exactly. The ability to measure productivity allows measurement of the allocative impact of regulations.[73]

The empirical literature on misallocation grew from the seminal contribution of Hsieh and Klenow (2009). This gave way to the *wedge* (alternatively, *gap* or *friction*) approach that has been applied in many settings. The main premise of the Hsieh and Klenow approach is that a frictionless economy should see full equalization of inputs' marginal revenue products across production units. Dispersion thereof indicates the existence of

[73]The discussion in Section 5.3 underscores that the estimation of a production function requires an explicit treatment of the friction to credibly identify the productivity residual.

frictions in output or input markets that prevent the optimal allocation of resources. And indeed, Hsieh and Klenow find massive aggregate productivity losses due to misallocation. Specifically, if capital and labor inputs in the Chinese and Indian manufacturing sectors were reallocated to equalize marginal products to the extent observed in the United States, the implied TFP gains would be 30-50 percent in China and 40-60 in India. In other words, both countries could see enormous output growth without any additional factors required.

While the "equate marginal revenue products" intuition of the approach is extremely incisive, this approach requires a host of assumptions on conduct, demand, and production. Hsieh and Klenow (2009) assume that producers are monopolistically competitive and face CES demand curves, producing with identical Cobb-Douglas constant returns to scale production functions and facing identical input prices and no factor adjustment costs.[74] This follows the tradition of much of the empirical work in IO. As with all structural work, the devil is in the details; in this case the set of assumptions used to compute marginal revenue products and infer misallocation is the center of attention.

Adjustment Costs and Volatility The presence of adjustment costs and uncertainty about future productivity pose a challenge for this approach. If there are capital adjustment costs, the dispersion in the marginal revenue product of capital is no longer informative about misallocation. And once we allow for the fact that producers face uncertainty about their sales per input process, we also naturally obtain equilibrium dispersion of inputs' marginal revenue products. The presence of either or both of these issues break the prima facie link between observed MRP dispersion and resource misallocation. Asker, Collard-Wexler, and De Loecker (2014) makes exactly this point using both reduced-form empirical evidence and simulations from a fully specified model estimated by data moment conditions.

There is abundant evidence in the literature documenting volatility of productivity and the presence of capital adjustment costs. The implication is that in countries or industries with higher volatility, we should expect to see higher dispersion of inputs' marginal revenue products. Asker, Collard-Wexler, and De Loecker (2014) confirm this prediction in reduced-form evidence and show using a model that this can explain about 60 to 90 percent of observed dispersion.

This result has two important, and related implications. One, static measures of

[74]See Haltiwanger, Kulick, and Syverson (2018) for detailed discussion of this environment and the implications for interpreting findings using this approach.

distortions or misallocation are limited in their ability to detect productivity-detracting frictions that might be targeted by policy. Two, identification of distortions affecting the allocation of resources from time-series patterns in the data (or, put differently, the ones that impact investment decisions) are often preferable. In any case, linkages among productivity analyses of the type discussed in this chapter, volatility, and features of a market's operating environment are well worth analyzing further.

7 Looking Ahead

In this section, we briefly discuss some of the most active and novel portions of the literature. While these bodies of work are in their infancy, we expect that they will be busy areas of inquiry moving forward.

7.1 Market Power and Productivity Data

Recent work has put forward a framework to analyze market power (by measuring price-cost margins) using the very same data typically used to study productivity. This has the potential to further integrate the productivity literature with what one might perceive as the standard approach in IO, as well as to provide additional micro-founded measurement of a variety of performance indicators of interest to policymakers. We sketch out here the current interface of the market power and productivity literatures and summarize a few recent applications. This approach, combined with additional information and assumptions on factor and product markets, brings another tool to the applied researcher's toolkit for studying imperfect competition.[75]

The industrial organization literature's typical concern regarding market power is the quantity distortion. Firms with market power charge too much and produce too little, creating deadweight loss and reducing consumer surplus. While antitrust authorities across the world increasingly rely on demand analysis coupled with a conduct assumption to evaluate mergers and acquisitions, cartels, and other forms of market power abuse, the analysis of potential cost effects is currently less developed. This results in reliance on rather arbitrary guesstimates about whether cost effects would offset market power effects on pricing. IO researchers and antitrust practitioners justifiably spend a lot of effort mod-

[75]In reference to Table 1, we consider cases in row B, capturing models of imperfect competition, in either the product or factor market.

eling and estimating demand to predict merger effects acting through markups. However, there is little systematic modeling of expected productivity/cost effects. Productivity researchers could offer considerable value added here.[76]

There is another important connection between the market power and productivity literatures. Besides creating deadweight loss and lost consumer surplus from quantity restrictions, cartels and monopolies can also have allocative efficiency effects on productivity. Market power can skew the distribution of production among firms in a way that raises total industry costs. In the parlance, market power creates not just losses from "triangles" but "rectangles" as well. These rectangles can be large. Misallocation, as discussed in the previous section, has been extensively studied in the recent macroeconomic literature. This work has developed evidence that there is considerable misallocation of output across producers with different productivity levels. As a result, industry costs are higher than they need to be.

However, research looking at market power as a specific source of misallocation is still rather scant. One possible reason for this is the influential study of Harberger (1954) and its followers. Harberger concluded that the rates of return on capital across U.S. (manufacturing) industries during the 1920s were not sufficiently dispersed to generate any meaningful aggregate distortions attributable to market power. The intuition for this inference is that market power operates like a tax, where the implicit tax rate is reflected in the rates of return on capital (profits). If these rates are equal, then there can be no scope for misallocation of the incremental resource (production) unit. This analysis and its conclusion that market power scarcely impacts economy-wide outcomes became the default view held by many economists for decades. Harberger's focus on deadweight loss triangles has persisted in much contemporary work on market power. A recent literature started to challenge this view, and aided with rich micro production and cost data researchers have started to look into the prevalence of market-power-related distortions, impacting both deadweight loss and productive inefficiency. We briefly present the underlying framework that has allowed researchers to leverage standard productivity data to say something about market power in a given market or industry.

[76]The lack of cost-side evidence is also noted by Whinston (2008). In fact, he relies on Olley and Pakes (1996) just to draw even some tentative conclusions.

7.1.1 Measuring Market Power Using Production Data

Estimating the size of the wedge between price and marginal costs is often the place IO economists start when measuring market power. This wedge, *the markup*, measures a producer's ability to raise the price above the marginal cost of production and at a basic level indicates the presence of market power, or a least the possibility. Bresnahan (1989) discusses the challenges of measuring market power, noting in particular what by now can be safely called the conventional wisdom that *measuring markups is difficult because marginal costs are rarely directly observed*. We agree, but it is still very useful to embrace the rich production and cost information used in the productivity literature to aid in measuring marginal costs.

Let us first define the markup as the price-to-marginal cost ratio:[77]

$$\mu \equiv \frac{P}{c} \tag{47}$$

There exist three main approaches to measure markups. First, the accounting approach relies on directly observable gross (or net) margins of profits. While this is straightforward to implement, it suffers from well-known problems, chief among them the assumption that average cost equals marginal cost. This imposes strong restrictions on firm-level cost structures.[78] Under this setup, the markup equals the profit rate. This is often an undesirable assumption in many applications of interest to IO economists, as it rules out economies of scale, network effects, and so on—cost structures that may be relevant across a wide range of industries and markets.

The second approach comes from the New Empirical IO literature (see Bresnahan (1989)) and relies on the specification of a demand system that delivers price-elasticities of demand. Combined with assumptions on how firms compete, the demand approach delivers measures of markups through the first order condition associated with optimal pricing. This approach has a long tradition in the field of IO, and has been widely market tested.[79] It is also well-known that this approach restricts attention to a particular model

[77]Some work instead considers markups in absolute terms, $P - c$, or as a Lerner index, $\frac{P-c}{P}$. While there are obvious and straightforward conversions among these three metrics, one has to keep in mind which definition is being used when comparing results across different settings.

[78]In essence, the simplicity of the accounting approach is to multiply the price-cost ratio through by total output (Q) and obtain the ratio of revenue (PQ) and total cost (cQ), both of which are commonly reported. Underlying this are the same assumptions as those used to rely on factor (cost) shares: constant returns to scale in production and the absence of economies of scale; i.e., there are no fixed costs.

[79]See Pakes (2021) for an overview.

of conduct (most often, a Bertrand-Nash static pricing game) from which the first order condition price equations are derived, while requiring information on product-level prices and quantities for all relevant products in a market (which often the researcher must define).[80]

De Loecker and Warzynski (2012) propose an alternative and complementary approach to markup estimation, labeled the "production approach." The method builds on the insights of Hall (1988) and relies on a different first order condition to measure markups: cost minimization of a variable input of production. Implementation requires the output elasticity of the variable input and its revenue share. These were the topic of Section 5 in this chapter; however, this approach moves the focus away altogether from trying to recover the productivity residual.[81]

The key assumption behind this approach is that, in a given period, producers minimize cost by optimally choosing those inputs that are free from frictions, the statically chosen factors (as opposed to the dynamically chosen factors, which face adjustment costs and other frictions).

This framework leads to the following expression to compute the markup using production and cost data:

$$\mu_{it} = \theta_{it}^{V} \frac{P_{it} Q_{it}}{P_{it}^{V} X_{it}^{V}}, \tag{48}$$

with θ_{it}^{V} the output elasticity (of input X^{V}), importantly it is in general to producer-time specific.[82]

The flexibility of the approach is recognized when one notes that the markup expression is derived without specifying conduct in the product market or a particular demand system. With this approach to markup estimation, there are in principle multiple first order conditions, one for each variable input in production, that yield an expression for the markup. See De Loecker and Warzynski (2012), De Loecker, Eeckhout, and Unger (2020), and Raval (2019) for a discussion. Regardless of which variable input is used, there are two key ingredients needed in order to measure the markup: its revenue share and output

[80] An older literature relies on observed cost data to infer markups. Roberts and Dunne (1992) and Roberts and Supina (2000) are two early studies that used firm production data to measure marginal costs and use them to infer markups.

[81] We refer the reader to De Loecker (2011b) and De Loecker and Scott (2016) for a detailed discussion of both approaches, and the various trade-offs. Chapters 1 and 2 in this Volume [TO BE CONFIRMED] discuss the demand approach in great detail.

[82] In the case of our leading Cobb-Douglas example this would correspond to the parameter β_V.

elasticity. The approaches discussed above in this chapter generate estimates of the output elasticity. However, the validity of the production approach to recover markups is quite distinct from the auxiliary assumptions leveraged to estimate the production function.[83]

The standard derivation discussed in De Loecker and Warzynski (2012) assumes that firms take input prices as given. This does not preclude input providers charging markups, potentially leading to double-marginalization. The production approach, however, can accommodate departures from price-taking by considering multiple variable inputs, allowing for a non-zero input-price elasticity. For recent applications of this approach see Morlacco (2017), Mertens (2020) and Rubens (2020), where input buyer power and labor market power in the form of monopsony are identified alongside product market power in a variety of settings.

The markup formula (48) derived under the production approach highlights that the marginal cost of production is derived from a single variable input, without imposing any particular substitution elasticity with respect to other inputs (variable or fixed) or returns to scale. It is instructive to contrast it to the accounting approach introduced above. Only in the case of constant returns to scale and either a single variable input (V), or *only* variable inputs in the production function (thus excluding fixed costs), will the accounting-based markup be correct.

While the production approach holds in general, and thus also for multi-product firms (for each product), deploying this framework requires one to confront the standard input allocation problem. The main challenge is, as stated in Section 5.2, the product-specific input share cannot usually be observed. Additional assumptions or data are required; see, for example, De Loecker et al. (2016), who restrict attention to product-level production functions.

Applications The production approach to markups measurement has been used in a variety of applications, both in and outside of IO. It has been used to document rising markups and provide facts in the recent debate around the surge of market power and industry concentration in the U.S. and other world regions. While it has to confront the many measurement issues involved with production data and the challenges that come with measuring output elasticities, it has opened the debate and generated a renewed

[83]For example, Raval tests the hypothesis of a joint set of assumptions, capturing the specification of the production function, the presence of adjustment costs across input choices, and factor demand conditions, and the underlying behavioral model of cost minimization with respect to a given input.

interest in the productivity literature. Most notably, De Loecker, Eeckhout, and Unger (2020) rely on this approach to measure share-weighted aggregate markups for the U.S. economy during the period 1955-2016, and find a steady rise. The rise is correlated with greater measured fixed costs of production and increasing profitability. Moreover, the growth in the aggregate comes about from a reallocation process of market shares from relatively low to relatively high markup firms. This suggests a particular force at work, namely, increased alignment of output and markups over time in the economy. This ties in with the separate accounts of increased concentration in U.S. industries.[84]

Through this body of work, the production approach has connected productivity data to a larger global debate around market power, declining labor shares, increased globalization, and a variety of other labor market issues such as monopsony.[85] We see this as a unique opportunity for IO to contribute to a larger debate on the overall state of competition, as well as to help identify sources and implications for competition policy, by using a combination of single-industry and cross-industry studies. The field can leverage its rich and diverse toolbox of measurement, estimation, and modeling techniques.

7.1.2 Integrating Product and Factor Markets Using Productivity Data

The production approach to markups relies on insights and practices of the production function estimation literature, and therefore the use of producer-level output and input data. At the same time, it directly interacts with the objects of interest in the *demand approach* literature. For lack of better terminology, and with an obvious oversimplification, the "demand approach" refers to the practice (discussed briefly above) of studying market power using an estimated demand system to recover cross-price elasticities, paired with assumptions on conduct and market structure, that yield a first-order condition for pricing. This delivers measures of marginal costs and markups without observing any cost data, following the approach suggested by Rosse (1970) and Bresnahan (1987), and hence the phrase *estimating cost without cost data*. While the latter has traditionally relied on consumer-level data to learn about competition across firms in a given market, it relies on assumptions of the relevant upstream (factor) markets. This is precisely where the productivity data can be used to learn more about the relationships along the supply chain. We briefly discuss two promising avenues of research that leverage the very data

[84]Though there are many well-known issues relying on concentration ratio (e.g., HHI). See Syverson (2019) and Berry, Gaynor, and Scott Morton (2019) for more discussion.

[85]See Autor et al. (2020) and Krueger (2018).

sources we have discussed in this chapter, to learn about features of factor markets and vertical aspects of industries.

Vertical Linkages The integration of these approaches, including a combination of a variety of data sources, to learn about vertical linkages has started in earnest only recently. Productivity along buyer-supplier links has been studied to this point mostly in the context of international trade. Some IO-related exceptions include Hortaçsu and Syverson (2007), Forbes and Lederman (2010), and Atalay, Hortaçsu, and Syverson (2014).

The natural orientation of the data, with most production-approach data recorded at the producer level (e.g., the beer brewer or car manufacturer) and demand data at the retail level (e.g., supermarket or dealership), naturally links different parts of the production and distribution chain. Researchers can deploy this combined approach to study the interactions of productivity, factor demand, and margins among producers, wholesalers, and retailers. For example, De Loecker and Scott (2016) combines the production and demand approach to measure the extent to which the U.S. retail market (selling beer) is competitive.

Labor Market Power While much of IO has traditionally focused on product market competition, there has recently been a marked increase in interest in monopsony power in factor markets, particularly labor. IO economists have deployed various approaches, including the production approach (extended to incorporate input market power), to do so. De Loecker, Eeckhout, and Mongey (2021) relate increasing aggregate markups to wages and the overall labor share in the US. Prager and Schmitt (2021) document the downward impact on nurses' wages in hospital markets exposed to merger activity. Rubens (2021) exploits a Chinese government-mandated consolidation in cigarette manufacturing to estimate the monopsony power in upstream tobacco farming. Goolsbee and Syverson (2019) estimate the extent of universities' market power in the labor market for faculty. Azar, Berry, and Marinescu (2019) apply modeling insights from the differentiated product demand literature to analogous situations created by differentiated jobs in the labor market.

After developing mostly separately for a long time, we think the time is ripe for researchers to harness the combined capabilities of the productivity and demand literatures.

7.2 Technological Change and Market-Level Outcomes

The measurement of technological change and discovery of its drivers and implications has been a central component of the productivity literature since its inception. As discussed before, the predominant model of technological change in applications involves exogeneous Hicks-neutral productivity growth. In reality, however, technological change may come about in different ways. We briefly mention two recent literatures that consider departures from this standard setup.

7.2.1 Factor-Biased Technological Change

Hicks-neutral productivity influences the marginal products of all factors the same way. Departures from this, involving factor-biased technological differences, have been treated (mostly) theoretically in the macroeconomics literature for some time. These kinds of frameworks have started to creep into the micro productivity literature.

Raval (2019) provides evidence that there is a nontrivial amount of, and variation in, labor augmenting productivity among U.S. manufacturers. Further, this factor-specific productivity, just like typically measured TFP, is persistent at the producer level and correlated with producers' exporter status, size, and growth. Doraszelski and Jaumandreu (2018) estimate the factor-bias of technological change among Spanish producers and find important roles for a labor-augmenting productivity process. Interestingly, this labor-favoring productivity is more closely tied to firms' R&D than is their measured Hicks-neutral productivity. Zhang (2019) finds that labor-augmenting productivity growth explains over half of the large decline in labor's share of income in the Chinese steel industry during the 2000s. In the methodological vein, Demirer (2020) extends proxy-variable production function estimation approaches to allow for non-neutral productivity.

7.2.2 Endogenous Productivity Growth

Producer-level productivity exhibits obvious empirical dynamics. Production function estimation approaches incorporate such dynamics, and indeed often model it explicitly (e.g., assuming productivity follows a Markov process), and allow that process to vary systematically with observables. However, much of the literature treats the process as exogenous. The tension with this treatment is that it is clear producers would have incentives to improve their productivity level if doing so were cost effective, and introspection

about producers' actual behaviors strongly indicates that they do indeed invest resources in attempts to do precisely that.

The obvious conceptual resolution to this tension is to endogenize the productivity process, allowing it to be influenced by the choices of the producer, like R&D. Doing so would in some sense return to an earlier part of the productivity literature that sought to explain productivity differences using proxies for producers' investments in productivity growth (e.g., the accumulated stock of R&D spending). This earlier work did not explicitly model the endogenous choice and benefit-cost tradeoff underlying the productivity-enhancing investments, but at its core was a notion that such investments were consequential and important to understand. While fully endogenous productivity is not yet the standard treatment in the literature, there are several studies that have made strides in this direction. Prominent examples include Doraszelski and Jaumandreu (2013), Bøler, Moxnes, and Ulltveit-Moe (2015), Peters et al. (2017), and Humlum (2019).

8 Conclusion

The arrival of detailed producer-level production and cost data covering a broad set of industries and countries has left an enormous mark on the research literature. It has brought models of firm behavior and industry performance closer to the data. Economists in multiple fields have increasingly interacted with each other, finding a common interest in understanding production and performance at the micro level and describing their influence on industry and economic aggregates. We have provided, in this chapter, a sample of this type of research. We believe further progress is possible and likely to yield many insights. We encourage interested researchers to follow this existing work.

References

Abramovitz, Moses (1956). "Resource and Output Trends in the United States since 1870". *American Economic Review* 46, pp. 5–23.

Acemoglu, Daron (2002). "Directed Technical Change". *The Review of Economic Studies* 69.4, pp. 781–809.

Acemoglu, Daron (2003). "Labor- and Capital-Augmenting Technical Change". *Journal of the European Economic Association* 1.1, pp. 1–37.

Ackerberg, Daniel A (2020). "Timing Assumptions and Efficiency: Empirical Evidence in a Production Function Context". *Mimeo, University of Texas at Austin*.

Ackerberg, Daniel A, Benkard, C Lanier, Berry, Steven, and Pakes, Ariel (2007). "Econometric Tools for Analyzing Market Outcomes". *Handbook of Econometrics* 6, pp. 4171–276.

Ackerberg, Daniel A, Caves, Kevin, and Frazer, Garth (2015). "Identification Properties of Recent Production Function Estimators". *Econometrica* 83.6, pp. 2411–51.

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

Ackerberg, Daniel A, Frazer, Garth, Kim, Kyoo il, Luo, Yao, and Su, Yingjun (2020a). "Under-Identification of Structural Models Based on Timing and Information Set Assumptions".

Ackerberg, Daniel A, Frazer, Garth, Luo, Yao, and Su, Yingjun (2020b). "Under-Identification of Structural Models Based on Timing and Information Set Assumptions". *Mimeo, University of Texas at Austin*.

Ackerberg, Daniel A and Hahn, Jinyong (2015). "Some Non-Parametric Identification Results using Timing and Information Set Assumptions". *Working Paper*.

Allcott, Hunt, Collard-Wexler, Allan, and O'Connell, Stephen D (2016). "How do Electricity Shortages Affect Industry? Evidence from India". *American Economic Review* 106.3, pp. 587–624.

Arellano, Manuel and Bond, Stephen (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". *The Review of Economic Studies* 58.2, pp. 277–97.

Asker, John, Collard-Wexler, Allan, and De Loecker, Jan (2014). "Dynamic Inputs and Resource (Mis) Allocation". *Journal of Political Economy* 122.5, pp. 1013–63.

Asker, John, Collard-Wexler, Allan, and De Loecker, Jan (2019). "(Mis) Allocation, Market Power, and Global Oil Extraction". *American Economic Review* 109.4, pp. 1568–615.

Asplund, Marcus and Nocke, Volker (2006). "Firm Turnover in Imperfectly Competitive Markets". *The Review of Economic Studies* 73.2, pp. 295–327.

Atalay, Enghin (2014). "Materials Prices and Productivity". *Journal of the European Economic Association* 12.3, pp. 575–611.

Atalay, Enghin, Hortaçsu, Ali, and Syverson, Chad (2014). "Vertical Integration and Input Flows". *American Economic Review* 104.4, pp. 1120–48.

Autor, David, Dorn, David, Katz, Lawrence F, Patterson, Christina, and Van Reenen, John (2020). "The Fall of the Labor Share and the Rise of Superstar Firms". *The Quarterly Journal of Economics* 135.2, pp. 645–709.

Azar, José, Berry, Steven, and Marinescu, Ioana Elena (2019). "Estimating Labor Market Power".

Backus, Matthew (2020). "Why is Productivity Correlated with Competition?" *Econometrica* 88.6, pp. 2415–44.

Baily, Martin Neil, Hulten, Charles, Campbell, David, Bresnahan, Timothy F, and Caves, Richard E (1992). "Productivity Dynamics in Manufacturing Plants". *Brookings Papers on Economic Activity. Microeconomics* 1992, pp. 187–267.

Balat, Jorge, Brambilla, Irene, and Sasaki, Yuya (2016). "Heterogeneous Firms: Skilled-labor Productivity and the Destination of Exports". *Working Paper*.

Bartelsman, Eric J and Doms, Mark (2000). "Understanding Productivity: Lessons from Longitudinal Microdata". *Journal of Economic Literature* 38.3, pp. 569–94.

Basu, Susanto and Fernald, John G. (1997). "Returns to Scale in U.S. Production: Estimates and Implications". *Journal of Political Economy* 105.2, pp. 249–83.

Becker, Randy A, Haltiwanger, John, Jarmin, Ron S, Klimek, Shawn D, and Wilson, Daniel J (2006). "Micro and Macro Data Integration: The Case of Capital". *A new architecture for the US national accounts*. University of Chicago Press, pp. 541–610.

Bernard, Andrew B, Redding, Stephen J, and Schott, Peter K (2010). "Multiple-Product Firms and Product Switching". *American Economic Review* 100.1, pp. 70–97.

Berry, Steven, Gaynor, Martin, and Scott Morton, Fiona (2019). "Do Increasing Markups Matter? Lessons From Empirical Industrial Organization". *Journal of Economic Perspectives* 33.3, pp. 44–68.

Berry, Steven, Kortum, Samuel, and Pakes, Ariel (1996). "Environmental Change and Hedonic Cost Functions for Automobiles". *Proceedings of the National Academy of Sciences* 93.23, pp. 12731–38.

Berry, Steven, Levinsohn, James, and Pakes, Ariel (1995). "Automobile Prices in Market Equilibrium". *Econometrica: Journal of the Econometric Society*, pp. 841–90.

Berry, Steven and Reiss, Peter (2007). "Chapter 29 Empirical Models of Entry and Market Structure". Handbook of Industrial Organization 3. Ed. by M. Armstrong and R. Porter, pp. 1845–86.

Bilir, L Kamran and Morales, Eduardo (2020). "Innovation in the Global Firm". *Journal of Political Economy* 128.4, pp. 1566–625.

Bloom, Nicholas (2009). "The Impact of Uncertainty Shocks". *Econometrica* 77.3, pp. 623–85.

Bloom, Nicholas and Van Reenen, John (2007). "Measuring and Explaining Management Practices Across Firms and Countries". *Quarterly Journal of Economics* 122.4, 1351–1408.

Blundell, Richard and Bond, Stephen (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models". *Journal of Econometrics* 87.1, pp. 115–43.

Bond, Stephen and Söderbom, Måns (2005). "Adjustment Costs and the Identification of Cobb Douglas Production Functions". *Institute for Fiscal Studies* WP O5/04.

Borenstein, Severin, Bushnell, James B, and Wolak, Frank A (2002). "Measuring Market Inefficiencies in California's Restructured Wholesale Electricity Market". *American Economic Review* 92.5, pp. 1376–405.

Braguinsky, Serguey, Ohyama, Atsushi, Okazaki, Tetsuji, and Syverson, Chad (2015). "Acquisitions, Productivity, and Profitability: Evidence from the Japanese Cotton Spinning Industry". *American Economic Review* 105.7, pp. 2086–119.

Bresnahan, Timothy F (1987). "Competition and Collusion in the American Automobile Industry: The 1955 Price War". *Journal of Industrial Economics* 35.4, pp. 457–82.

Bresnahan, Timothy F (1989). "Empirical Studies of Industries with Market Power". *Handbook of Industrial Organization* 2, pp. 1011–57.

Bresnahan, Timothy F, Brynjolfsson, Erik, and Hitt, Lorin M (2002). "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence". *The Quarterly Journal of Economics* 117.1, pp. 339–76.

Burkhardt, Jesse (2019). "The Impact of the Renewable Fuel Standard on US Oil Refineries". *Energy Policy* 130, pp. 429–37.

Bøler, Esther Ann, Moxnes, Andreas, and Ulltveit-Moe, Karen Helene (2015). "R&D, International Sourcing, and the Joint Impact on Firm Performance". *American Economic Review* 105.12, pp. 3704–39.

Chandra, Amitabh, Finkelstein, Amy, Sacarny, Adam, and Syverson, Chad (2016). "Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector". *American Economic Review* 106.8, pp. 2110–44.

Christensen, Laurits R, Jorgenson, Dale, and Lau, Lawrence J (1973). "Transcendental Logarithmic Production Frontiers". *The Review of Economics and Statistics* 55.1, pp. 28–45.

Collard-Wexler, Allan (2013). "Demand Fluctuations in the Ready-Mix Concrete Industry". *Econometrica* 81.3, pp. 1003–37.

Collard-Wexler, Allan and De Loecker, Jan (2015). "Reallocation and Technology: Evidence from the US Steel Industry". *American Economic Review* 105.1, pp. 131–71.

Collard-Wexler, Allan and De Loecker, Jan (2016). "Production Function Estimation with Measurement Error in Inputs". *NBER Working Paper* 22437.

Cooper, Russell, Haltiwanger, John, and Willis, Jonathan L (2015). "Dynamics of Labor Demand: Evidence from Plant-Level Observations and Aggregate Implications". *Research in Economics* 69.1, pp. 37–50.

Cooper, Russell W and Haltiwanger, John C (2006). "On the Nature of Capital Adjustment Costs". *The Review of Economic Studies* 73.3, pp. 611–33.

Cooper, William W, Seiford, Lawrence M, and Zhu, Joe (2011). "Data Envelopment Analysis: History, Models, and Interpretations. In: Cooper W, Seiford L., Zhu J (eds) Handbook on Data Envelopment Analysis." *International Series in Operations Research and Management Science* 164, pp. 1–39.

Crouzet, Nicolas and Eberly, Janice (2020). "Rents and Intangible Capital: A Q+ Framework". *Mimeo, Northwestern University*.

De Loecker, Jan (2011a). "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity". *Econometrica* 79.5, pp. 1407–51.

De Loecker, Jan (2011b). "Recovering Markups from Production Data". *International Journal of Industrial Organization* 29.3, pp. 350–55.

De Loecker, Jan (2013). "Detecting Learning by Exporting". *American Economic Journal: Microeconomics* 5.3, pp. 1–21.

De Loecker, Jan, Eeckhout, Jan, and Mongey, Simon (2021). "Quantifying Market Power and Business Dynamism in the Macroeconomy". 28761 (May 2021).

De Loecker, Jan, Eeckhout, Jan, and Unger, Gabriel (2020). "The Rise of Market Power and the Macroeconomic Implications". *The Quarterly Journal of Economics* 135.2, pp. 561–644.

De Loecker, Jan, Fuss, Catherine, and Van Biesebroeck, Jo (2018). "Markup and Price Dynamics: Linking Micro to Macro". *NBB Working Paper* 357.

De Loecker, Jan, Goldberg, Pinelopi K, Khandelwal, Amit K, and Pavcnik, Nina (2016). "Prices, Markups, and Trade Reform". *Econometrica* 84.2, pp. 445–510.

De Loecker, Jan and Goldberg, Pinelopi Koujianou (2014). "Firm Performance in a Global Market". *Annual Review of Economics* 6.1, pp. 201–27.

De Loecker, Jan and Scott, Paul T (2016). "Estimating Market Power. Evidence from the US Brewing Industry". *NBER Working Paper* 22957.

De Loecker, Jan and Warzynski, Frederic (2012). "Markups and Firm-Level Export Status". *American Economic Review* 102.6, pp. 2437–71.

Demirer, Mert (2020). "Production Function Estimation with Factor-Augmenting Technology: An Application to Markups". *MIT Working Paper*.

Dhyne, Emmanuel, Petrin, Amil, Smeets, Valérie, and Warzynski, Frederic (2020). "Theory for Extending Single-Product Production Function Estimation to Multi-Product Settings". *Aarhus University Working Paper*.

Dhyne, Emmanuel, Petrin, Amil, and Warzynski, Frederic (2014). "Deregulation and Spillovers in Multi-Product Production Settings". *Mimeo, Aarhus University*.

Diewert, W Erwin (1973). "Functional Forms for Profit and Transformation Functions". *Journal of Economic Theory* 6.3, pp. 284–316.

Doraszelski, Ulrich and Jaumandreu, Jordi (2013). "R&D and Productivity: Estimating Endogenous Productivity". *Review of Economic Studies* 80.4, pp. 1338–83.

Doraszelski, Ulrich and Jaumandreu, Jordi (2018). "Measuring the Bias of Technological Change". *Journal of Political Economy* 126.3, pp. 1027–84.

Dunne, Timothy, Roberts, Mark J, and Samuelson, Larry (1989). "The Growth and Failure of US Manufacturing Plants". *The Quarterly Journal of Economics* 104.4, pp. 671–98.

Eckel, Carsten and Neary, J Peter (2010). "Multi-Product Firms and Flexible Manufacturing in the Global Economy". *The Review of Economic Studies* 77.1, pp. 188–217.

Ericson, Richard and Pakes, Ariel (1995). "Markov-Perfect Industry Dynamics: A Framework for Empirical Work". *The Review of Economic Studies* 62.1, pp. 53–82.

Eslava, Marcela and Haltiwanger, John C (2020). "The Life-Cycle Growth of Plants: The Role of Productivity, Demand and Wedges". *NBER Working Paper* 27184.

Feenstra, Robert C (2003). "A Homothetic Utility Function for Monopolistic Competition Models, Without Constant Price Elasticity". *Economics Letters* 78.1, pp. 79–86.

Forbes, Silke and Lederman, Mara (2010). "Does Vertical Integration Affect Firm Performance? Evidence from the Airline Industry". *RAND Journal of Economics* 41.4, pp. 765–90.

Forlani, Emanuele, Martin, Ralf, Mion, Giordiano, and Muûls, Mirabelle (2016). "Unraveling Firms: Demand, Productivity and Markups Heterogeneity". *CEPR Discussion Paper* DP11058.

Foster, Lucia, Haltiwanger, John, and Krizan, C.J. (2006). "Market Selection, Reallocation, and Restructuring in the U.S. Retail Trade Sector in the 1990s". *The Review of Economics and Statistics* 88.4, pp. 748–58.

Foster, Lucia, Haltiwanger, John, and Syverson, Chad (2008). "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review* 98.1, pp. 394–425.

Fox, Jeremy T., Hadad, Vitor, Hoderlein, Stefan, Petrin, Amil, and Sherman, Robert P. (2017). "Heterogeneous Production Functions, Panel Data and Productivity Dispersion".

Fox, Jeremy T and Smeets, Valérie (2011). "Does Input Quality Drive Measured Differences in Firm Productivity?" *International Economic Review* 52.4, pp. 961–89.

Ganapati, Sharat (2021). "The Modern Wholesaler: Global Sourcing, Domestic Distribution and Scale Economies". *Center for Economic Studies, US Census Bureau Working Papers*.

Gandhi, Amit, Navarro, Salvador, and Rivers, David A (2020). "On the Identification of Gross Output Production Functions". *Journal of Political Economy* 128.8, pp. 2973–3016.

Goldberg, Pinelopi Koujianou, Khandelwal, Amit Kumar, Pavcnik, Nina, and Topalova, Petia (2010). "Imported Intermediate Inputs and Domestic Product Growth: Evidence from India". *The Quarterly Journal of Economics* 125.4, pp. 1727–67.

Goldmanis, Maris, Hortaçsu, Ali, Syverson, Chad, and Emre, Önsel (2010). "E-Commerce and the Market Structure of Retail Industries". *The Economic Journal* 120.545, pp. 651–82.

Goolsbee, Austan and Syverson, Chad (2019). "Monopsony Power in Higher Education: A Tale of Two Tracks". Working Paper Series 26070.

Grieco, Paul LE, Li, Shengyu, and Zhang, Hongsong (2016). "Production Function Estimation with Unobserved Input Price Dispersion". *International Economic Review* 57.2, pp. 665–90.

Grieco, Paul LE and McDevitt, Ryan C (2017). "Productivity and Quality in Health Care: Evidence from the Dialysis Industry". *The Review of Economic Studies* 84.3, pp. 1071–105.

Griliches, Zvi (1967). "Production Functions in Manufacturing: Some Preliminary Results". NBER Chapters, pp. 275–340.

Griliches, Zvi and Mairesse, Jacques (1995). "Production Functions: The Search for Identification". *NBER Working Paper* 5067.

Hall, Robert E (1988). "Intertemporal Substitution in Consumption". *Journal of Political Economy* 96.2, pp. 339–57.

Haltiwanger, John, Kulick, Robert, and Syverson, Chad (2018). "Misallocation Measures: The Distortion that Ate the Residual". *NBER Working Paper* 24199.

Haltiwanger, John C (1997). "Measuring and Analyzing Aggregate Fluctuations: The Importance of Building from Microeconomic Evidence". *Federal Reserve Bank of St. Louis Review* 79.3, p. 55.

Harberger, Arnold C (1954). "The Welfare Loss from Monopoly". *American Economic Review* 44.2, pp. 77–87.

Hendel, Igal and Spiegel, Yossi (2014). "Small Steps for Workers, a Giant Leap for Productivity". *American Economic Journal: Applied Economics* 6.1, pp. 73–90.

Hopenhayn, Hugo A (2014). "Firms, Misallocation, and Aggregate Productivity: A Review". *Annual Review of Economics* 6.1, pp. 735–70.

Hortaçsu, Ali and Syverson, Chad (2007). "Cementing Relationships: Vertical Integration, Foreclosure, Productivity, and Prices". 12894.

Hsieh, Chang-Tai and Klenow, Peter J (2009). "Misallocation and Manufacturing TFP in China and India". *The Quarterly Journal of Economics* 124.4, pp. 1403–1448.

Humlum, Anders (2019). "Robot Adoption and Labor Market Dynamics". *Working Paper*.

Itoga, Takaaki (2019). "Within-Firm Reallocation and the Impacts of Trade under Factor Market Imperfection". *Mimeo, Penn State University*.

Jovanovic, Boyan (1982). "Selection and the Evolution of Industry". *Econometrica* 50.3, pp. 649–70.

Kim, Kyoo il, Luo, Yao, and Su, Yingjun (2019). "A Robust Approach to Estimating Production Functions: Replication of the ACF Procedure". *Journal of Applied Economics* 34.4, pp. 612–19.

Kim, Kyoo il, Petrin, Amil, and Song, Suyong (2016). "Estimating Production Functions with Control Functions when Capital is Measured with Error". *Journal of Econometrics* 190.2, pp. 267–79.

Klette, Tor Jakob and Griliches, Zvi (1996). "The Inconsistency of Common Scale Estimators when Output Prices are Unobserved and Endogenous". *Journal of Applied Econometrics* 11.4, pp. 343–61.

Krueger, Alan B (2018). "Reflections on Dwindling Worker Bargaining Power and Monetary Policy". *Presentation at the Jackson Hole Economic Symposium*.

Kugler, Maurice and Verhoogen, Eric (2012). "Prices, Plant Size, and Product Quality". *The Review of Economic Studies* 79.1, pp. 307–39.

Lau, Lawrence J (1976). "A Characterization of the Normalized Restricted Profit Function". eng. *Journal of Economic Theory* 12.1, pp. 131–63.

Levinsohn, James and Melitz, Marc (2002). "Productivity in a Differentiated Products Market Equilibrium". *Unpublished Manuscript* 9, pp. 12–25.

Levinsohn, James and Petrin, Amil (2003). "Estimating Production Functions using Inputs to Control for Unobservables". *The Review of Economic Studies* 70.2, pp. 317–41.

Li, Tong and Sasaki, Yuya (2017). "Constructive Identification of Heterogeneous Elasticities in the Cobb-Douglas Production Function".

Lucas, Robert (1978). "On the Size Distribution of Business Firms". *Bell Journal of Economics* 9.2, pp. 508–23.

Mairesse, Jacques and Jaumandreu, Jordi (2005). "Panel-Data Estimates of the Production Function and the Revenue Function: What Difference Does it Make?" eng. *The Scandinavian journal of economics*. Scandinavian Journal of Economics 107.4, pp. 651–72.

McFadden, Daniel (1978). *Cost Revenue and Profit Functions' in Fuss, M and D. Mc Fadden (ed.) Production Economics: A Dual Approach to Theory and Applications, Vol-1.*

Melitz, Marc J (2003). "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity". *Econometrica* 71.6, pp. 1695–725.

Melitz, Marc J and Ottaviano, Gianmarco IP (2008). "Market Size, Trade, and Productivity". *The Review of Economic Studies* 75.1, pp. 295–316.

Melitz, Marc J and Polanec, Sašo (2015). "Dynamic Olley-Pakes Decomposition with Entry and Exit". *RAND Journal of Economics* 46.2, pp. 362–75.

Melitz, Marc J and Trefler, Daniel (2012). "Gains from Trade when Firms Matter". *Journal of Economic Perspectives* 26.2, pp. 91–118.

Mertens, Matthias (2020). "Labor Market Power and the Distorting Effects of International Trade". *International Journal of Industrial Organization* 68.

Morlacco, Monica (2017). "Market Power in Input Markets: Theory and Evidence from French Manufacturing". *Yale University Technical Report*.

Nerlove, Marc (1961). "Returns to Scale in Electricity Supply". *ed. by C. Christ et al. Stanford: Stanford University Press*, 167–200.

Nishida, Mitsukuni, Petrin, Amil, and Polanec, Sašo (2014). "Exploring Reallocation's apparent Weak Contribution to Growth". eng. *Journal of Productivity Analysis* 42.2, pp. 187–210.

Olley, G. Steven and Pakes, Ariel (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". *Econometrica* 64.6, pp. 1263–97.

Ornaghi, Carmine (2006). "Assessing the Effects of Measurement Errors on the Estimation of Production Functions". *Journal of Applied Econometrics* 21.6, pp. 879–91.

Orr, Scott (2019). "Within-Firm Productivity Dispersion: Estimates and Implications". *University of British Columbia Working Paper*.

Pakes, Ariel (2021). "A Helicopter Tour of Some Underlying Issues in Empirical Industrial Organization". *Annual Review of Economics* 13, pp. 397–421.

Panzar, John C. (1989). "Technological Determinants of Firm and Industry Structure". 1. Ed. by R. Schmalensee and R. Willig, pp. 3–59.

Pavcnik, Nina (2003). "What Explains Skill Upgrading in Less Developed Countries?" *Journal of Development Economics* 71.2, pp. 311–28.

Peters, Bettina, Roberts, Mark J, Vuong, Van Anh, and Fryges, Helmut (2017). "Estimating Dynamic R&D Choice: An Analysis of Costs and Long-run Benefits". *RAND Journal of Economics* 48.2, pp. 409–37.

Peters, Ryan H and Taylor, Lucian A (2017). "Intangible Capital and the Investment-Q Relation". *Journal of Financial Economics* 123.2, pp. 251–72.

Petrin, Amil and Levinsohn, James (2012). "Measuring Aggregate Productivity Growth using Plant-Level Data". *RAND Journal of Economics* 43.4, pp. 705–25.

Pozzi, Andrea and Schivardi, Fabiano (2016). "Demand or Productivity: What Determines Firm Growth?" *RAND Journal of Economics* 47.3, pp. 608–30.

Prager, Elena and Schmitt, Matt (2021). "Employer Consolidation and Wages: Evidence from Hospitals". *American Economic Review* 111.2, pp. 397–427.

Raval, Devesh (2019). "Testing the Production Approach to Markup Estimation". *Mimeo, University of Texas at Austin*.

Reiss, Peter C. and Wolak, Frank A (2007). "Structural Econometric Modeling: Rationales and Examples from Industrial Organization". 6A. Ed. by JJ. Heckman and E.E. Leamer.

Restuccia, Diego and Rogerson, Richard (2008). "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments". *Review of Economic Dynamics* 11.4, pp. 707–20.

Roberts, Mark J, Xu, Daniel Yi, Fan, Xiaoyan, and Shengxing, Zhang (2017). "The Role of Firm Factors in Demand, Cost, and Export Market Selection for Chinese Footwear Producers". *The Review of Economic Studies* 85.4, pp. 2429–61.

Rosse, James N (1970). "Estimating Cost Function Parameters without Using Cost Data: Illustrated Methodology". *Econometrica* 38.2, pp. 256–75.

Rubens, Michael (2020). "Ownership Consolidation, Monopsony Power and Efficiency: Evidence from the Chinese Tobacco Industry". *Mimeo, University of Leuven*.

Rubens, Michael (2021). "Market Structure, Oligopsony Power, and Productivity". *Mimeo, University of Leuven*.

Saunders, Adam and Brynjolfsson, Erik (2016). "Valuing IT-Related Intangible Assets". *MIS Quarterly* 40.1, pp. 83–110.

Slavtchev, Viktor, Bräuer, Richard, and Mertens, Matthias (2020). "Import Competition and Firm Productivity: Evidence from German Manufacturing". *IWH Discussion Papers* 1/20.

Smeets, Valérie and Warzynski, Frederic (2013). "Estimating Productivity with Multi-Product Firms, Pricing Heterogeneity and the Role of International Trade". *Journal of International Economics* 90.2, pp. 237–44.

Solow, Robert M (1957). "Technical Change and the Aggregate Production Function". *Review of Economics and Statistics* 39.3, pp. 312–20.

Stiebale, Joel and Vencappa, Dev (2018). "Acquisitions, Markups, Efficiency, and Product Quality: Evidence from India". *Journal of International Economics* 112, pp. 70–87.

Syverson, Chad (2004a). "Market Structure and Productivity: A Concrete Example". *Journal of Political Economy* 112.6, pp. 1181–222.

Syverson, Chad (2004b). "Product Substitutability and Productivity Dispersion". *Review of Economics and Statistics* 86.2, pp. 534–50.

Syverson, Chad (2011). "What Determines Productivity?" *Journal of Economic Literature* 49.2, pp. 326–65.

Syverson, Chad (2019). "Macroeconomics and Market Power: Context, Implications, and Open Questions". *Journal of Economic Perspectives* 33.3, pp. 23–43.

Valmari, Nelli (2016). "Estimating Production Functions of Multiproduct Firms". *ETLA Working Paper* 37.

Van Biesebroeck, Johannes (2003). "Productivity Dynamics with Technology Choice: An Application to Automobile Assembly". *The Review of Economic Studies* 70.1 (Jan. 2003), pp. 167–98.

Walters, Alan A (1963). "Production and Cost functions: An Econometric Survey". *Econometrica* 31.1/2, pp. 1–66.

Westlake, S and Haskel, J (2017). "The Rise of the Intangible Economy: Capitalism without Capital". *Princeton University Press*.

Whinston, Michael D (2008). "Lectures on Antitrust Economics". *MIT Press Books*.

Wolak, Frank A (2003). "Measuring Unilateral Market Power in Wholesale Electricity Markets: The California Market, 1998-2000". *American Economic Review* 93.2, pp. 425–30.

Zhang, Hongsong (2019). "Non-neutral Technology, Firm Heterogeneity, and Labor Demand". *Journal of Development Economics* 140.C, pp. 145–68.