Prices, Markups and Trade Reform

Jan De Loecker† Pinelopi K. Goldberg ‡ Amit K. Khandelwal§ Nina Pavcnik¶

First Draft: February 2012
This Draft: April 2014

Abstract

This paper examines how prices, markups and marginal costs respond to trade liberalization. We develop a framework to estimate markups from production data with multi-product firms. This approach does not require assumptions on the market structure or demand curves faced by firms, nor assumptions on how firms allocate their inputs across products. We exploit quantity and price information to disentangle markups from quantity-based productivity, and then compute marginal costs by dividing observed prices by the estimated markups. We use India’s trade liberalization episode to examine how firms adjust these performance measures. Not surprisingly, we find that trade liberalization lowers factory-gate prices and that output tariff declines have the expected pro-competitive effects. However, the price declines are small relative to the declines in marginal costs, which fall predominantly because of the input tariff liberalization. The reason for this incomplete cost pass-through to prices is that firms offset their reductions in marginal costs by raising markups. Our results demonstrate substantial heterogeneity and variability in markups across firms and time and suggest that producers benefited relative to consumers, at least immediately after the reforms.

Keywords: Markups, Production Function Estimation, Marginal Cost, Pass-through, Input Tariffs, Trade Liberalization

†Princeton University, Fisher Hall, Prospect Ave, Princeton, NJ 08540, email: jdeloeck@princeton.edu
‡Yale University, 37 Hillhouse, New Haven, CT 06520, email: penny.goldberg@yale.edu
§Columbia Business School, Uris Hall, 3022 Broadway, New York, NY 10027 email: ak2796@columbia.edu
§§Dartmouth College, 6106 Rockefeller Hall, Hanover, NH 03755, email: nina.pavcnik@dartmouth.edu

*The work for this project was carried out while Goldberg was a Fellow of the Guggenheim Foundation, De Loecker was a visitor of the Cowles Foundation at Yale University and a visiting Professor at Stanford University, and Khandelwal was a Kenen Fellow at the International Economics Section at Princeton University. The authors thank the respective institutions for their support. We are grateful to Steve Berry, Elhanan Helpman, Ariel Pakes, Andres Rodriguez-Clare and Frank Wolak for useful discussions at early stages of this project and seminar participants at several institutions and conferences. We also thank the Editor and three anonymous referees for insightful comments and suggestions.
1 Introduction

Trade reforms have the potential to deliver substantial benefits to economies by forcing a more efficient allocation of resources. A large body of theoretical and empirical literature has analyzed the mechanisms behind this process. When trade barriers fall, aggregate productivity rises as less productive firms exit and the remaining firms expand (e.g., Melitz (2003) and Pavcnik (2002)) and take advantage of cheaper or previously unavailable imported inputs (e.g., Goldberg et al. (2010a) and Halpern et al. (2011)). Trade reforms have also been shown to reduce markups (e.g., Levinsohn (1993) and Harrison (1994)). Based on this evidence, we should expect trade reforms to exert downward pressure on firm prices. However, we have little direct evidence on how prices respond to liberalization because they are rarely observed during trade reforms. We fill this gap by developing a unified framework to estimate jointly markups and marginal costs from production data, and examine how prices, and their underlying markup and cost components, adjust during India's comprehensive trade liberalization.

Our paper makes three main contributions. First, we develop a unified framework to estimate markups and marginal costs of multi-product firms across a broad set of manufacturing industries. Since these measures are unobserved, we must impose some structure on the data. However, our approach does not require assumptions on consumer demand, market structure or the nature of competition common in industrial organization studies. This flexibility is particularly appealing in settings when one wants to infer the full distribution of markups across firms and products over time in different manufacturing sectors. Since prices are observed, we can directly recover marginal costs from the markup estimates. Our approach is quite general and since data containing this level of detail are becoming increasingly available, this methodology is useful to researchers studying other countries and industries. The drawback of this approach is that we are unable to perform counterfactual simulations since we do not explicitly model consumer demand and firm pricing behavior.

The second and key contribution of our study is towards the methodology to estimate production functions. In order to infer markups, the proposed approach requires estimates of production functions. Typically, these estimates have well-known biases if researchers use revenue rather than quantity data. Estimates of “true” productivity (or marginal costs) are confounded by demand shocks and markups, and these biases may be severe (see Foster et al. (2008)). De Loecker (2011) demonstrates that controlling for demand shocks substantially attenuates the productivity increases in response to trade reforms in the European Union textile industry. That paper addresses the bias arising from unobserved output prices (the so-called output price bias) by introducing a CES demand system in the analysis. In contrast to that approach, we address the output price-bias by estimating a quantity-based production function using data that contain the prices and quantities of firms’ products over time. The focus on a quantity-based production function highlights the need for the estimation to address two additional biases that have not received much attention in the literature: the bias stemming from the unobserved allocation of inputs across products within multi-product firms and the bias stemming from unobserved input prices (or the use of (quality)
differentiated inputs) by firms - the so-called input price bias. Our study contributes an approach to address these biases.

Third, existing studies that have analyzed the impact of trade reforms on markups have focused exclusively on the competitive effects from declines in output tariffs (e.g., Levinsohn (1993) and Harrison (1994)). Comprehensive reforms also lower tariffs on imported inputs and previous work, particularly on India, has emphasized this aspect of trade reforms (e.g., see Goldberg et al. (2009)). These two tariff reductions represent distinct shocks to domestic firms. Lower output tariffs increase competition by changing the residual demand that firms face. Conversely, firms benefit from lower costs of production when input tariffs decline. It is important to account for both channels of liberalization to understand the overall impact of trade reforms on prices and markups. In particular, declines in markups depend on the extent to which firms pass these cost savings to consumers, the pass-through being influenced by both the market structure and nature of demand. For example, in models with monopolistic competition and CES demand, markups are constant and so by assumption, pass-through of tariffs on prices is complete. Arkolakis et al. (2012) demonstrate that several of the influential trade models assume constant markups and by doing so, abstract away from the markup channel as a potential source of gains from trade. This is the case in Ricardian models that assume perfect competition, such as Eaton and Kortum (2002), and models with monopolistic competition such as Krugman (1980) and its heterogeneous firm extensions like Melitz (2003). There are models that can account for variable markups by imposing some structure on demand and market structure.\footnote{See Goldberg (1995), Bernard et al. (2003), Goldberg and Verboven (2005), Atkeson and Burstein (2008), Melitz and Ottaviano (2008), Feenstra and Weinstein (2010), Nakamura and Zerom (2010), Edmonds et al. (2011), Goldberg and Hellerstein (2013), Arkolakis et al. (2012), Mayer et al. (2014) and Atkin and Donaldson (2014).} While these studies allow for richer patterns of markup adjustment, the empirical results on markups and pass-through ultimately depend on the underlying parametric assumptions imposed on consumer demand and nature of competition. Ideally, we want to understand how trade reforms affect markups without having to rely on explicit parametric assumptions of the demand systems and/or market structures, which themselves may change with trade liberalization.

The structure of our analysis is as follows. We use production data to infer markups by exploiting the optimality of firms’ variable input choices. Our approach is based on Hall (1988) and De Loecker and Warzynski (2012), but we extend their methodology to account for multi-product firms and to take advantage of observable price data and physical quantity of products. The key assumption we need to infer markups is that firms minimize cost; then, markups are the deviation between the elasticity of output with respect to a variable input and that input’s share of total revenue. We obtain this output elasticity from estimates of production functions across many industries. As noted above, in contrast to many studies, we utilize physical quantity data rather than revenues to estimate the production functions.\footnote{Foster et al. (2008) also use quantity data in their analysis of production functions, but they focus on a set of homogeneous products.} This alleviates the concern that the production function estimation is contaminated by prices, yet presents different challenges that we discuss in detail in Section 3. Most importantly, using physical quantity data forces us to conduct the analysis at
the product level since without a demand system to aggregate across products, prices and physical quantities are only defined at the product level.

The approach we propose calls for an explicit treatment of multi-product firms. We show how to exploit data on single-product firms along with a sample selection correction to obtain consistent estimates of the production functions. The benefit of using single-product firms at the production function estimation stage is that it does not require assumptions on how firms allocate inputs across products, something we do not observe in our data.\(^3\) This approach assumes that the physical relationship between inputs and outputs is the same for single- and multi-product firms that manufacture the same product. That is, a single-product firm uses the same technology to produce rickshaws as a multi-product firm that produces rickshaws and cars. While this assumption may appear strong, it is already implicitly employed in all previous work that pools data across single- and multi-product firms (e.g., Olley and Pakes (1996) or Levinsohn and Petrin (2003)). Importantly, the assumption of the same physical production structure does not rule out economies of scope, which can operate through higher (factor-neutral) productivity of multi-product firms, the spreading of fixed costs across multiple products, or lower input prices for multi-product firms (as long as they are not related to input quantities). Once we estimate the production functions from the single-product firms, we show how to back out allocation of inputs across products within a multi-product firm. We obtain the markups for each product manufactured by firms by dividing the output elasticity of materials by the materials share of total revenue.\(^4\) Finally, we divide prices by the markups to obtain marginal costs.

The estimation of the production function provides plausible results and highlights the importance of addressing the input price bias. We also observe that firms have lower markups and higher marginal costs on products that are farther from their core competency, a finding consistent with recent heterogeneous models of multi-product firms. Foreshadowing the impact of the trade liberalizations, we find that changes in marginal costs are not perfectly reflected in changes in prices because of variable markups (i.e., incomplete pass-through).

Our main results focus on how prices, marginal costs, and markups adjust during India’s trade liberalization. As has been discussed extensively in earlier work, the nature of India’s reform provides an identification strategy that alleviates the standard endogeneity concerns associated with trade liberalization. Perhaps not surprisingly, we observe price declines during the reform period, but these declines appear modest relative to the size of the reform. On average, prices fall 18 percent despite average output tariff declines of 62 percentage points. Marginal costs, however, decline on average by 35 percent due primarily to input tariff liberalization; this finding is consistent with earlier

\(^3\)Suppose a firm manufactures three products using raw materials, labor and capital. To our knowledge, no dataset covering manufacturing firms reports information on how much of each input is used for each product. One way around this problem is to assume input proportionality. For example, Foster et al. (2008) allocate inputs based on products’ revenue shares. Their approach is valid under perfect competition or the assumption of constant markups across all products produced by a firm. While these assumptions may be appropriate for the particular homogenous good industries they study, we study a broad class of differentiated products where these assumptions may not apply. Moreover, our study aims to estimate markups without imposing such implicit assumptions.

\(^4\)For multi-product firms, we use the estimated input allocations in the markup calculation.
work demonstrating the importance of imported inputs in India’s trade reform. The predominant force driving down marginal costs are lower input tariffs reducing the costs of imported inputs, rather than output tariffs reducing X-inefficiencies. Since our prices decompose exactly into their underlying cost and markup components, we can show that the reason the relatively large decline in marginal costs did not translate to equally large price declines was because markups increased: on average, the trade reform raised relative markups by 17 percent. The results imply that firms offset the cost declines from input tariff reductions by raising markups, and the net effect is that the reform has an attenuated impact on prices. The increases in markups do not imply that the trade reforms caused firms to collude or engage in less competitive behavior. Rather, the results simply show that prices do not respond fully to cost, a finding that has been studied extensively in the exchange rate literature and is consistent with any model with variable markups. Finally, we observe that firms’ ability to raise markups even further is mitigated by the pro-competitive impact of output tariff declines, particularly for those firms with very high initial markups. Our analysis is based on data representative of larger firms, so our results are representative of these larger firms.

Our results suggest that the most likely beneficiaries of the trade liberalization in the short-run are domestic Indian firms who benefit from lower production costs while simultaneously raising markups. The short-run gains to consumers appear small, especially considering that we observe factory-gate prices rather than retail prices. However, the additional short-run profits accrued to firms may have spurred innovation in Indian manufacturing, particularly in the introduction of many new products, that benefit consumers in the long run. These new products accounted for about a quarter of overall manufacturing growth (see Goldberg et al. (2010b)). In earlier work, we showed that the new product introductions were concentrated in sectors with disproportionately large input tariff declines that allowed firms access to new, previously unavailable imported materials (see Goldberg et al. (2010a)). In the present paper, we find that firms with larger increases in average markups were more likely to introduce new products, which suggests that higher profits may have financed the development of new products that contributed to long run gains to consumers. A more detailed investigation of this channel is beyond the scope of the present paper.

In addition to the papers discussed earlier, our work is related to a wave of recent papers that focus on productivity in developing countries, such as Bloom and Van Reenen (2007) and Hsieh and Klenow (2009). The low productivity in the developing world is often attributed to lack of competition (see Bloom and Van Reenen (2007) and Bloom and Van Reenen (2010)) or the presence of policy distortions that result in a misallocation of resources across firms (Hsieh and Klenow (2009)). Against this background, it is natural to ask whether there is any evidence that an increase in competition or a removal of distortions reduces production costs. India’s reforms are an excellent context to study these questions because of the nature of the reform and the availability of detailed data. Trade protection is a policy distortion that distorts resource allocation. Limited competition benefits some firms relative to others, and the high input tariffs are akin to the capital

The relative importance of input and output tariffs is consistent with Amiti and Konings (2007) and Topalova and Khandelwal (2011) who find that firm-level productivity changes in Indonesia and India, respectively, were predominantly driven by input tariff declines.
distortions examined by Hsieh and Klenow (2009). Our results suggest that the removal of barriers on inputs lowered production costs, so the reforms did indeed deliver gains in the form of lower production costs. However, the overall picture is more nuanced as firms do not appear to pass the entirety of the cost savings to consumers in the form of lower prices. Our findings highlight the importance of jointly studying changes in prices, markups and costs to understand the full distributional consequences of trade liberalization.

The remainder of the paper is organized as follows. In the next section, we provide a brief overview of India’s trade reform and the data used in the analysis. In Section 3 we lay out the general empirical framework that allows us to estimate markups, and marginal costs. Section 3.1 presents the theoretical framework, Section 3.2 presents the empirical methodology to estimate the production function and discusses identification, and Section 3.3 explains the process to recover the allocation of inputs across products for multi-product firms. Section 4 presents the results and Section 5 concludes.

2 Data and Trade Policy Background

We first describe the Indian data since it dictates our empirical methodology. We also describe key elements of India’s trade liberalization that are important for our identification strategy. Given that the Indian trade liberalization has been described in a number of papers (including several by a subset of the present authors), we keep the discussion of the reforms brief.

2.1 Production and Price data

We use the Prowess data that is collected by the Centre for Monitoring the Indian Economy (CMIE). Prowess includes the usual set of variables typically found in firm-level production data, but has important advantages over the Annual Survey of Industries (ASI), India’s manufacturing census. First, unlike the repeated cross section in the ASI, Prowess is a panel that tracks firm performance over time. Second, the data span India’s trade liberalization from 1989-2003. Third, Prowess records detailed product-level information for each firm. This enables us to distinguish between single-product and multi-product firms, and track changes in firm scope over the sample period. Fourth, Prowess collects information on quantity and sales for each reported product, so we can construct the prices of each product a firm manufactures. These advantages make Prowess particularly well-suited for understanding the mechanisms of firm-level adjustments in response to trade liberalizations that are typically hidden in other data sources, and deal with measurement issues that arise in most studies that estimate production functions.

Prowess enables us to track firms’ product mix over time because Indian firms are required by the 1956 Companies Act to disclose product-level information on capacities, production and sales in their annual reports. As discussed extensively in Goldberg et al. (2010b), several features of the database give us confidence in its quality. Product-level information is available for 85 percent of the manufacturing firms, which collectively account for more than 90 percent of Prowess’ manufacturing
output and exports. Since product-level information and overall output are reported in separate modules, we can cross-check the consistency of the data. Product-level sales comprise 99 percent of the (independently) reported manufacturing sales. We refer the reader to Appendix C and Goldberg et al. (2010a,b) for a more detailed discussion of the data.

The definition of a product is based on the CMIE's internal product classification. There are 1,400 products in the sample for estimation. Table 1 reports basic summary statistics by two-digit NIC (India’s industrial classification system) sector. As a comparison, the U.S. data used by Bernard et al. (2010), contain approximately 1,500 products, defined as five-digit SIC codes across 455 four-digit SIC industries. Thus, our definition of a product is similar to earlier work that has focused on the U.S. Table 2 provides a few examples of products available in our data set. In our terminology, we will distinguish between “sectors” (which correspond to two-digit NIC aggregates), “industries” (which correspond to four-digit NIC aggregates) and “products” (the finest disaggregation we observe); we emphasize that since the “product” definition is available at a highly disaggregated level, unit values are plausibly interpreted as “prices” in our application.

The data also have some disadvantages. Unlike Census data, the CMIE database is not well suited for understanding firm entry and exit. However, Prowess contains mainly medium large Indian firms, so entry and exit is not necessarily an important margin for understanding the process of adjustment to increased openness within this subset of the manufacturing sector.

We complement the production data with tariff rates from 1987 to 2001. The tariff data are reported at the six-digit Harmonized System (HS) level and were compiled by Topalova (2010). We pass the tariff data through India’s input-output matrix for 1993-94 to construct input tariffs. We concord the tariffs to India’s national industrial classification (NIC) schedule developed by Debroy and Santhanam (1993). Formally, input tariffs are defined as $\tau_{it}^{\text{input}} = \sum_k a_{ki} \tau_{kt}^{\text{output}}$, where $\tau_{kt}^{\text{output}}$ is the tariff on industry $k$ at time $t$, and $a_{ki}$ is the share of industry $k$ in the value of industry $i$.

2.2 India’s Trade Liberalization

A key advantage of our approach is that we examine the impact of openness by relying on changes in trade costs induced by a large-scale trade liberalization. India’s post-independence development strategy was one of national self-sufficiency and heavy government regulation of the economy. India’s trade regime was amongst the most restrictive in Asia, with high nominal tariffs and non-tariff barriers. In response to a balance-of-payments crisis, India launched a dramatic liberalization of the economy as part of an IMF structural adjustment program in August 1991. An important part of this reform was to abandon the extremely restrictive trade policies it had pursued since independence.

Several features of the trade reform are crucial to our study. First, the external crisis of 1991,
which came as a surprise, opened the way for market oriented reforms (Hasan et al. (2007)). The liberalization of the trade policy was therefore unanticipated by firms in India and not foreseen in their decisions prior to the reform. Moreover, reforms were passed quickly as sort of a “shock therapy” with little debate or analysis to avoid the inevitable political opposition (see Goyal (1996)). Industries with the highest tariffs received the largest tariff cuts implying that both the average and standard deviation of tariffs across industries fell.

While there was significant variation in the tariff changes across industries, Topalova and Khandelwal (2011) show that tariff changes through 1997 were uncorrelated with pre-reform firm and industry characteristics such as productivity, size, output growth during the 1980s and capital intensity. The tariff liberalization does not appear to have been targeted towards specific industries and appears relatively free of usual political economy pressures until 1997 (which coincides with an election that changed political power). We estimate the production function and markups on the full sample, but restrict our analysis of the trade reform to the 1989-1997 period when trade policy did not respond to pre-existing industry- or firm-level trends. We again refer the reader to previous publications that have used this trade reform for a detailed discussion (Topalova and Khandelwal (2011); Topalova (2010); Sivadasan (2009); Goldberg et al. (2010a,b)).

3 Methodology: Recovering Markups and Marginal Costs

This section describes the framework to estimate markups and marginal costs using product- and firm-level production data. Section 3.1 presents the theoretical framework and explicitly states the assumptions required to implement the approach. The computation of markups and marginal costs requires estimates of production function coefficients and information about the allocation of inputs across products. Section 3.2 describes the methodology to estimate the production function and identification. Once the production function parameters are estimated, Section 3.3 explains how we recover the allocation of inputs across products for multi-product firms. In section 3.4 we discuss how we compute markups and marginal costs. Section 3.5 comments on the assumptions required to implement our methodology.

3.1 Theoretical Framework

Consider a production function for a firm $f$ producing a product $j$ at time $t$:

$$Q_{fjt} = F_{jt}(V_{fjt}, K_{fjt})\Omega_{ft}$$  

(1)

where $Q$ is physical output, $V$ is a vector of variable inputs that the firm can freely adjust and $K$ is a vector of fixed inputs that face adjustment costs. The firm’s productivity is denoted $\Omega_{ft}$. A firm

---

8Some commentators (e.g., Panagariya (2008)) noted that once the balance of payments crisis ensued, market-based reforms were inevitable. While the general direction of the reforms may have been anticipated, the precise changes in tariffs were not. Our empirical strategy accounts for this shift in broad anticipation of the reforms, but exploits variation in the sizes of the tariff cuts across industries.
produces a discrete number of products $J_{ft}$. Collect the inputs into a vector $X = \{V, K\}$. Let $W_{fjvt}$ denote the price of a variable input $v$ and $W_{fjkt}$ denote the price of a dynamic input $k$, with $v = \{1, ..., V\}$ and $k = \{1, ..., K\}$.

We begin by characterizing conceptual assumptions necessary to estimate markups and marginal costs for multi-product firms. We refer to these assumptions as conceptual because they are independent of the particular data and setting. Implementing the approach requires additional assumptions dictated by particular features of our data and our focus on India’s trade reforms (e.g., functional form and identification assumptions), and we describe these in the next section. The approach requires the following assumptions:

**Assumption 1: The production technology is product-specific.** Our notation reflects this assumption. The production function $F(.)$ is indexed by product $j$. This assumption implies that a single-product firm and a multi-product firm that produce the same product have the same production technology, although their productivities $\Omega_{ft}$ might differ.

**Assumption 2: $F_{jt}(.)$ is continuous and twice differentiable w.r.t. at least one element of $V_{fj}$, and this element of $V_{fj}$ is a static (i.e., freely adjustable or variable) input in the production of product $j$.** This assumption restricts the technology so that the firm can adjust its output quantity by changing a particular variable input.\(^9\) Furthermore, this assumption implies that firm cost minimization involves at least one static first order condition with respect to a variable input of production.

**Assumption 3: Hicks-neutral productivity $\Omega_{ft}$ is log-additive and firm-specific.** This assumption implies that a multi-product firm has the same productivity $\Omega_{ft}$ in the production of all its products.\(^{10}\) This assumption follows the tradition of modeling productivity in the multi-product firm literature in this manner (e.g., Bernard et al. (2011)). For single-product firms, this assumption is of course redundant.

**Assumption 4: Expenditures on all variable and fixed inputs are attributable to products.** This assumption implies that we can always write the expenditure on input $X$ attributable to product $j$ as $W_{fjtx_{fj}} = \tilde{\rho}_{fjt} \sum_j (W_{fjxt_{fj}})$ where $W_{fjxt}$ is the price for input $X$ with $X \in X$, and $\tilde{\rho}_{fjt}$ is the share of input expenditures attributable to product $j$ with the restriction that $\sum_j \tilde{\rho}_{fjt} = 1$. Note that $\tilde{\rho}_{fjt}$ is not observed in the data. Assumption 4 allows for economies (or diseconomies) of scope in costs of production; we discuss this distinction below in Section 3.5.

---

\(^9\)Assumption 2 rules out a fixed proportion technology (e.g., Leontief) in all variable inputs. The assumption seems reasonable at the level of aggregation of our data. We observe total labor, capital and intermediate inputs at the firm level, and so there is ample room for firms to substitute, say, workers for capital while keeping output constant.

\(^{10}\)In principle, we can allow for $F_{jt}(V_{fjt}, K_{fjt}, \Omega_{fjt})$ to derive a theoretical expression for markups. However, assumption 3 is required to estimate markups for multi-product firms.
Assumption 5: The state variables of the firm are

\[ s_{ft} = \{J_{ft}, K_{f,j=1,t}, \ldots, K_{f,J_{ft},t}, \Omega_{ft}, G_f, r_{ft} \} \]

The state variables include the number of products produced \((J_{ft})\), the dynamic inputs for all products \((K_{fjt})\), productivity \((\Omega_{ft})\), location information \((G_f)\), and all payoff relevant serially correlated variables, such as tariffs and the firm’s export status \((EXP_{ft})\), which we collect in \(r_{ft}\).

Assumption 6: Firms minimize short-run costs taking output quantity and input prices \(W_{fjt}\) at time \(t\) as given. Firms face a vector of variable input prices \(W^v_{fjt} = W^v_i(\nu_{fjt}, G_f, a_{fjt-1})\), which depends on the quality \(\nu_{fjt}\) of product \(j\), exogenous factors \(G_f\) (e.g., geography), and firm/product-level actions \(a_{fjt-1}\) taken prior to time \(t\). The latter can capture pre-negotiated input prices through contracts, for example, as long as the contracts do not specify input prices as a function of input quantities. The important assumption is that a firm’s variable input price does not depend on input quantity. This assumption rules out static sources of market power in input markets. We discuss this assumption in more detail at the end of this subsection.

We consider the firm’s cost minimization problem conditioning on state variables. From assumptions 2 and 6, firms minimize costs with respect to variable inputs. Assumptions 4 and 6 imply that costs are separable across products since a firm’s product mix is a dynamic choice and pre-determined at time \(t\) when variable inputs are chosen. Hence, we can minimize costs product-by-product for multi-product firms.

The associated Lagrangian function for any product \(j\) at time \(t\) is:

\[
L(V_{fjt}, K_{fjt}, \lambda_{fjt}) = \sum_{v=1}^{V} W^v_{fjt} V^{v}_{fjt} + \sum_{k=1}^{K} W^k_{fjt} K^{k}_{fjt} + \lambda_{fjt} [Q_{fjt} - Q_{fjt}(V_{fjt}, K_{fjt}, \Omega_{ft})] \tag{2}
\]

The first order condition for any variable input \(V^v_{fjt}\) used on product \(j\), is

\[
\frac{\partial L_{fjt}}{\partial V^v_{fjt}} = W^v_{fjt} - \lambda_{fjt} \frac{\partial Q_{fjt}(\cdot)}{\partial V^v_{fjt}} = 0, \tag{3}
\]

where the marginal cost of production at a given level of output is \(\lambda_{fjt}\) since \(\frac{\partial L_{fjt}}{\partial Q_{fjt}} = \lambda_{fjt}\). Rearranging terms and multiplying both sides by \(\frac{V_{fjt}}{Q_{fjt}}\), provides the following expression:

\[
\frac{\partial Q_{fjt}(\cdot)}{\partial V^v_{fjt}} \frac{V^{v}_{fjt}}{Q_{fjt}} = \frac{1}{\lambda_{fjt}} \frac{W^v_{fjt} V^{v}_{fjt}}{Q_{fjt}}. \tag{4}
\]

The left-hand side of the above equation represents the elasticity of output with respect to variable input \(V^v_{fjt}\) (the “output elasticity”). Define the markup \(\mu_{fjt}\) as \(\mu_{fjt} \equiv \frac{P_{fjt}}{\lambda_{fjt}}\).

This cost-minimization condition can be rearranged to express the markup for each product \(j\)
\[
\mu_{fjt} = \theta^v_{fjt} \left( \frac{P_{fjt} Q_{fjt}}{W^v_{fjt} V^v_{fjt}} \right) = \theta^v_{fjt} (\alpha^v_{fjt})^{-1}
\]

where \( \theta^v_{fjt} \) denotes the output elasticity on variable input \( V^v \) and \( \alpha^v_{fjt} \) is the share of expenditure on input \( V^v \) allocated to product \( j \) in the total sales of product \( j \). This expression forms the basis for our approach to compute markups. To compute the markup, we need the output elasticity on \( V^v \) for product \( j \), and the share of the input’s expenditure allocated to product \( j \) in the total sales of product \( j \), \( \alpha^v_{fjt} \).

The expression for the markup in (5) looks similar to the one derived in De Loecker and Warzynski (2012) with one crucial difference: all variables are indexed by \( j \). This seemingly small distinction has significant ramifications for the analysis and precludes us from using the existing approach in De Loecker and Warzynski (2012) to obtain the subcomponents of (5). De Loecker and Warzynski (2012) focus on firm-level markups and implement the conventional production function methodology using revenue data. Because of their focus and data, they do not need to confront the challenges posed by multi-product firms. Specifically, the firm-specific expenditures shares are directly observed in their data and the output elasticity is obtained by estimating a firm-level production function using deflated revenues. In contrast, our framework utilizes product-specific information on quantities and prices. This forces us to conduct the analysis at the product-level because aggregation to the firm-level is not possible without an explicit model of market demand.

The focus on products rather than firms calls for an explicit treatment of multi-product firms. In a multi-product setting, both components in equation (5) are unobserved. In contrast to a single-product firm setting, we must estimate the output elasticity separately for each product manufactured by each firm. Furthermore, the product-specific input expenditure shares \( \alpha^v_{fjt} \) cannot be calculated from the data because firms do not report the input expenditure allocations \( \tilde{\rho}_{fjt} \).\(^{11}\) Our framework, presented below, confronts these two challenges by proposing a methodology for estimating production functions that explicitly deals with multi-product firms and allows one to impute the input expenditure allocations across the products of a multi-product firm.

An additional advantage of focusing on products rather than firms is that once we derive estimates of product-level markups, we can calculate marginal costs using information on product-level prices, which are observed directly in the data:

\[
mc_{fjt} = \frac{P_{fjt}}{\mu_{fjt}}.
\]

A brief discussion of the assumptions underlying the analysis is in order. Assumptions 1-5 have been explicitly or implicitly assumed throughout the literature estimating production functions.\(^{12}\) For example, Assumption 1 is made implicitly whenever researchers pool single- and multi-product firm data to estimate production functions, which is almost always the case. The only difference is

\(^{11}\)We are unaware of any data set that provides this information for all inputs.

\(^{12}\)See Ackerberg et al. (2006) for an overview of this literature.
that the standard approach uses firm-level deflated sales and expenditure data; this practice does not force the researcher to confront multi-product firms in the data since the analysis is conducted at the firm level. Our framework strictly nests this approach, but since we use price data, and because prices are only defined at the product level (unless one is willing to make additional assumptions on demand that will allow aggregation to the firm level), we must specify physical production functions at the product level. We therefore explicitly state the assumptions that underlie the treatment of multi-product firms (Assumptions 1, 3 and 4).

Variants of Assumption 4 have been invoked in the few studies that have addressed the price bias in production function estimation (e.g., Foster et al. (2008) and De Loecker (2011)). Foster et al. (2008) allocate input expenditures according to revenue shares, while De Loecker (2011) allocates them based on the number of products. These variants are considerably stronger than, and are strictly nested within, Assumption 4. Relaxing these input allocation assumptions is one of the methodological contributions of this paper.

The product-by-product short-run cost minimization with respect to variable inputs in (2) follows from Assumptions 2, 4 and 6. Assumption 2 assures the existence of a variable input and is essential for our approach. If all inputs are dynamic, we can still estimate the production function, but we cannot derive markups using the approach we described above. However, the assumption that there is at least one factor of production that the firm can freely adjust over the period of a year (we have annual production data) is both plausible and standard in empirical work.

Our framework allows for economies (or diseconomies) of scope. While physical synergies in production are ruled out by Assumption 1, other forms of economies (or diseconomies) of scope are consistent with Assumptions 1 and 4. Economies of scope can operate through the Hicks-neutral productivity shocks $\Omega_{ft}$, by spreading of the fixed costs associated with dynamic inputs (e.g., capital) across multiple products, and/or through pre-negotiated firm-level contracts for input prices $W_{vft}$, as long as these input prices do not depend on quantity of inputs. We discuss economies of scope in more detail in Section 3.5.

Finally, an important assumption we maintain throughout the analysis is that input prices do not depend on input quantities (Assumption 6). While restrictive, this assumption is more general than the one employed in almost all production function studies, in which it is assumed that all firms face the same input prices (in contrast, we allow for input prices to differ across firms because of locational differences and/or quality differentiation). If firms have monopsony power in input markets, Assumption 6 will be violated. In this case, one can show that our approach will tend to understate the level of markups. However, the approach can still be used to trace and explain changes in markups, as long as there are no contemporaneous changes in firms’ monopsony power, or, even if there are such changes, as long as changes in firms’ monopsony power are uncorrelated with trade policy changes. Appendix D provides a detailed discussion of the conditions under which our approach is valid in the case of monopsony power.\footnote{That is, $J^{-1} \sum_{j} W_{vft} K_{fj}$ falls as the number of products increases.}

\footnote{In principle, one could make the argument that trade policy might lead to exit of smaller, less productive firms, which might give monopsony power to the remaining firms in the market. In practice, we do not observe firm exit in}
In sum, our approach to recover estimates of markups and marginal costs requires estimates of the parameters of the production function $F_{jt}(.)$ at the product level and the input allocations $\hat{\rho}_{fjt}$ across products within each multi-product firm. Section 3.2 discusses the production function estimation method and the identification strategy we employ in order to obtain the output elasticities for both single- and multi-product firms.

### 3.2 Estimation

We take logs of equation (1) and allow for log-additive measurement error and/or unanticipated shocks to output ($\epsilon_{fjt}$). Log output is given by: $q_{fjt} = \ln(Q_{fjt}\exp(\epsilon_{fjt}))$. Letting $x_{fjt}$ be the vector of (log) physical inputs, $x_{fjt} = \{v_{fjt}, k_{fjt}\}$, and $\omega_{ft}$ be $\ln(\Omega_{ft})$, we obtain:

$$q_{fjt} = f_{jt}(x_{fjt}; \beta) + \omega_{ft} + \epsilon_{fjt}. \quad (7)$$

By writing the production function in terms of physical output rather than revenue, we exploit separate information on quantities and prices that are available in the data. The use of physical output in equation (7) eliminates concerns of a price bias that arises if output is constructed by deflating firm revenues by an industry-level price index.\(^{15}\)

Unobserved productivity $\omega_{ft}$ potentially leads to well known simultaneity and selection biases. These two biases have been the predominant focus of the production function estimating literature and we follow the insights of Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2006) in addressing them. Note that if we theoretically had data on the physical inputs $(v_{fjt}, k_{fjt})$ for all products, these existing approaches to estimating production functions would in principle suffice to obtain consistent estimates of the production function coefficients $\beta$.

In reality, no dataset records product-specific inputs, so estimating equation (7) requires dealing with two additional issues: (a) we do not observe input allocations across products in multi-product firms; and (b) we observe industry-wide deflated firm-level input expenditures rather than firm-level input quantities. The latter is not merely a measurement problem because firms typically rely on differentiated inputs to manufacture differentiated products, so physical input and output are not readily comparable across firms.

To understand the implications of these two issues for estimation, let $\hat{x}_{ft}$ denote the (observed) vector of deflated input expenditures, deflated by a sector-specific price index. From Assumption 4, product-level input quantities, $x_{fjt}$, for each input $x$ relate to firm-level expenditures as follows:

$$x_{fjt} = \rho_{fjt} + \hat{x}_{ft} - w^{x}_{fjt} \quad (8)$$

where $\rho_{fjt} = \ln(\hat{\rho}_{fjt})$ is the (log) share of firm input expenditures allocated to product $j$ and $w^{x}_{fjt}$ denote our sample, so we do not consider such a scenario as a likely explanation for our empirical results. We have explored heterogeneity in our results by identifying business groups in our sample who may have some degree of monopsony power, but we do not find differential effects with respect to the impacts of tariffs on their prices, markups and marginal costs (results available upon request).\(^{15}\)

\(^{15}\)For a detailed discussion, see De Loecker (2011) and Foster et al. (2008).
notes the deviation of the unobserved (log) firm-product-specific input price from the (log) industry-wide input price index. By substituting this expression for physical inputs into equation (7) and defining \( w_{fjt} \) as the vector of log firm-product-specific input prices, we obtain:

\[
q_{fjt} = f_{jt}(\tilde{x}_{ft}; \beta) + A_t(\rho_{fjt}, \tilde{x}_{ft}, \beta) + B_t(w_{fjt}, \rho_{fjt}, \tilde{x}_{ft}, \beta) + \omega_{ft} + \epsilon_{fjt} \tag{9}
\]

Compared to equation (7), there are two additional unobserved terms in (9). First, the term \( A_t(.) \) that arises from the unobserved product-level input allocations \( \rho_{fjt} \) and second, the term \( B_t(.) \) that captures unobserved firm-product-specific input prices \( w_{fjt} \). The exact form of terms \( A(.) \) and \( B(.) \) depends on the functional form of \( f(.) \). Both terms depend on the vector of coefficients \( \beta \), the input expenditures \( \tilde{x}_{ft} \), and the unobserved product-level input allocation shares \( \rho_{fjt} \). It is evident from (9) that even after controlling for the unobserved productivity \( \omega_{ft} \) using standard estimation techniques, the presence of the terms \( A(.) \) and \( B(.) \) leads to biased production function coefficients since both terms are correlated with the deflated input expenditures \( \tilde{x}_{ft} \). We refer to the bias arising from the term \( A(.) \) as the "input allocation" bias and the bias arising from \( B(.) \) as the "input price" bias. The methodology we develop in this subsection addresses these biases.

Neither the "input allocation" nor the "input price" bias have received much attention in the literature on production function estimation to date because the standard practice regresses deflated sales on deflated expenditures at the firm level. \(^{18}\) \( \) De Loecker and Goldberg (2014) discuss the conditions under which these biases interact so as to produce reasonable estimates. But although such estimates may look plausible, this does not imply that the coefficients are consistent estimates of the production function. Failing to correct these biases traces the elasticity of sales with respect to input expenditures, but that elasticity is not useful in our approach because equation (5) requires the elasticity of output quantities with respect to input quantities.

To deal with these biases, we proceed in four steps. Subsection 3.2.1 explains how the estimation addresses the unobserved input allocation bias. Subsection 3.2.2 explains how to address the bias arising from unobserved input prices. Subsection 3.2.3 explains our treatment of the unobserved productivity shock and selection correction. Subsection 3.2.4 explains the moment conditions and further elaborates on identification and estimation. The first two steps are new to the literature on production function estimation; the last two steps build on existing work.

\(^{16}\)We allow for multi-product firms to face different input prices in the production of their various products. Accordingly, the input prices \( w \) are indexed by both \( f \) and \( j \). This would be the case if a multi-product firm manufactured products of different qualities that relied on inputs of different qualities; see subsection 3.2.2 for a discussion of the relationship between output and input quality.

\(^{17}\)To simplify notation, we will always use \( w_{fjt} \) to denote the deviations of firm-product-specific input prices from industry input price indexes. Similarly, from now on, we will use the term "firm input prices" to denote firm-specific deviations from industry averages.

\(^{18}\)Katayama et al. (2009) is the only study to our knowledge that acknowledges the existence of the input price bias.
3.2.1 Unobserved Input Allocations: The Use of Single-Product Firms

Assumptions 1 and 4 imply that a firm’s technology used to produce product $j$ is independent of the other products manufactured by the firm. This also implies that a multi-product firm uses the same technology as a single-product firm producing the same product.\(^{19}\) We can therefore rely on single-product firms to estimate the product-level production function in (9), without having to address the unobserved input allocations in multi-product firms. For single-product firms, $A(\cdot) = 0$ because by definition, $\hat{\rho}_{fjt} = 1$. Since estimation is based on the single-product sample, we omit the product subscript $j$ for the remainder of the exposition of the estimation algorithm.

Equation (9) simplifies to:

$$q_{ft} = f_t(\tilde{x}_{ft}; \beta) + B_t(w_{ft}, \tilde{x}_{ft}; \beta) + \omega_{ft} + \epsilon_{ft}. \quad (10)$$

The approach of using the single-product firm estimates to infer the production function coefficients for all firms raises the concern that the estimates may suffer from a selection bias since we rely only on single-product firms in the estimation. The selection bias arises if firms’ choice to add a second product and become multi-product depends on the unobserved firm productivity $\omega_{ft}$ and/or firms’ input use. Our estimation procedure utilizes the selection correction insights from Olley and Pakes (1996) to address this potential selection bias in two ways. First, we use an unbalanced panel that consists of firms that are single-product at a given point in time. At time $t$, the unbalanced panel includes both firms who always remain single-product firms and those that manufacture a single product at $t$ but add additional products at a later date. This feature of the sample is important since many firms start off as single-product firms and add products during our sample. The use of the unbalanced panel fully addresses the non-random event that a firm becomes a multi-product producer based on unobserved productivity $\omega_{ft}$.\(^{20}\) Second, to account for the possibility that the productivity threshold determining the transition of a firm from single- to multi-product status is correlated with production inputs (in particular, capital), we additionally apply a sample selection correction procedure. We describe the details of the sample selection correction procedure in subsection 3.2.3.\(^{21}\)

As the notation in (10) indicates, it is in principle possible to estimate separate production functions by year. In practice, our sample is not large enough to allow for time-varying production functions. Therefore, the production function we take to the data is not indexed by $t$. We consider

---

\(^{19}\)For example, imagine a single-product firm produces a t-shirt using a particular technology, and another single-product firm produces carpets using a different combination of inputs. We assume that a multi-product firm that manufactures both products will use each technology on its respective product, rather than some third technology.

\(^{20}\)This non-random event of adding a second product results in a sample selection issue analogous to the non-random exit of firms discussed in Olley and Pakes (1996). In their context, Olley and Pakes (1996) are concerned about the left tail of the productivity distribution; here, a balanced panel of single-product firms would censor the right tail of the productivity distribution. The use of the unbalanced panel of single-product firms improves upon this selection problem.

\(^{21}\)Firms in our sample very rarely drop products, so we do not observe the reverse transition from multi- to single-product status. We refer the reader to Goldberg et al. (2010b) for a detailed analysis of product adding and dropping in our data. Unlike Olley and Pakes (1996), we are also not concerned with firm exit. Firm exit is rare in our data because Prowess covers the medium and large firms in India.
three inputs in the (deflated) input expenditure vector $\tilde{x}_{ft}$: labor ($\tilde{l}$), intermediate inputs ($\tilde{m}$) and capital ($\tilde{k}$). It is clear from equation (10) that we still need to correct for the term related to unobserved firm-specific input price variation, $B(w_{ft}, \tilde{x}_{ft}, \beta)$, and the unobserved firm-level productivity ($\omega_{ft}$) in order to obtain consistent estimates of the production function parameters $\beta$, and hence the output elasticities that are used to compute markups and marginal costs. We turn to these issues next.

### 3.2.2 Unobserved Input Prices

The treatment of unobserved input prices is important for two reasons. First, we need to control for them in $B(w_{ft}, \tilde{x}_{ft}, \beta)$ in equation (10) to recover consistent estimates of the production function parameters $\beta$.\(^{22}\) Second, the input demand equation that is used to control for productivity $\omega_{ft}$ naturally depends on input prices (see next subsection 3.2.3).

In our framework (see Assumption 6), firm-specific input price variation can arise through exogenous variation in input prices across local input markets ($G_f$) and/or variation in input quality ($\nu_{ft}$).\(^{23}\) This implies that two firms in the same industry that produce in the same location only face the exact same input prices if they buy the exact same input quality. We propose an approach to control for unobserved input price variation across firms using information on observables, particularly (but not exclusively) output prices. The intuition is that output prices contain information about input prices. For example, using data from Colombia that uniquely record price information for both inputs and outputs, Kugler and Verhoogen (2011) document that producers of more expensive products also use more expensive inputs.

We provide a formal model that rationalizes our approach to control for input prices in Appendix A. We show that in a large class of models of consumer demand and imperfect competition used in the Industrial Organization and International Trade literatures, we can proxy for unobserved input prices using a function of the firm’s output price, market share, and product dummies. Here, we sketch the main argument and provide the economic intuition underlying our empirical strategy.

The main premise is that manufacturing high quality products requires high quality inputs, and that high quality inputs are expensive. We further assume complementarity in input quality: manufacturing high quality products requires combining high quality materials with high quality labor and capital. This is a common assumption in the literature and underlies ‘O-Ring’-type theories of production (e.g., Kremer (1993), Verhoogen (2008) and Kugler and Verhoogen (2011)). This complementarity implies that the prices of all inputs facing a firm can be expressed as a function of a single index of product quality. Appendix A shows that input prices are an increasing function of product quality in this setting. Accordingly, we can control for input price variation across firms using differences in output quality across firms.

\(^{22}\)This subsection considers single-product firms since we use only these firms to estimate the production functions, but all relationships described below also apply to multi-product firms (in which case all relevant variables should be indexed by $j$).

\(^{23}\)We abstract from lagged action variables $a_{ft-1}$, since we do not have rich enough data to measure these (e.g., past contracts specifying input prices independent of quantities).
Given that input prices are an increasing function of input quality, which is an increasing function of output quality, we can use the variables proxying for output quality (i.e., output price, market share and product dummies) to proxy for input prices. Formally, we write input prices $w_{ft}^x$ as a function of output quality $\nu_{ft}$ and firm location $G_f$:\footnote{We remind the reader that we have defined the input price $w_{ft}^x$ for input $x$ as the deviation of the actual input price from the relevant input price index (i.e., the weighted industry mean), and therefore $w_{ft}^x = 0$ for the producer paying exactly the (weighted) average $\bar{w}^x_t$. Formally $w_{ft}^x = w_{ft}^x - \bar{w}^x_t$, where $^*$ denotes the actual input price faced by firm $f$ for its product $j$ at time $t$.}

$$w_{ft}^x = w_t(\nu_{ft}, G_f).$$ (11)

This expression for input prices generalizes Assumption 6 to all inputs. Under the assumption of input quality complementarity, the unobserved input price variation across all inputs can be captured by a single control function.

Using the results from Appendix A we get:

$$w_{ft}^x = w_t(p_{ft}, m_{sf}, D_f, G_f, EXP_{ft}),$$ (12)

where $p_{ft}$ is the output price of the firm, $m_{sf}$ is a vector of market shares, $D_f$ captures the vector of product dummies, and $EXP_{ft}$ denotes the export status of a firm.\footnote{We include the export status of a firm to allow for market demand conditions to differ from the domestic market. In our data we do not observe the product-destination trade flows for each firm. Otherwise this information could be included here.} It is important to note that our approach to control for unobserved input quality does not assume that products are only vertically differentiated. It allows for horizontal differentiation, but horizontal differentiation is costless. In contrast, differentiation along the vertical dimension requires higher quality inputs that have higher input prices. This assumption is common in trade models (e.g., Verhoogen (2008) and Khandelwal (2010)). Moreover, because we model output quality as a flexible function of output prices, market share, and product dummies, the approach does not commit to a particular demand function since it encompasses a large class of demand models used in the literature. For example, in a purely vertical differentiation model, there is a one-to-one mapping between product quality and product prices, so output prices perfectly proxy for quality; in this case, one would not require controls for market share or product characteristics. In the simple logit model, quality is a function of output prices and market shares (see Khandelwal (2010) for a detailed exposition). In more general models, such as the nested logit or random coefficients models, quality is a function of additional variables, such as product characteristics, conditional market shares, etc. While product characteristics usually cannot be observed in firm-level data, product dummies accommodate these more general demand specifications as in Berry (1994). Finally, using output prices as a proxy for quality does not imply that we assume complete pass-through of input to output prices; the degree of pass-through is dictated by the (unspecified) underlying demand and market structure and by the firm behavioral assumptions. Accordingly, the approach is consistent with any degree of pass-through between input and output prices.
The final step is to substitute the input price control function from (12) into the expression for $w_{ft}$ in $B(w_{ft}, \tilde{x}_{ft}, \beta)$ in equation (10), we get:

$$B(w_{ft}, \tilde{x}_{ft}, \beta) = B((p_{ft}, ms_{ft}, D_f, G_f, EXP_{ft}) \times \tilde{x}_{ft}; \beta, \delta) \quad (13)$$

A few words on notation are in order. The function $B(.)$ is different from the input price function $w(.)$ as described in equation (12). The function $B(.)$ depends on the input prices $w_{ft}$ and will therefore take as arguments the elements of $w(.)$. However, it also contains interactions of the input prices ($w_{ft}$) with the vector of deflated input expenditures $\tilde{x}_{ft}$. We use the notation $\tilde{x}_{ft}$ to highlight the fact that the input price term $w(.)$ enters also by itself, without being interacted with the input expenditures $\tilde{x}_{ft}$, and thus we include a constant term: $\tilde{x}_{ft} = \{1, \tilde{x}_{ft}\}$. The notation highlights that the use of the input price control function requires us to estimate an additional parameter vector $\delta$ alongside the production function parameters $\beta$.

### 3.2.3 Unobserved Productivity and Selection Correction

The only remaining source of potential bias in (10) is the unobserved firm-level productivity $\omega_{ft}$. Firms’ choices of inputs and number of products are in part affected by this (to the econometrician) unobserved productivity, potentially leading to simultaneity and selection bias in estimation. We control for unobserved productivity $\omega_{ft}$ in (10) using a control function based on a static input demand equation. In addition, we implement a selection correction for the potential selection bias stemming from the use of single-product firms in the estimation procedure, discussed in subsection 3.2.1. We describe both procedures here.

We follow the literature on production function estimation, as initiated by Olley and Pakes (1996) and extended by Levinsohn and Petrin (2003), and control for unobserved productivity $\omega_{ft}$ in (10) using a static input demand equation. The materials demand function in our setting will take as arguments all state variables of the firm noted in Assumption 5, including productivity, and all additional variables that affect a firm’s demand for materials. These include firm location ($G_f$), output prices ($p_{ft}$), product dummies ($D_f$), market shares ($ms_{ft}$), input prices ($w_{ft}$), the export status of a firm ($EXP_{ft}$) and the input ($\tau_{it}^\text{input}$) and output tariffs ($\tau_{it}^\text{output}$) that the firm faces on the product it produces. From (12) input prices are themselves a function of output price, market share and product dummies\(^{26}\), so materials demand is given by:

$$\tilde{m}_{ft} = m_t(\omega_{ft}, \tilde{k}_{ft}, \tilde{l}_{ft}, G_f, p_{ft}, D_f, ms_{ft}, EXP_{ft}, \tau_{it}^\text{input}, \tau_{it}^\text{output}) \quad (14)$$

We collect all the variables determining intermediate input demand, except for the input expenditures and unobserved productivity, in $z_{ft} = \{G_f, p_{ft}, D_f, ms_{ft}, EXP_{ft}, \tau_{it}^\text{input}, \tau_{it}^\text{output}\}$. The number of products ($J_{ft}$) is omitted from the set of state variables since the sample we use for

\(^{26}\text{Note that we consider (log) intermediate input expenditure, defined as the sum (in logs) of the intermediate input demand and the input price. This implies that the materials expenditure function }\tilde{m}_t(\cdot)\text{ takes as arguments the same variables as the physical materials demand function }m_t(\cdot): m_{ft} = m_t(w_{ft}^\text{input}, \cdot)\text{ and }\tilde{m}_{ft} = m_t(\cdot) + w_{ft}^\text{input} = \tilde{m}_t(w_{ft}^\text{input}, \cdot),\text{ where }w_{ft}^\text{input}\text{ is the input price.}
estimation contains only single-product firms. The subscript $i$ on the tariff variables denotes an industry to indicate that tariffs vary at a higher level of aggregation than products. Inverting (14) gives our control function for productivity:\(^{27}\)

$$\omega_{ft} = h_t(\tilde{x}_{ft}, z_{ft}).$$  \hspace{1cm} (15)

Our approach also encompasses a selection correction to address the potential selection bias stemming from the use of only single-product firms in the estimation discussed in subsection 3.2.1. The selection bias arises if a firm's choice to add a second product and become a multi-product firm depends on unobserved firm productivity $\omega_{ft}$ in equation (10) and/or the firm's input use. Following Olley and Pakes (1996), who address the selection bias due to plant exit in their setting, we model the probability that a firm continues to produce one product non-parametrically as a function of the firm's productivity forecast and all state variables $s_{ft}$.

The underlying model behind our sample selection correction is one where the number of products manufactured by firms increases with productivity. Several multi-product firm models generate this correlation, with Mayer et al. (2014) matching our setup most closely. In that model, the number of products a firm produces is an increasing step function of the firms' productivity. Firms have a productivity draw which determines their core product. Conditional on entry, the firm produces this core product and incurs an increasingly higher marginal cost of production for each additional product it manufactures. This structure generates a competence ladder that is characterized by a set of cutoff points, each associated with the introduction of an additional product.\(^{28}\)

The cutoff point relevant to our sample selection procedure is the one associated with the introduction of a second product. We denote this cutoff by $\tilde{\omega}_{ft}$. Firms with productivity that exceeds $\tilde{\omega}_{ft}$ are multi-product firms that produce two (or more) products while firms below $\tilde{\omega}_{ft}$ remain single-product producers and are included in the estimation sample.

If the threshold $\tilde{\omega}_{ft}$ is independent of the right-hand side variables in the production function in equation (10), there is no selection bias and we obtain consistent estimates of production function coefficients (as long as we use the unbalanced panel of single product firms, i.e., the sample of firms that are single-product at any point in time, but may become multi-product in the future). A bias arises when the threshold is a function of capital and/or labor. For example, it is possible that even conditional on productivity, a firm with more capital finds it easier to finance the introduction of

---

\(^{27}\)As discussed in Olley and Pakes (1996), the proxy approach does not require knowledge of the market structure for the input markets; it simply states that input demand depends on the firm's state variables and variables affecting input demand. By using a static control to proxy for productivity, we do not have to revisit the underlying dynamic model and prove invertibility when modifying Olley and Pakes (1996) for our setting to include additional state variables (e.g., tariffs). See De Loecker (2011) and Ackerberg et al. (2006) for an extensive discussion. A recent literature has discussed alternative estimation procedures that do not rely on this inversion. In the absence of shocks to output $\epsilon_{ft}$, this can be accomplished without any extra assumptions. However, these shocks end up being important, especially when estimating physical output production functions where they absorb unit fixed effects.

\(^{28}\)Alternative models such as Bernard et al. (2010) introduce firm-product-specific demand shocks that generate product switching (e.g., product addition and dropping) in each period. We avoid this additional complexity since product dropping is not a prominent feature of our data (Goldberg et al. (2010b)). Moreover, in Section 4 we find strong support that firms' marginal costs are lower on their core competent products (products that have higher sales shares).
an additional product; or, a firm that employs more workers may have an easier time expanding into new product lines. In these cases, firms with more capital and/or labor are less likely to be single-product firms, even conditional on productivity, and this generates a negative bias in the capital and labor coefficients.

To address the selection bias, we allow the threshold \( \bar{\omega}_{ft} \) to be a function of the state variables \( s_{ft} \) and the firm’s information set at time \( I_{t-1} \) (we assume the decision to add a product is made in the previous period). The selection rule requires that the firm make its decision to add a product based on a forecast of these variables in the future. Define an indicator function \( \chi_{ft} \) to be equal to 1 if the firm remains single-product (SP) and 0 otherwise. The selection rule can be written as:

\[
\Pr(\chi_{ft} = 1) = \Pr[\omega_{ft} \leq \bar{\omega}_{ft}(s_{ft}) | \omega_{ft-1}] = \kappa_{t-1}(\bar{\omega}_{ft}(s_{ft}), \omega_{ft-1}) = \kappa_{t-1}(\tilde{x}_{ft-1}, i_{ft-1}, z_{ft-1}) \equiv SP_{ft}
\]

Note that the variables included in \( z \) are a subset of the state variables that appear in \( s \) (the latter include the dynamic inputs that are part of \( \tilde{x} \)). We use the fact that the threshold at \( t \) is predicted using the firm’s state variables at \( t-1 \), the accumulation equation for capital, and \( \omega_{ft} = h_{t}(\tilde{x}_{ft}, z_{ft}) \) from equation (15) to arrive at the last equation.\(^{29}\) As in Olley and Pakes (1996), we have two different indexes of firm heterogeneity, the productivity and the productivity cutoff point. Note that \( SP_{ft} = \kappa_{t-1}(\omega_{ft-1}, \bar{\omega}_{ft}) \) and therefore \( \omega_{ft} = \kappa_{t-1}^{-1}(\omega_{ft-1}, SP_{ft}) \).

### 3.2.4 Productivity Process, Moment Conditions, and Identification

To estimate the parameter vectors \( \beta \) and \( \delta \), we follow Ackerberg et al. (2006) and form moments based on the innovation in the productivity shock \( \xi_{ft} \). We consider the following law of motion for productivity:

\[
\omega_{ft} = g(\omega_{ft-1}, \tau_{it-1}^{\text{output}}, \tau_{it-1}^{\text{input}}, EXP_{ft-1}, SP_{ft}) + \xi_{ft}.
\]

The tariff variables and export dummy are included in the law of motion to account for the fact that trade policy and exporting may affect productivity. As De Loecker (2013) shows, if one expects these variables to have an effect on productivity, then the theoretically consistent treatment is to include them directly in the law of motion. Otherwise, their omission may lead to biased production function coefficients. Of course, the fact that these variables are allowed to have an impact on productivity does not mean that they will in fact have an effect. It is entirely possible that the empirical estimates indicate that the trade variables have no effect on productivity. Hence, including trade variables in the law of motion does not assume a particular result regarding the

\[^{29}\text{The accumulation equation for capital is } K_{ft} = (1 - \delta)K_{ft-1} + I_{ft-1}, \text{ where } \delta \text{ is the depreciation rate of capital. This specification takes into account that firms hire and/or fire workers based on their labor force at time } t-1 \text{ and their forecast of future demand and costs captured by } z \text{ and } \omega. \text{ So all variables entering the nonparametric function } \kappa_{t-1}(\cdot) \text{ help predict the firm’s employment at time } t.\]
effects of tariffs or exporting on productivity.

Trade related variables are expected to affect productivity both through exporting and importing channels. For example, a large literature suggests “learning by exporting” effects. Likewise, trade economists have postulated that a reduction in output tariffs that exposes firms to intensified import competition may lead to reduction in X-inefficiencies and adoption of better management practices. In this case, output tariff reductions may lead to productivity improvements. On the input side, input tariff reductions may lead to the import of new, previously unavailable intermediate products, which will lead to increases in productivity (see Halpern et al. (2011) for a formalization of this argument). We emphasize that the specification we adopt for the law of motion for productivity in equation (18) allows for these mechanisms to generate productivity improvements, but by no means assumes the result. The inclusion of the probability that a firm remains single-product in the next period $SP_{ft}$ in the law of motion addresses the selection correction from equation (16).

To form moments based on the innovation in the productivity shock in (18), one needs to express the productivity $\omega_{ft}$ as a function of data and parameters. Plugging the expressions for the input price correction from (13) and for unobserved productivity from (15) into the production function equation (10), we get:

$$q_{ft} = \phi_t(\tilde{x}_{ft}, z_{ft}) + \epsilon_{ft}, \quad (19)$$

Estimation of (19) enables one to get rid of unanticipated shocks and/or measurement error $\epsilon_{ft}$. We note that although the variables proxying for input prices (see equation (12)) also enter the input demand equation in equation (15), this has no implications for the identification of the production function parameters. The only purpose of the first stage estimation is to purge the output quantity data from unanticipated shocks and/or measurement error (i.e., purge $\epsilon_{ft}$ in equation (10)). For example, output prices ($p_{ft}$) enter this first stage both to control for unobserved productivity and input price differences, but we do not need to distinguish between them when forecasting output. Note that even if we observed (quality-corrected) input prices, we would still include output prices and the function $\phi_t(\cdot)$ would reflect this.

The first stage of the estimation in (19) yields an estimate of predicted output $\hat{\phi}_{ft}$. One can then express productivity $\omega_{ft}$ as a function of data and parameters. In particular, using equations (10), (13) and (19) we have:

$$\omega_{ft}(\beta, \delta) = \hat{\phi}_{ft} - f(\tilde{x}_{ft}; \beta) - B((p_{ft}, ms_{ft}, D_{f}, G_{f}, EXP_{ft}) \times \tilde{x}_{ft}; \delta), \quad (20)$$

where the last term, the function $B(\cdot)$, represents the input price control function.\textsuperscript{32}

\textsuperscript{30}We could set $\epsilon_{ft} = 0$; in this case, we no longer need to invert the input demand function to control for unobserved productivity. However, we feel that the input demand specification addresses first-order empirical issues with the data: measurement error in output and differences in units across products within sectors, which are absorbed by unit fixed effects in the first stage.

\textsuperscript{31}In practice we approximate the function $\phi_t(\cdot)$ with a third-order polynomial in all its elements, with the exception of product dummies. We add the product dummies linearly to avoid having to estimate all cross terms. This seems innocuous since the first stage $R^2$ is very close to one.

\textsuperscript{32}We approximate $B(\cdot)$ with a flexible third-order polynomial. At this point the reader might find it useful to consider a special case of a Cobb-Douglas production function and a vertical differentiation model of consumer
It is important to note that even though the input expenditures $\bar{x}_{ft}$ enter both the production function $f(.)$ and the input price control function $B(.),$ the coefficients of the production function $\beta$ are identified because $\bar{x}_{ft}$ enter the input price control function in (13) only interacted with input prices, or put differently, the input expenditures do not enter the input price function $w(.)$ in (12). This identification insight does not rest on any functional form assumptions; it results from the fact that the control function for quality, and hence input prices, rests on the demand side alone and hence does not include input expenditures.

The main parameters of interest to compute markups are the vector of production function coefficients $\beta.$ However, from (13), note that the parameter vector $\delta$ allows us to identify the input prices: after we have estimated $\beta$ and $\delta,$ we can recover the input prices from equation (12). To estimate the parameter vectors $\beta$ and $\delta,$ we form moments based on the innovation in the productivity shock $\xi_{ft}$ in law of motion in equation (18). We use (20) to project $\omega_{ft}(.)$ on the elements of $g(.)$ to obtain the innovation $\xi_{ft}$ as a function of the parameters $\xi_{ft}(\beta, \delta):$

$$\xi_{ft}(\beta, \delta) = \omega_{ft}(\beta, \delta) - E \left( \omega_{ft-1}(\beta, \delta), \tau_{ft-1}^{\text{output}}, \tau_{ft-1}^{\text{input}}, EXP_{ft-1}, SP_{ft} \right)$$

(21)

The moments that identify the parameters are:

$$E(\xi_{ft}(\beta, \delta)Y_{ft}) = 0,$$

where $Y_{ft}$ contains lagged materials, current capital and labor, and their higher order and interaction terms, as well as lagged output prices, lagged market shares, lagged tariffs, and their appropriate interactions with the inputs.

This method identifies the production function coefficients by exploiting the fact that current shocks to productivity will immediately affect a firm’s materials choice while labor and capital do not immediately respond to these shocks; moreover, the degree of adjustment can vary across firms and time. These moments that rely on adjustment costs in inputs are by now standard in this literature. In our context, we assume that firms freely adjust materials and treat capital and labor as dynamic inputs that face adjustment costs. In other settings, one may choose to treat labor as a flexible input. Since materials are the flexible input, we use lagged materials when we construct moments.\cite{34}

We use lagged output prices, market shares, and tariffs and their interactions with appropriately lagged inputs to form additional moment conditions to identify jointly the production function coefficients $\beta$ and the coefficients $\delta$ capturing the input price variation. For example, the parameter related to the output price is identified off the moment $E(\xi_{fp}) = 0; \,$ this moment condition is based on the insight that current prices do react to productivity shocks, so we need to use lagged demands. In this special case equation (20) reduces to: $\omega_{ft}(\beta, \delta) = \phi_{ft} - \bar{x}_{ft}^\beta - \Gamma w_{ft}(p_{ft}; \delta),$ where $\Gamma$ denotes the returns to scale parameter. Please see Appendix B for details.\cite{33}

\cite{33}In other words, we specify the function $w(.)$ and therefore the $\delta$ parameters are a function of both the production function coefficients $\beta,$ and the parameters in $w(.)$.

\cite{34}In our setting, input tariffs are serially correlated and since they affect input prices, input prices are serially correlated over time, creating a link between current and lagged intermediate input usage.
output prices which exploit the serial correlation of prices.

We estimate the model using a GMM procedure on a sample of firms that manufacture a single product for at least three consecutive years.\textsuperscript{35} We choose three years since the moment conditions require at least two years of data because of the lagged values; we add an additional (third) year to allow for potential measurement error in the precise timing of a new product introduction. We discuss the timing assumptions further in subsection 3.5.2. In principle, one could run the estimation separately for each product. In practice, our sample size is too small to allow estimation at the product level, so we estimate (10) at the two-digit sector level.\textsuperscript{36}

Estimation of equation (10) requires choosing a functional form for $f$. We adopt a translog specification because of its flexibility.\textsuperscript{37} Specifically, the translog offers the advantage that it generates output elasticities that are not constant over time and across firms (though the production coefficients are constrained to be the same across years and firms); hence, large firms can have different elasticities than small firms. The exact functional form for $f(.)$ does not generate any identification results. The crucial assumption is that productivity enters in a log-additive fashion (Assumption 3 in Section 3.1).

Finally, the standard errors on the coefficients are obtained using block-bootstrapping, where we draw an entire firm time series. Since our ultimate objective is to estimate the impact of the trade reforms on markups and marginal costs, we correct the standard errors of the regressions in Section 4 by block-bootstrapping over our entire empirical procedure.

### 3.3 Recovering Input Allocations

As shown in equations (5) and (6), computing markups and marginal costs requires the product-specific output elasticity and product-specific revenue shares on a variable input (in our case, materials). We obtain the output elasticity from the estimation outlined in Section 3.2 based on single-product firms, but we do not know the product-specific revenue shares of inputs for multi-product firms. Here, we show how to compute the input allocations across products of a multi-product firm in order to construct $\alpha^{M}_{jt}$.

From Assumption 6, recall that $\rho_{fjt} = \ln \left( \frac{W_{fjt}X_{fjt}}{X_{ft}} \right)$ $\forall X \in \{V, K\}$, is product $j$’s input cost share. We solve for $\rho_{fjt}$ in multi-product firms as follows. We first eliminate unanticipated shocks and measurement error from the product-level output data by following the same procedure as in the first stage of our estimation routine for the single-product firms in (19). We project $q_{fjt}$ on the exact same variables used in the first stage of the estimation procedure, $\hat{q}_{fjt} \equiv E([q_{fjt} | \phi_{t}(\tilde{x}_{ft}, z_{ft})])$, which allows us to eliminate any measurement error and unanticipated shocks to output from the recorded output data.

\textsuperscript{35}We follow the procedure suggested by Wooldridge (2009) that forms moments on the joint error term ($\xi_{ft} + \epsilon_{ft}$).

\textsuperscript{36}This follows the standard practice in the literature where production functions are estimated at the industry level. For example, see Levinsohn and Petrin (2003).

\textsuperscript{37}The translog production function is $q_{ft} = \beta_{l}l_{ft} + \beta_{u}u_{ft} + \beta_{k}k_{ft} + \beta_{kk}k_{ft}^2 + \beta_{m}m_{ft} + \beta_{mm}m_{ft}^2 + \beta_{lk}l_{ft}k_{ft} + \beta_{lm}l_{ft}m_{ft} + \beta_{mk}m_{ft}k_{ft} + \beta_{lmk}l_{ft}m_{ft}k_{ft} + \omega_{ft}$. 

23
Given the aforementioned assumptions that productivity is firm-specific and log-additive and that inputs are divisible across products, we can rewrite the production function as: 
\[ \hat{q}_{fjt} = f(\tilde{x}_{ft}, \hat{\beta}, \hat{\omega}_{fjt}, \rho_{fjt}) + \omega_{ft} \]
and recover \( \{\rho_{fjt}\}_{j=1}^{J} \) using:
\[ \hat{q}_{fjt} - f(\tilde{x}_{ft}, \hat{\beta}, \hat{\omega}_{fjt}) = f_{2}(\tilde{x}_{ft}, \hat{\omega}_{fjt}, \rho_{fjt}) + \omega_{ft} \\
\sum_{j} \exp(\rho_{fjt}) = 1, \tag{23} \]
where \( f_{1} \) and \( f_{2} \) depend on the functional form of the production function and the input prices \( \hat{\omega}_{fjt} \) for each product \( j \) are computed based on the input price function (12). In other words, to recover the input allocations \( \rho_{fjt} \), we separate the production function into a component \( f_{1} \) that captures all terms that do not depend on \( \rho_{fjt} \) and a component \( f_{2} \) that collects all terms that involve \( \rho_{fjt} \). Because the input allocation shares have to sum up to 1 across all products in a multi-product firm, this yields a system of \( J_{ft} + 1 \) equations (where \( J_{ft} \) is the number of products produced by firm \( f \) at time \( t \)) in \( J_{ft} + 1 \) unknowns (the \( J_{ft} \) input allocations \( \rho_{fjt} \) and \( \omega_{ft} \)) for each firm-year pair.

Let \( \hat{\omega}_{fjt} = \hat{q}_{fjt} - f_{1}(\tilde{x}_{ft}, \hat{\beta}, w_{ft}) \). Applying our translog functional form to (23), we obtain:
\[ \hat{\omega}_{fjt} = \omega_{ft} + \hat{a}_{fjt}\rho_{fjt} + \hat{b}_{fjt}\rho_{fjt}^2 + \hat{c}_{fjt}\rho_{fjt}^3 \tag{25} \]
The terms \( \hat{a}_{fjt}, \hat{b}_{fjt}, \) and \( \hat{c}_{fjt} \) are functions of the estimated parameter vector \( \hat{\beta} \) and the estimated input price correction \( \hat{\omega}_{fjt} \).

For each year, we obtain the firm’s productivity and input allocations, the \( J + 1 \) unknowns (\( \omega_{ft}, \rho_{f1t}, \ldots, \rho_{fjt} \)), by solving a system of \( J + 1 \) equations:
\[
\begin{align*}
\hat{\omega}_{f1t} &= \omega_{ft} + \hat{a}_{f1t}\rho_{f1t} + \hat{b}_{f1t}\rho_{f1t}^2 + \hat{c}_{f1t}\rho_{f1t}^3 \\
\vdots \\
\hat{\omega}_{fJt} &= \omega_{ft} + \hat{a}_{fJt}\rho_{fJt} + \hat{b}_{fJt}\rho_{fJt}^2 + \hat{c}_{fJt}\rho_{fJt}^3 \\
\sum_{j=1}^{J} \exp(\rho_{fjt}) &= 1, \quad \exp(\rho_{fjt}) \leq 1 \quad \forall fjt \tag{29}
\end{align*}
\]
This system imposes the economic restriction that each input share can never exceed one and they must together sum up to one across products in a firm. We numerically solve this system for each firm in each year.

\[\begin{align*}
\hat{a}_{f1} &= \hat{\beta}_{k} + \hat{\beta}_{l} + 3\hat{\omega}_{f1t}^{2}\hat{\beta}_{lmk} + \hat{\omega}_{f1t}\left(\hat{\beta}_{kl} + 2\hat{\beta}_{tk} + \hat{\beta}_{tm} + \hat{k}_{ft}\hat{\beta}_{lmk} + \hat{m}_{ft}\hat{\beta}_{lmk} - 2\hat{\omega}_{fjt}\hat{\beta}_{lmk}\right) + \hat{\beta}_{m} + \hat{k}_{ft}\left(2\hat{\beta}_{kk} + \hat{\beta}_{lk} + \hat{\beta}_{mk} + 2\hat{\beta}_{mm}\right) \\
\hat{b}_{f1} &= \hat{\beta}_{kk} + \hat{\beta}_{lk} + \hat{\beta}_{lm} + \hat{\beta}_{mk}k_{ft} + \hat{\beta}_{lmk}\hat{m}_{ft} - 3\hat{\omega}_{fjt}\hat{\beta}_{lmk} + \hat{\beta}_{mk} + \hat{\beta}_{mm} \\
\hat{c}_{f1} &= \hat{\beta}_{lmk}
\end{align*}\]

\[\text{For the translog, these terms are}\]
3.4 Markups and Marginal Costs

We can now apply our framework to compute markups and marginal costs using the estimates of the production function coefficients ($\beta$) and the input allocations ($\rho$). We calculate the markup for each product-firm pair $f, j$ in each time period $t$ using:

$$\hat{\mu}_{fjt} = \hat{\theta}^M_{fjt} \frac{P_{fjt}Q_{fjt}}{\exp(\hat{\rho}_{fjt}) \tilde{X}^M_{ft}},$$

(30)

where $\hat{\theta}^M_{fjt} = \theta(\hat{\beta}, \hat{\tilde{x}}_{ft}, \hat{\tilde{w}}_{fj}, \hat{\rho}_{fjt})$ and $\tilde{X}^M_{ft}$ denotes the firm’s expenditure on materials.

The product-specific output elasticity for materials $\hat{\theta}^M_{fjt}$ is a function of the production function coefficients and the materials allocated to product $j$. Hence, it can be easily computed once the allocation of inputs across products has been recovered. Marginal costs $mc_{fjt}$ are then recovered by dividing price by the relevant markup according to equation (6).

Note that both markups and marginal costs are estimates since they depend on the estimated production function coefficients and the input cost allocation parameters, which are estimates themselves since they depend on the production function coefficients. Hence, the only source of uncertainty in our markup (and marginal cost) estimates comes from using estimated coefficients (the production function coefficients $\hat{\beta}$ and the input price correction coefficients $\hat{\delta}$). We account for the measurement error in these variables when we estimate the reduced form regressions in Section 4 by bootstrapping over the entire procedure. We execute the following steps in sequence: 1) estimate the production function, 2) recover the input allocations, 3) calculate markups (marginal costs), and 4) project markups and costs on trade policy variables. We then repeat this procedure 500 times, using bootstrapped (with replacement) samples that keep the sample size equal to the original sample size. This allows us to compute the bootstrapped standard error on the trade policy coefficients in Section 4.

3.5 Discussion

In addition to the conceptual assumptions discussed in Section 3.1, the actual implementation of the approach requires a set of assumptions to accommodate limitations of the data. Some of these limitations are specific to our data set (for example, we do not have information on physical labor units and wages, but only the wage bill) and may be of little general relevance. But other limitations are present in every firm-level data set and will need to be addressed by any study using such data. To our knowledge, no dataset reports the allocation of input expenditures across products in multi-product firms or contains the complete information on the firm-specific input prices (including firm-specific price of capital). The additional assumptions we impose are needed in order to deal with these features of the data. Apart from measurement issues, the assumptions we employ also address challenges that arise from product differentiation.

---

39 The expression for the materials output elasticity for product $j$ at time $t$ is: $\hat{\theta}^M_{fjt} = \hat{\beta}_m + 2\hat{\beta}_{mm}m_{fjt} + \hat{\beta}_{ml}l_{fjt} + \hat{\beta}_{mk}k_{fjt} + \hat{\beta}_{mkl}l_{fjt}k_{fjt}$. As before, to obtain the physical inputs, we rely on our estimates of the input prices $\hat{\tilde{w}}_{fjt}$ and the input allocation shares $\hat{\rho}_{fjt}$.
In this section we discuss these additional assumptions and our identification strategy. We start by discussing the way we deal with the unobserved input allocations in multi-product firms.

3.5.1 The Use of Single-Product Firms: Economies of Scope and Relationship to Cost Function Estimation

This subsection expands on the discussion of economies of scope in our setting and relates it to discussion of economies of scope in the cost function literature. Our approach does not rule out economies (or diseconomies) of scope, which may be important for multi-product firms. Panzar (1989) defines economies of scope in terms of cost. Baumol et al. (1983) speak of economies of scope in production if the cost function is sub-additive: 

\[ c_f(t(q_1, q_2)) \leq c_f(t(q_1)) + c_f(t(q_2)) \]

where \( c_f(t) \) is a firm’s cost curve. While our framework rules out differences in the production technology between single- and multi-product firms, it allows for economies of scope through cost synergies. The main sources of such economies are the possibility that multi-product firms have higher Hicks-neutral productivity than single-product firms (potentially because of differences in management practices or organizational structures) and the multiproduct firms’ ability to spread their fixed costs across multiple products.

An alternative way of explaining the assumptions underlying our approach is to express them in terms of the cost function rather than the production function. A multi-product firm faces the short-run cost function, written in a general form as:

\[
C(Q) = \Phi(\Omega)C(Q, W, \beta) + F(\iota(Q))
\]

where \( C \) denotes the total costs for a firm producing a vector of outputs \( Q \), \( \Phi(\Omega) \) denotes the impact of factor-neutral productivity on costs, \( W \) denotes a vector of input prices, \( F \) are the fixed costs (which would be zero in long-run cost function), and \( \iota(.) \) is an indicator that takes the value of 1 if a firm produces a particular product in the vector \( Q \) and is zero otherwise. The assumption we impose is that the function \( C(Q, W, \beta) \) is the same across single- and multi-product firms. However, costs between the two types of firms can still differ because of: 1) factor-neutral productivity differences reflected in \( \Phi(\Omega) \); 2) (in the short run) the amortization of fixed costs \( F \) across more products for multi-product firms. A third possibility is that factor prices \( W \) differ across the two types of firms because of pre-negotiated contracts; such differences are consistent with our assumptions regarding input prices as long as the contracts do not specify input prices as a function of input quantity.

We emphasize that we allow for economies of scope rather than assuming it. For example, our results could find no productivity differences between single- and multi-product firms, or find that multi-product firms are less productive implying diseconomies of scope. Likewise, finding economies of scope in the range of our data does not imply existence of economies of scope over any range of products produced by a firm; it is possible that economies of scope switch to diseconomies once a firm reaches a certain number of products. This paper does not attempt to provide a theory of multi-product firms. We simply point out that our approach does not \textit{a priori} rule out economies

26
or diseconomies of scope in the range of our data.

The representation of the cost function in (31) raises the natural question of why we do not exploit the duality between production and cost function and estimate a multi-product cost function. The main reason for focusing on the production function is that we do not have information on wages and the firm-specific user cost of capital, which are required to estimate a cost function. Furthermore, a multi-product cost function estimation would require additional identification assumptions in order to deal with the endogeneity of multiple product outputs on the right-hand side. Finally, even if one could come up with such identification assumptions, the product portfolios in our particular context are not stable. While Indian firms very rarely drop products, they often add products during this period (see Goldberg et al. (2010b)). These frequent additions require explicitly modeling a firm’s decision to add a particular product (in contrast, our approach requires us to model only the change from single- to multi-product status). Given these challenges, the approach to estimate production functions from single-product firms while accounting for the potential selection bias is an appealing alternative.

3.5.2 Control Function for Input Prices and Timing Assumptions

This subsection explains how the control function for input prices, the law of motion for productivity and the timing assumptions allow us to identify the coefficients. Recall that the identification strategy involves two control functions for the two unobservables: input prices and productivity:

\[ w_{ft} = w_t(p_{ft}, ms_{ft}, D_f, G_f, EXP_{ft}) \]  
\[ \omega_{ft} = g(\omega_{ft-1}, \tau_{it-1}^{\text{output}}, \tau_{it-1}^{\text{input}}, EXP_{ft-1}, SP_{ft}) + \xi_{ft}. \]  

While \( \omega_{ft} \) enters the production function (10) linearly, the input prices enter non-linearly as part of the term \( B(.) \). By substituting the input price control function into the expression for \( w \), we get equation (13).

First, note that we make use of the input price control function in the first stage of the estimation, when we purge the data from the noise \( \epsilon \). At this stage, we use materials as a proxy for productivity. Given that materials demand depends on input prices, it is important to control for the input prices using the control function specified above. However, the first stage has no implications for the identification of the production function coefficients; its sole purpose is to net out \( \epsilon \).

Next, consider the identification of the production function coefficients \( \beta \) and the coefficients associated with the input price correction term \( \delta \). These are identified off our timing assumptions. To review these assumptions, we assume that materials are a freely adjustable input and hence they will be correlated with contemporaneous productivity. Similarly, output prices will be correlated with current productivity. In contrast, capital and labor are dynamic inputs. Therefore, they will be uncorrelated with the productivity innovation \( \xi_{ft} \). We rely on these assumptions to form moment conditions.

There are two remaining identification issues that need to be discussed. First, as we noted earlier,
the term $B(.)$ will in general include input expenditures $\bar{x}_{ft}$. This raises the question of whether the production coefficients $\beta$ are identified. They are identified because the input expenditures $\bar{x}_{ft}$ enter the input price term $B(.)$ only through interaction with the input prices. It is because of the complexity of the translog that $\bar{x}_{ft}$ appear in $B(.)$ through interactions with input prices. In a Cobb-Douglas specification, the input expenditures do not appear in $B(.)$. In fact, under a constant returns to scale Cobb-Douglas production function the input correction term $B(.)$ simplifies to $w(.)$.40

The second question is how the coefficients on variables that enter both the law of motion for productivity and the input price control function are identified. One example of such a variable is the export dummy. The law of motion for productivity includes a dummy for exporting in $t - 1$, while it is also included in the input price control. The answer is that these coefficients are again identified off timing assumptions. We assume that productivity responds with a lag to changes in a firm’s environment, since it plausibly takes time for a firm to take the actions required to increase its efficiency (e.g., hiring better managers, adopting better management practices, changing organizational structure, importing new intermediate inputs, etc.). Accordingly, variables that may influence a firm’s productivity, such as tariffs or exporting, enter with a lag in the law of motion of productivity. In contrast, output and input prices respond immediately to changes in the economic environment. Accordingly, the variables included in the input price control function enter with their current values. As noted earlier, it is precisely because these variables enter with their current values that we face an identification problem; the current values will be correlated with $\xi_{ft}$ since by assumption they respond to contemporaneous shocks. It is this potential correlation that leads us to form moment conditions based on the lags, and not the current values, of the corresponding variables (the vector $Y_{ft}$ contains lagged output prices, lagged market shares, etc.).

4 Empirical Results

4.1 Output Elasticities, Marginal Costs and Markups

In this subsection, we present the output elasticities recovered from the production function estimation procedure. We describe how failing to correct for input price variation or account for the selection bias affects the parameters. Finally, we present and discuss our markup and marginal cost estimates.

The output elasticities are reported in Table 3.41 The nice feature of the translog is that unlike in a Cobb-Douglas production function, output elasticities can vary across firms (and across products within firms). We report both the average and standard deviation of the elasticities across sectors, and the final column reports the returns to scale. We note that a few sectors appear to have low returns to scale, but these are driven by outliers; Table 4 reports median output elasticities which

---

40 See Appendix B for details of the special case of Cobb-Douglas.

41 The output elasticity for capital and labor are defined analogous to the materials elasticity reported in Footnote 39.
are generally larger than the averages. Since the returns to scale vary across firms, it is possible for many firms in a sector to have increasing returns to scale, while the estimate of the industry-average returns to scale is close to one. At the firm level, 70 percent of the sample exhibits increasing returns to scale.

The left panel of Table 5 repeats the production function estimation without implementing the correction for the unobserved input price variation discussed in subsection 3.2.2. The uncorrected procedure yields nonsensical estimates of the production function. For example, the output elasticities and returns to scale are sometimes negative, very low or very high. These results are to be expected given that we estimate a quantity-based production function using deflated input expenditures, i.e., we relate physical output to input expenditures. It is clear that failing to account for input price variation yields distorted estimates. To understand the source of the distortion, consider the following concrete example from our data: in 1995, Ashnoor Textile Mills and Delight Handicrafts Palace sold 71,910 and 67,000,000 carpets, respectively. Ashnoor, however, had about three times higher input expenditures and three times higher revenues. It is easiest to understand the implications of this example for the estimates using a Cobb-Douglas specification. A quantity production function estimation that ignores input price variation would result in very large and negative output elasticities (more input expenditures result in lower quantity for Ashnoor). In the more general translog specification, it is impossible to sign this bias because there are three inputs which interact in complicated ways with each other and input prices, but it is clear that one needs to correct for input price variation across firms. By introducing the input price control, we are effectively comparing output quantities to input quantities, and the resulting output elasticities then look reasonable.

The importance of the input price correction is not apparent in the earlier literature, which traditionally estimates a Cobb-Douglas specification of the form: $q + p = \hat{x}\beta + \hat{\omega}$. This specification relates deflated sales to deflated expenditures and implies that $\hat{\omega} = \omega + p - w(\cdot)$. That is, the unobserved productivity measure includes both (unobserved) output price $p$ and (unobserved) input prices $w$. If one does not control for either output or input price variation (the typical practice in this literature until recently), there is no apparent problem as the two price biases tend to work in opposite directions.\(^{42}\) This leads to output elasticities that appear plausible without immediately calling for a correction.\(^{43}\) Of course, this does not mean that the two biases exactly cancel each other, so the final estimates will generally still be biased.

The right panel of Table 5 presents the mean output elasticities from an estimation of the production function that does not include the sample selection correction described in Section 3.2.3. The coefficients change slightly when the selection correction is not implemented. The stability of the coefficient estimates with and without selection correction for the unbalanced panel suggests that the use of the unbalanced panel of single-product firms (which includes firms that are always

\(^{42}\)See De Loecker and Goldberg (2014) for a detailed discussion of this issue.

\(^{43}\)In fact, when we estimate a firm-level revenue-based production function using the standard control function approach, we obtain production function coefficients that look similar to the previous literature (results available upon request).
single-product and firms that ultimately transition to a multi-product status) likely alleviates most of the concerns about the selection bias. This is consistent with the findings in Olley and Pakes (1996).

The markups are reported in Table 6. The mean and median markups are 2.24 and 1.18, respectively, but there is considerable variation across sectors and across products and firms within sectors. Some firms report markups below one for individual products, but multi-product firms maximize profits across products, so they may lose money on some products while being profitable on others. To get a better sense of the plausibility of our estimates, we aggregate the product-level markups to the firm level using the share of sales as weights. The firm-level markups are below one for only about 15 percent of the sample and the median firm-level markup is 1.43. In fact, we find a strong positive (and statistically significant) relationship between firm markups and reported accounting profits, measured as operating profits divided by total sales (results available upon request). Importantly, for our main results below, we rely on changes in markups over time by exploiting variation within firm-product pairs rather than variation in levels across firms.

The methodology provides measures of markups and marginal costs without a priori assumptions on the returns to scale. The estimates show that many firms are characterized by increasing returns to scale, so we expect to observe an inverse relationship between a product’s marginal cost and quantity produced. Accordingly, another way to assess the plausibility of the measures is to plot marginal costs against production quantities in Figure 1 (we de-mean each variable by product-year fixed effects in order to facilitate comparisons across firms). The figure shows indeed that marginal costs vary inversely with production quantities. The left panel of the figure shows that quantities and markups are positively related indicating that firms producing more output also enjoy higher markups (due to their lower marginal costs).

We also examine how markups and marginal costs vary across products within a firm. Our analysis here is guided by the recent literature on multi-product firms. Our correlations are remarkably consistent with the predictions of this literature, especially with those of Eckel and Neary (2010) and the multi-product firm extension of Melitz and Ottaviano (2008) developed by Mayer et al. (2014). A key assumption in these models is that multi-product firms each have a “core competency”. The core product has the lowest (within a firm) marginal cost. For the other products, marginal costs rise with a product’s distance from the core competency. Mayer et al. (2014) assume a linear demand system which implies that firms have non-constant markups across products. Furthermore, firms have their highest markups on their “core” products with markups declining as they move away from their main product. Figure 2 provides evidence supporting these implications. They plot the de-meaned markups and marginal costs against the sales share of the product within each firm (markups and marginal costs are de-meaned by product-year and firm-year fixed effects in order to make these variables comparable across products within firms). Marginal costs rise as a firm moves away from its core competency while the markups fall. In other words, the firm’s most profitable product (excluding any product-specific fixed costs) is its core product. Despite not imposing any assumptions on the market structure and demand system in our estimation, these correlations are
remarkably consistent with the predictions from the multi-product firm literature.

4.2 Pass-Through

Foreshadowing the results in the next subsection, we also find evidence of imperfect pass-through of costs on prices because of variable markups. This subsection explains how we estimate pass-through.

Consider the identity that decomposes the (log) price of a firm $f$ producing product $j$ into its two subcomponents: (log) marginal cost, $\ln mc_{fjt}$, and (log) markup, $\ln \mu_{fjt}$:

$$\ln P_{fjt} = \ln mc_{fjt} + \ln \mu_{fjt}$$  \hspace{1cm} (34)

This identity can also be written as:

$$\ln P_{fjt} = \ln \mu_{fj} + \ln mc_{fjt} + (\ln \mu_{fjt} - \ln \mu_{fj})$$  \hspace{1cm} (35)

where $\ln \mu_{fj}$ is the (time-invariant) average (log) markup for this particular firm-product pair and $(\ln \mu_{fjt} - \ln \mu_{fj})$ is the deviation of the markup from its average. If markups are constant, then the last term becomes zero. This is the case of complete pass-through: a proportional change in marginal cost is passed entirely to prices. If markups are variable, then marginal costs are correlated with the term in parenthesis and pass-through is incomplete. For example, if the price elasticity of demand is increasing in price, then an increase in marginal cost (which will tend to raise the price) will lead to an increase in the price elasticity of demand and a decrease in the markup. In this case, the marginal cost is negatively correlated with the (variable) markup and the pass-through of a marginal cost change onto price is below one. This correlation between marginal costs and markups is not an econometric issue since the equation above is an identity. Rather, it is a correlation dictated by economic theory: any model that implies variable markups will also imply a correlation between marginal cost and markup and result in incomplete pass-through.

To understand the implications of variable markups and incomplete pass-through in our setting, first consider the hypothetical case where marginal cost can be measured exactly. Suppose we run the following pass-through regression:

$$\ln P_{fjt} = a_{fj} + \zeta \ln mc_{fjt} + \varepsilon_{fjt}$$  \hspace{1cm} (36)

where $a_{fj}$ is a firm-product fixed effect. In this setup, the error term $\varepsilon_{fjt}$ has a structural interpretation. It reflects the deviation of the actual markup in period $t$ from the average (i.e., it corresponds to $(\ln \mu_{fjt} - \ln \mu_{fj})$).

If markups are constant, then we would expect to find that $\zeta = 1$ and $\varepsilon_{fjt} = 0$ (i.e., an exact fit). The firm-product fixed effect $a_{fj}$ would accurately measure the constant markup and the coefficient $\zeta$ would measure the pass-through of marginal cost to price which would be complete ($\zeta = 1$). The deviation of the actual markup from the average, $\varepsilon_{fjt}$, would be zero if markups were constant. Of course, in reality we would never get an exact fit of the regression line. But as long as $\varepsilon_{fjt}$ captures
random variation in price (due for example to recording errors) that is orthogonal to the marginal cost, we would estimate complete pass-through.

If markups are variable, then the error term $\varepsilon_{fjt}$ will be correlated with the marginal cost $\ln mc_{fjt}$.\textsuperscript{44} We again emphasize that this correlation is dictated by theory and not by econometrics. If the price elasticity facing the firm is increasing in price, then a marginal cost increase will lead to a price increase, which will raise the price elasticity and lower the markup. Hence, $\varepsilon_{fjt}$ and $\ln mc_{fjt}$ will be negatively correlated and the pass-through coefficient $\zeta$ will be below one. This is the case of incomplete pass-through.

When observing marginal cost, the coefficient $\zeta$ reflects markup variability and pass-through. There would be no need to instrument for marginal costs. In fact, instrumenting marginal costs is conceptually incorrect because the correlation between marginal costs and the structural error of the regression (i.e., the markup) is precisely what the coefficient $\zeta$ is supposed to capture. However, in our application (and almost every other empirical study), we only observe an estimate of marginal cost, $\ln \hat{mc}_{fjt} = \ln mc_{fjt} + \sigma_{fjt}$. The pass-through regression becomes

$$\ln P_{fjt} = a_{fj} + \zeta \ln \hat{mc}_{fjt} + (\varepsilon_{fjt} - \zeta \sigma_{fjt}) = a_{fj} + \zeta \ln \hat{mc}_{fjt} + u_{fjt}$$

(37)

Measurement error results in a downward bias in the pass-through coefficient $\zeta$ leading us to conclude, potentially erroneously, that pass-through is incomplete. We therefore require instruments to address measurement error in marginal costs. It is important to note that in this setting, instruments must be uncorrelated with the measurement error, $\sigma_{fjt}$. However, we do not require that they are uncorrelated with the part of the error term that reflects the deviation in markup, $\varepsilon_{fjt}$. Indeed, such a condition would be inconsistent with the exercise which is precisely to measure the correlation between marginal cost and markup, that is the correlation between $\hat{mc}_{fjt}$ and $\varepsilon_{fjt}$.

We instrument for marginal cost in equation (37) with input tariffs and lagged marginal cost. Both variables are certainly correlated with marginal cost. The former should be uncorrelated with the measurement error in our marginal cost estimate, but input tariffs do not vary at the firm level. The advantage of lagged marginal cost is that it varies at the firm-product-year level. Although lagged marginal costs contain measurement error, we have no reason to expect this measurement error to be serially correlated.

Table 7 presents the pass-through results from estimating (37).\textsuperscript{45} OLS results are reported in column 1, and the coefficient is 0.360. The second column instruments marginal costs with both lagged marginal cost and input tariffs, and consistent with measurement error, the IV estimate increases. The coefficient rises to 0.427 and is statistically significant. In case one is concerned about first-order serial correlation in measurement error, the third column uses input tariffs and two-period lagged marginal cost as the instruments, and the IV estimate is now 0.572. Thus, the results seem robust to the use of alternative instruments and consistently point to low pass-through.

\textsuperscript{44}Variable markups can be generated in many different ways through various combinations of market structure, firm behavior and demand function. See Goldberg and Hellerstein (2013) for a discussion.

\textsuperscript{45}As noted in Section 3.4, we report bootstrap standard errors.
This imperfect pass-through means that shocks to marginal costs, for example shocks from trade liberalization, do not lead to proportional changes in factory-gate prices because of changes in markups. We examine this markup adjustment in detail in the subsequent section.

4.3 Prices, Markups and Trade Liberalization

We now examine how prices, markups and marginal costs adjusted as India liberalized its economy. As discussed in Section 2, we restrict the analysis to 1989-1997 since tariff movements after this period appear correlated with industry characteristics.

We begin by plotting the distribution of raw prices in 1989 and 1997 in Figure 3. Here, we include only firm-product pairs that are present in both years, and we compare the prices over time by regressing them on firm-product pair fixed effects and year dummies and plotting the residuals. As before, we remove outliers in the bottom and top 3rd percentiles. This comparison of the same firm-product pairs over time exploits the same variation as our regression analysis below. The figure shows that the distribution of (real) prices did not change much between 1989 and 1997. This might at first be a surprising result given nature of India’s economic reforms during this period that were designed to reduce entry barriers and increase competition in the manufacturing sector. As a first pass, the figure suggests that prices did not move much despite the reforms.

Of course, the figure includes only firm-product pairs that are present at the beginning and end of the sample, and summarizes aggregate trends, thereby not controlling for sector-specific factors that could influence prices beyond the trade reforms. We use the entire sample and control for macroeconomic trends in the following specification:

\[ p_{fjt} = \lambda_{fj} + \lambda_{st} + \lambda_{1} \tau_{it}^{\text{output}} + \eta_{fjt}. \]  

We exploit variation in prices and output tariffs within a firm-product over time through the firm-product fixed effects (\( \lambda_{fj} \)) and control for macroeconomic fluctuations through sector-year fixed effects \( \lambda_{st}. \) Since the trade policy measure varies at the industry level, we cluster our standard errors at this level. We report the price regression with just year fixed effects in column 1 of Table 8. The coefficient on the output tariff is positive implying that a 10 percentage point decline is associated with a small 1.29 percent decline in prices. Between 1989 and 1997, output tariffs fall on average by 62 percentage points; this results in a precisely estimated average price decline of 8 percent (=62*0.129). This is a small effect of the trade reform on prices and it is consistent with the raw distributions plotted in Figure 3. The basic message remains the same if we control more flexibly for trends with sector-year fixed effects in column 2. The results imply that the average

---

46 One could try to capture the net impact of tariff reforms using the effective rate of protection measure proposed by Corden (1966). However, this measure is derived in a setting with perfect competition and an infinite export-demand and import-supply elasticities which imply perfect pass-through. As we show below, these assumptions are not satisfied in our setting, so that the concept of the “effective rate of protection” is not well defined in our case.

47 Recall from Section 2 that tariffs vary at a 4-digit level, while sector is defined as a 2-digit industry.

48 Our result is consistent with Topalova (2010) who finds that a 10 percentage point decline in output tariffs results in a 0.96 percent decline in wholesale prices in India during this period.
decline in output tariffs led to a 10 (=62*.161) percent relative drop in prices.

These results show that although the trade liberalization led to lower factory-gate prices, the decline is more modest than we would have expected given the magnitude of the tariff declines. Since earlier studies (Goldberg et al. (2010a), Topalova and Khandelwal (2011)) have emphasized the importance of declines in input tariffs in shaping firm performance, we separate the effects of output tariffs and input tariffs on prices. Output tariff liberalization reflects primarily an increase in competition, while the input tariff liberalization should provide access to lower cost (and more variety of) inputs. We run the analog of the regression in (38), but separately include input and output tariffs:

\[ p_{fjt} = \lambda_{fj} + \lambda_{st} + \lambda_1 \tau^{\text{output}}_{it} + \lambda_2 \tau^{\text{input}}_{it} + \eta_{fjt}. \]  

(39)

The results are shown in column 1 of Table 9. There are two interesting findings that are important for understanding how trade affects prices in this liberalization episode. First, there is a positive and statistically significant coefficient on output tariffs. This result is consistent with the common intuition that increases in competitive pressures through lower output tariffs will lead to price declines. The effect is traditionally attributed to reductions in markups and/or reductions in X-inefficiencies within the firm. The point estimates imply that a 10 percentage point decline in output tariffs results in a 1.49 percent decline in prices. On the other hand, the coefficient on input tariffs is small and noisy. Holding input tariffs fixed and reducing output tariffs, we would observe a precisely estimated decline in prices. Overall, average output tariffs and input tariffs fall by 62 and 24 percentage points, respectively, and using the point estimates in column 1, this implies that prices fall on average by 18.1 percent (a decline that is statistically significant).

We use the estimates of markups and costs to examine the mechanisms behind these moderate changes in factory-gate prices. We begin by plotting the distribution of markups and costs in Figure 4. Like Figure 3, this figure considers only firm-product pairs that appear in both 1989 and 1997. The figure indicates that between 1989 and 1997, the marginal cost distribution shifted left indicating an efficiency gain. However, this marginal cost decline is offset by a corresponding rightward shift in the markup distribution. Since (log) marginal costs and (log) markups exactly sum to (log) prices, the net effect results in little changes to prices. Hence, the raw data point towards imperfect pass-through of cost declines to prices. As before, these patterns are only suggestive and presented only for illustrative purposes, given that the figures do not condition on the policy and other changes that took place over this period.

We re-run specification (39) using marginal costs and markups as the dependent variables to formally analyze these relationships. Since prices decompose exactly to the sum of marginal costs and markups, the coefficients in columns 2 and 3 sum to their respective coefficients in column 1 in Table 9. We first focus on marginal costs regressions reported in column 2. The coefficient on outliers in the top and bottom 3rd percent of the markup distribution. We trim to ensure that the results are not driven by outliers. Nevertheless, the results are robust (e.g., magnitudes change slightly but statistical significance is unaffected) to alternative trims (e.g., the top and bottom 1st) and to not trimming at all (results are available upon request).
output tariffs is statistically insignificant, suggesting that marginal costs are insensitive to output tariff liberalization. However, the coefficient on input tariffs is both positive and large in magnitude. This is strong evidence that improved access to cheaper and more variety of imported inputs results in large cost declines. The final row of Table 9 reports the average effect on marginal costs using the average declines in input and output tariffs. On average, marginal costs fell 35.2 percent.

This magnitude of the marginal costs decline is sizable and would translate to larger prices declines if markups are constant. However, Figure 4 suggests that markups rose during this period, and in column 3 of Table 9, we directly examine how input and output tariffs affected markups. The coefficient on input tariffs is large and negative implying that input tariff liberalization resulted in higher markups. The results indicate that firms offset the beneficial cost reductions from improved access to imported inputs by raising markups. The overall effect, taking into account the average declines in input and output tariffs between 1989 and 1997, is that markups, on average, increased by 17.0 percent. This increase offsets about half of the average decline in marginal costs, and as a result, the overall effect of the trade reform on prices is moderated.\(^{50}\)

Although tempting, it is misleading to draw conclusions about the pro-competitive effects of the trade reform from the markup regressions in column 3 of Table 9. The reason is that one needs to control for impacts of output tariff liberalization on marginal costs in order to isolate pro-competitive effects. For example, if output tariffs affect costs through changes in X-inefficiencies, firms may adjust markups in response to these cost changes. The simultaneous effects that tariffs have on both costs and markups make it difficult to identify pro-competitive effects of the reform based on specification in column 3.

To isolate pro-competitive effects, we need to control for simultaneous shocks to marginal costs. We do this by re-running the markup regression but controlling flexibly for marginal costs. Conditioning on marginal costs, the output tariff coefficient isolates the direct pro-competitive effect of the trade liberalization on markups. We report the results in Table 10.\(^{51}\) Indeed, the coefficient on output tariffs in column 1 is positive and significant; this provides direct evidence that output tariff liberalization exerted pro-competitive effects on markups. The way to interpret the results in column 1 is to consider the markups on two products in different industries. Conditional on any (potentially differential) impact of the trade reforms on their respective costs, the product in the industry that experiences a 10 percentage point larger decline in output tariffs will have a 1.25 percent relative decline in markups.\(^{52}\) Column 2 instruments marginal costs to account for measurement error (see discussion in Section 4.2) with input tariffs and a second-order polynomial in lag marginal costs, and the coefficient declines but remains statistically significant. This analysis demonstrates

\(^{50}\) These results are robust to controlling for other contemporaneous policy changes in India, e.g., delicensing (results available upon request).

\(^{51}\) To control for marginal costs as flexibly as possible, we use a second-order polynomial for marginal costs and suppress these coefficients in Table 10. We find very similar results if we simply include marginal costs as the only control (results are available upon request).

\(^{52}\) In unreported results, we include input tariffs in the regression. As discussed earlier, input tariffs should affect markups only through the imperfect transmission of their impact on costs through improved access to imported inputs. Once we control for marginal costs, input tariffs should have no effect on markups and that is what we find.
that although India’s trade reform led to large cost reductions leading firms to respond by raising markups. Once we control for these cost effects, output tariff reductions do exert pro-competitive effects by putting downward pressure on markups.

The pro-competitive effects might differ across products. For example, output tariffs may exert more pressure on products with high markups prior to the reform. We explore this heterogeneity by creating a time-invariant indicator for firm-product pairs in the top decile of their industry’s markup distribution in the first year that a product-pair is observed in the data. We interact output tariffs with this indicator to allow for differential effects of output tariffs on markups for these high markup products. The results are reported in column 3 of Table 10. The table shows a very strong effect of output tariffs on these high markup products: a 10 percentage point decline in output tariffs leads to a 1.04 percent fall in markups for products initially below the 90th percentile in the markup distribution. For high markup products, the same policy reform results in an additional 4.40 percent decline in markups. In short, once we control for the incomplete pass-through of costs, output tariffs reduce markups and these reductions are substantially more pronounced on products with initially high markups. We observe similar patterns when we account for measurement error in marginal costs by instrumenting in column 4.

4.4 Interpretation of Results: Variable Markups and Incomplete Pass-through

Our results call for a nuanced evaluation of the effects of the Indian trade liberalization on markups. While we do find evidence that the tariff reductions have pro-competitive effects, especially at the right tail of the markup distribution, our results suggest that the most significant effect of the reforms is to reduce costs to producers. Due to variable markups, cost reductions are not passed through completely to consumers.

This last finding raises the question of why prices do not fully respond to cost reductions. Our results here relate to a voluminous literature on price rigidities and incomplete pass-through in macroeconomics and international macroeconomics. While this literature has focused primarily on exchange rate pass-through, its findings are equally relevant to tariff reductions given that exchange rate and tariff changes have similar effects on firm profits. Structural approaches within this literature explain incomplete pass-through through a combination of demand side and market structure assumptions. As discussed in Section 4.2, there is a large class of potential models (i.e., combinations of demand side and market structure assumptions) that can generate this phenomenon. Incomplete pass-through requires the demand elasticity perceived by the firm to be rising in price, so any model that delivers a demand elasticity increasing in price will also deliver incomplete pass-through. For example, this pattern can be generated in a setting with a linear consumer demand and monopolistic competition as in Melitz and Ottaviano (2008). Alternatively, one could assume CES preferences and Cournot (e.g., Atkeson and Burstein (2008)), or nested logit and Bertrand (e.g., Goldberg (1995) or Goldberg and Verboven (2005)); or random coefficients and Bertrand (e.g., Goldberg and Hellerstein (2013) or Nakamura and Zerom (2010)). Which assumptions are appropriate depends on the industry under investigation. Against this background, the advantage
of our approach is precisely the fact that it establishes the existence of incomplete pass-through and explores its implications for trade policy without committing to a particular structure. Such structure may be defensible in the context of Industrial Organization case studies which rely on a careful study of the industry under consideration and its institutional setting to inform their assumptions. But it is less defensible in the context of an analysis of the entire Indian manufacturing sector that includes many heterogeneous industries, each likely characterized by different demand and market conditions. Our study demonstrates that variable markups generate incomplete cost pass-through in many different sectors, but it cannot answer the question of which fundamentals in each case generate variable markups. To answer this last question, one would need to impose more structure along the lines of the aforementioned studies, yet doing so would undermine the fundamental rationale and advantage of our approach.

Our results suggest that the trade reforms benefited producers relatively more than consumers, at least in the short run. However, this does not necessarily imply that the reform lowered consumer welfare, especially in the long run. There is an active literature studying the relationship between competition, firm profitability and innovation (e.g., see Aghion et al. (2005)). In Goldberg et al. (2010a), we show that firms introduced many new products—accounting for about a quarter of output growth—during this period. If the cost reductions (and associated markup increases) induced by the trade reform spurred this product growth, the benefits to consumers are potentially substantially larger. We also observe a positive correlation between changes in firm markups and product introductions (results available upon request). This suggests that firms used the input tariff reductions and associated profit increases to finance the development of new products, implying potential long-term gains to consumers. A complete analysis of this mechanism and the impact on welfare lies beyond the scope of this current paper.

5 Conclusion

This paper examines the adjustment of prices, markups and marginal costs in response to trade liberalization. We take advantage of detailed price and quantity information to estimate markups from quantity-based production functions. Our approach does not require any assumptions on the market structure or demand curves that firms face. This feature of our approach is important in our context since we want to analyze how markups adjust to trade reforms without imposing ex ante restrictions on their behavior. An added advantage of our approach is that since we observe firm-level prices in the data, we can directly compute firms’ marginal costs once we have estimates of the markups.

Estimating quantity-based production functions for a broad range of differentiated products introduces new methodological issues that we must confront. We propose an identification strategy based on estimating production functions on single-product firms. The advantage of this approach

---

53 These findings are consistent with Peters (2012) who develops a model with imperfect competition that generates heterogeneous markups which determine innovation incentives.
is that we do not need to take a stand on how inputs are allocated across products within multi-product firms. We also demonstrate how to correct for a bias that arises when researchers do not observe input price variation across firms, an issue that becomes particularly important when estimating quantity-based production functions.

The large variation in markups suggests that trade models that assume constant markups may be missing an important channel when quantifying the gains from trade. Furthermore, our results highlight the importance of analyzing the effects of both output and input tariff liberalization. We observe large declines in marginal costs, particularly due to input tariff liberalization. However, prices do not fall by as much. This imperfect pass-through occurs because firms offset the cost declines by raising markups. Conditional on marginal costs, we find pro-competitive effects of output tariffs on markups. Our analysis is based on data representative of larger firms, so our results are representative of these larger firms. Our results suggest that trade liberalization can have large, yet nuanced effects, on marginal costs and markups. Understanding the welfare consequences of these results using models with variable markups is an important topic for future research.

Our results have broader implications for thinking about the trade and productivity across firms in developing countries. The methodology produces quantity-based productivity measures that can be compared with revenue-based productivity measures. Hsieh and Klenow (2009) discuss how these measures can inform us about distortions and the magnitude of misallocation within an economy. Importantly, our methodology can deliver quantity-based productivity measures purged of substantial variation in markups across firms, which potentially improves upon our understanding of the role of misallocation in generating productivity dispersion. We leave the analysis of the role of misallocation on the distribution of these performance measures for future research.

References


Arkolakis, C., A. Costinot, D. Donaldson, and A. Rodríguez-Clare (2012). The elusive pro-competitive effects of trade. *mimeo, Yale University.*


**Tables and Figures**

<table>
<thead>
<tr>
<th>Sector</th>
<th>Share of Sample</th>
<th>Output</th>
<th>All Firms</th>
<th>Single-Product Firms</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Food products and beverages</td>
<td></td>
<td>10%</td>
<td>300</td>
<td>137</td>
<td>135</td>
</tr>
<tr>
<td>17 Textiles, Apparel</td>
<td></td>
<td>9%</td>
<td>331</td>
<td>196</td>
<td>78</td>
</tr>
<tr>
<td>21 Paper and paper products</td>
<td></td>
<td>3%</td>
<td>76</td>
<td>59</td>
<td>32</td>
</tr>
<tr>
<td>24 Chemicals</td>
<td></td>
<td>25%</td>
<td>462</td>
<td>216</td>
<td>483</td>
</tr>
<tr>
<td>25 Rubber and Plastic</td>
<td></td>
<td>5%</td>
<td>149</td>
<td>102</td>
<td>83</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td></td>
<td>6%</td>
<td>119</td>
<td>88</td>
<td>60</td>
</tr>
<tr>
<td>27 Basic metals</td>
<td></td>
<td>17%</td>
<td>232</td>
<td>142</td>
<td>101</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td></td>
<td>2%</td>
<td>76</td>
<td>55</td>
<td>45</td>
</tr>
<tr>
<td>29 Machinery and equipment</td>
<td></td>
<td>7%</td>
<td>171</td>
<td>83</td>
<td>186</td>
</tr>
<tr>
<td>31 Electrical machinery, communications</td>
<td></td>
<td>6%</td>
<td>95</td>
<td>55</td>
<td>102</td>
</tr>
<tr>
<td>34 Motor vehicles, trailers</td>
<td></td>
<td>9%</td>
<td>62</td>
<td>40</td>
<td>95</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>100%</td>
<td>2,073</td>
<td>1,173</td>
<td>1,400</td>
</tr>
</tbody>
</table>

Notes: Table reports summary statistics for the sample. The first column reports the share of output by sector in 1995. Columns 2 and 3 report the number of firms and number of single-product firms manufacturing products in the sector in 1995. Column 4 reports the number of products over the full sample, 1989-2003.
Table 2: Example of Sector, Industry and Product Classifications

<table>
<thead>
<tr>
<th>NIC Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Basic Metal Industries (Sector s)</td>
</tr>
<tr>
<td>2710</td>
<td>Manufacture of Basic Iron &amp; Steel (Industry i)</td>
</tr>
<tr>
<td>130101010000</td>
<td>Pig iron</td>
</tr>
<tr>
<td>130101020000</td>
<td>Sponge iron</td>
</tr>
<tr>
<td>130101030000</td>
<td>Ferro alloys</td>
</tr>
<tr>
<td>130106004000</td>
<td>Welded steel tubular poles</td>
</tr>
<tr>
<td>130106004090</td>
<td>Steel tubular structural poles</td>
</tr>
<tr>
<td>130106005000</td>
<td>Tube &amp; pipe fittings</td>
</tr>
<tr>
<td>130106100000</td>
<td>Wires &amp; ropes of iron &amp; steel</td>
</tr>
<tr>
<td>130106100030</td>
<td>Stranded wire</td>
</tr>
<tr>
<td>2731</td>
<td>Casting of iron and steel (Industry i)</td>
</tr>
<tr>
<td>130106003000</td>
<td>Castings &amp; forgings</td>
</tr>
<tr>
<td>130106003010</td>
<td>Castings</td>
</tr>
<tr>
<td>130106003010</td>
<td>Steel castings</td>
</tr>
<tr>
<td>130106003012</td>
<td>Cast iron castings</td>
</tr>
<tr>
<td>130106003013</td>
<td>Maleable iron castings</td>
</tr>
<tr>
<td>130106003014</td>
<td>S.G. iron castings</td>
</tr>
<tr>
<td>130106003019</td>
<td>Castings, nec</td>
</tr>
</tbody>
</table>

Notes: This table is replicated from Goldberg et al. (2010b). For NIC 2710, there are a total of 111 products, but only a subset are listed in the table. For NIC 2731, all products are listed in the table.
Table 3: Average Output Elasticities, by Sector

<table>
<thead>
<tr>
<th>Sector</th>
<th>Observations in Production Function Estimation</th>
<th>Labor</th>
<th>Materials</th>
<th>Capital</th>
<th>Returns to Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>15 Food products and beverages</td>
<td>774</td>
<td>0.20</td>
<td>0.67</td>
<td>0.23</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.13]</td>
<td>[0.20]</td>
<td>[0.05]</td>
<td>[0.07]</td>
</tr>
<tr>
<td>17 Textiles, Apparel</td>
<td>1,574</td>
<td>0.13</td>
<td>0.77</td>
<td>0.14</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.06]</td>
<td>[0.09]</td>
<td>[0.08]</td>
<td>[0.14]</td>
</tr>
<tr>
<td>21 Paper and paper products</td>
<td>470</td>
<td>0.25</td>
<td>0.65</td>
<td>0.02</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.16]</td>
<td>[0.16]</td>
<td>[0.10]</td>
<td>[0.09]</td>
</tr>
<tr>
<td>24 Chemicals</td>
<td>1,552</td>
<td>0.22</td>
<td>0.73</td>
<td>0.17</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.07]</td>
<td>[0.05]</td>
<td>[0.07]</td>
<td>[0.07]</td>
</tr>
<tr>
<td>25 Rubber and Plastic</td>
<td>705</td>
<td>0.13</td>
<td>0.67</td>
<td>0.09</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.23]</td>
<td>[0.31]</td>
<td>[0.28]</td>
<td>[0.67]</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td>632</td>
<td>0.25</td>
<td>0.43</td>
<td>0.02</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.34]</td>
<td>[0.22]</td>
<td>[0.43]</td>
<td>[0.66]</td>
</tr>
<tr>
<td>27 Basic metals</td>
<td>947</td>
<td>0.11</td>
<td>0.70</td>
<td>0.06</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.20]</td>
<td>[0.18]</td>
<td>[0.20]</td>
<td>[0.41]</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td>392</td>
<td>0.05</td>
<td>0.50</td>
<td>-0.03</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.35]</td>
<td>[0.34]</td>
<td>[0.34]</td>
<td>[1.01]</td>
</tr>
<tr>
<td>29 Machinery and equipment</td>
<td>702</td>
<td>0.32</td>
<td>0.57</td>
<td>0.34</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.16]</td>
<td>[0.06]</td>
<td>[0.15]</td>
<td>[0.19]</td>
</tr>
<tr>
<td>31 Electrical machinery &amp; communications</td>
<td>761</td>
<td>0.08</td>
<td>0.64</td>
<td>0.02</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.17]</td>
<td>[0.24]</td>
<td>[0.15]</td>
<td>[0.47]</td>
</tr>
<tr>
<td>34 Motor vehicles, trailers</td>
<td>386</td>
<td>0.44</td>
<td>0.55</td>
<td>0.46</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.59]</td>
<td>[0.25]</td>
<td>[0.63]</td>
<td>[1.19]</td>
</tr>
</tbody>
</table>

Notes: Table reports the output elasticities from the production function. The first column reports the number of observations for each production function estimation. Columns 2-4 report the average estimated output elasticity with respect to each factor of production for the translog production function for all firms. Standard deviations of the output elasticities are reported in brackets. The 5th column reports the average returns to scale, which is the sum of the preceding three columns.
### Table 4: Median Output Elasticities, by Sector

<table>
<thead>
<tr>
<th>Sector</th>
<th>Labor (1)</th>
<th>Materials (2)</th>
<th>Capital (3)</th>
<th>Returns to Scale (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Food products and beverages</td>
<td>0.20</td>
<td>0.69</td>
<td>0.23</td>
<td>1.11</td>
</tr>
<tr>
<td>17 Textiles, Apparel</td>
<td>0.13</td>
<td>0.75</td>
<td>0.13</td>
<td>1.01</td>
</tr>
<tr>
<td>21 Paper and paper products</td>
<td>0.25</td>
<td>0.66</td>
<td>0.04</td>
<td>0.93</td>
</tr>
<tr>
<td>24 Chemicals</td>
<td>0.21</td>
<td>0.73</td>
<td>0.16</td>
<td>1.10</td>
</tr>
<tr>
<td>25 Rubber and Plastic</td>
<td>0.18</td>
<td>0.75</td>
<td>0.14</td>
<td>1.06</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td>0.29</td>
<td>0.46</td>
<td>0.14</td>
<td>0.84</td>
</tr>
<tr>
<td>27 Basic metals</td>
<td>0.14</td>
<td>0.72</td>
<td>0.09</td>
<td>0.97</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td>0.15</td>
<td>0.63</td>
<td>0.09</td>
<td>0.88</td>
</tr>
<tr>
<td>29 Machinery and equipment</td>
<td>0.32</td>
<td>0.57</td>
<td>0.33</td>
<td>1.18</td>
</tr>
<tr>
<td>31 Electrical machinery &amp; communications</td>
<td>0.12</td>
<td>0.69</td>
<td>0.06</td>
<td>0.93</td>
</tr>
<tr>
<td>34 Motor vehicles, trailers</td>
<td>0.36</td>
<td>0.50</td>
<td>0.39</td>
<td>1.24</td>
</tr>
</tbody>
</table>

Notes: Table reports the median output elasticities from the production function. Columns 2-4 report the median estimated output elasticity with respect to each factor of production for the translog production function for all firms. The 5th column reports the median returns to scale.

### Table 5: Output Elasticities, Input Price Variation and Sample Selection

<table>
<thead>
<tr>
<th>Sector</th>
<th>Labor without Correcting for Input Price Variation (1)</th>
<th>Materials without Correcting for Input Price Variation (2)</th>
<th>Capital without Correcting for Input Price Variation (3)</th>
<th>Returns to Scale without Correcting for Input Price Variation (4)</th>
<th>Labor without Correcting for Sample Selection (1)</th>
<th>Materials without Correcting for Sample Selection (2)</th>
<th>Capital without Correcting for Sample Selection (3)</th>
<th>Returns to Scale without Correcting for Sample Selection (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Food products and beverages</td>
<td>0.66</td>
<td>0.21</td>
<td>1.71</td>
<td>2.62</td>
<td>0.19</td>
<td>0.64</td>
<td>0.24</td>
<td>1.05</td>
</tr>
<tr>
<td>17 Textiles, Apparel</td>
<td>-0.02</td>
<td>0.58</td>
<td>-0.05</td>
<td>0.41</td>
<td>0.12</td>
<td>0.75</td>
<td>0.13</td>
<td>1.01</td>
</tr>
<tr>
<td>21 Paper and paper products</td>
<td>0.03</td>
<td>-0.09</td>
<td>-0.19</td>
<td>-0.48</td>
<td>0.27</td>
<td>0.62</td>
<td>0.03</td>
<td>0.92</td>
</tr>
<tr>
<td>24 Chemicals</td>
<td>0.71</td>
<td>0.67</td>
<td>-0.74</td>
<td>0.57</td>
<td>0.21</td>
<td>0.75</td>
<td>0.15</td>
<td>1.11</td>
</tr>
<tr>
<td>25 Rubber and Plastic</td>
<td>0.03</td>
<td>0.02</td>
<td>0.13</td>
<td>0.38</td>
<td>0.19</td>
<td>0.76</td>
<td>0.11</td>
<td>1.04</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td>0.27</td>
<td>0.30</td>
<td>0.92</td>
<td>1.43</td>
<td>0.24</td>
<td>0.48</td>
<td>0.18</td>
<td>0.90</td>
</tr>
<tr>
<td>27 Basic metal</td>
<td>-0.27</td>
<td>0.92</td>
<td>0.02</td>
<td>0.85</td>
<td>0.12</td>
<td>0.74</td>
<td>0.12</td>
<td>0.99</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td>-1.28</td>
<td>-0.67</td>
<td>2.18</td>
<td>0.26</td>
<td>0.16</td>
<td>0.65</td>
<td>0.08</td>
<td>0.90</td>
</tr>
<tr>
<td>29 Machinery and equipment</td>
<td>0.05</td>
<td>0.22</td>
<td>-0.21</td>
<td>0.29</td>
<td>0.29</td>
<td>0.56</td>
<td>0.34</td>
<td>1.15</td>
</tr>
<tr>
<td>31 Electrical machinery, communication</td>
<td>-1.49</td>
<td>-0.08</td>
<td>0.31</td>
<td>-0.26</td>
<td>0.08</td>
<td>0.76</td>
<td>0.05</td>
<td>0.97</td>
</tr>
<tr>
<td>34 Motor vehicles, trailers</td>
<td>0.02</td>
<td>-0.46</td>
<td>1.63</td>
<td>0.84</td>
<td>0.26</td>
<td>0.53</td>
<td>0.31</td>
<td>1.07</td>
</tr>
</tbody>
</table>

Notes: The left table reports the median output elasticities from production function estimations that do not account for input price variation. The right panel reports the median output elasticities from production function estimations that do not account for sample selection.
Table 6: Markups, by Sector

<table>
<thead>
<tr>
<th>Sector</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Food products and beverages</td>
<td>1.56</td>
<td>1.09</td>
</tr>
<tr>
<td>17 Textiles, Apparel</td>
<td>1.48</td>
<td>1.22</td>
</tr>
<tr>
<td>21 Paper and paper products</td>
<td>1.08</td>
<td>0.98</td>
</tr>
<tr>
<td>24 Chemicals</td>
<td>1.97</td>
<td>1.25</td>
</tr>
<tr>
<td>25 Rubber and Plastic</td>
<td>2.45</td>
<td>1.31</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td>3.22</td>
<td>1.51</td>
</tr>
<tr>
<td>27 Basic metals</td>
<td>2.57</td>
<td>1.11</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td>2.73</td>
<td>1.10</td>
</tr>
<tr>
<td>29 Machinery and equipment</td>
<td>1.93</td>
<td>1.00</td>
</tr>
<tr>
<td>31 Electrical machinery, communications</td>
<td>3.82</td>
<td>1.30</td>
</tr>
<tr>
<td>34 Motor vehicles, trailers</td>
<td>5.06</td>
<td>1.16</td>
</tr>
<tr>
<td>Average</td>
<td>2.24</td>
<td>1.18</td>
</tr>
</tbody>
</table>

Notes: Table displays the mean and median markup by sector for the sample 1989-2003. The table trims observations with markups that are above and below the 3rd and 97th percentiles within each sector.

Table 7: Pass-Through of Costs to Prices

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Marginal Cost</td>
<td>0.360***</td>
<td>0.427***</td>
<td>0.572**</td>
</tr>
<tr>
<td></td>
<td>0.042</td>
<td>0.084</td>
<td>0.251</td>
</tr>
<tr>
<td>Observations</td>
<td>21,122</td>
<td>15,887</td>
<td>12,232</td>
</tr>
<tr>
<td>Within R-squared</td>
<td>0.30</td>
<td>0.25</td>
<td>0.09</td>
</tr>
<tr>
<td>Firm-Product FEs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Instruments</td>
<td>-</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>First-Stage F-test</td>
<td>-</td>
<td>59</td>
<td>7</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is (log) price. Column 1 is an OLS regression on log marginal costs. Column 2 instruments marginal costs with input tariffs and lag marginal costs. Column 3 instruments marginal costs with input tariffs and two-period lag marginal costs. The regressions exclude outliers in the top and bottom 3rd percent of the markup distribution. All regressions include firm-product fixed effects. The regressions use data from 1989-1997. The standard errors are bootstrapped and are clustered at the firm level. Significance: * 10 percent, ** 5 percent, *** 1 percent.
### Table 8: Prices and Output Tariffs, Annual Regressions

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Output Tariff</strong></td>
<td>0.129 **</td>
<td>0.161 ***</td>
</tr>
<tr>
<td></td>
<td>0.058</td>
<td>0.055</td>
</tr>
<tr>
<td><strong>Within R-squared</strong></td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>21,122</td>
<td>21,122</td>
</tr>
<tr>
<td><strong>Firm-Product FEs</strong></td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td><strong>Year FEs</strong></td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td><strong>Sector-Year FEs</strong></td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overall Impact of Trade Liberalization</th>
<th>-8.0 **</th>
<th>-10.0 ***</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.6</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is a firm-product’s (log) price. Column 1 includes year fixed effects and Column 2 includes sector-year fixed effects. The regressions exclude outliers in the top and bottom 3rd percent of the markup distribution. All regressions include firm-product fixed effects and use data from 1989-1997. Standard errors are clustered at the industry level. The final row uses the average 62% decline in output tariffs from 1989-1997 to compute the mean and standard error of the impact of trade liberalization on prices. That is, for each column the mean impact is equal to the -0.62 * 100 * (coefficient on output tariffs). Significance: * 10 percent, ** 5 percent, *** 1 percent.
Table 9: Prices, Costs and Markups and Tariffs

<table>
<thead>
<tr>
<th></th>
<th>Log Prices&lt;subｆｉｔ&lt;/sub&gt;</th>
<th>Log Marginal Cost&lt;subｆｉｔ&lt;/sub&gt;</th>
<th>Log Markup&lt;subｆｉｔ&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Output Tariff&lt;subｆｉｔ&lt;/sub&gt;</td>
<td>0.149 ***</td>
<td>0.084</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>0.056</td>
<td>0.085</td>
<td>0.077</td>
</tr>
<tr>
<td>Input Tariff&lt;subｆｉｔ&lt;/sub&gt;</td>
<td>0.372</td>
<td>1.253 **</td>
<td>-0.881 *</td>
</tr>
<tr>
<td></td>
<td>0.303</td>
<td>0.523</td>
<td>0.489</td>
</tr>
<tr>
<td>Within R-squared</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Observations</td>
<td>21,122</td>
<td>21,122</td>
<td>21,122</td>
</tr>
<tr>
<td>Firm-Product FEs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Sector-Year FEs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Overall Impact of Trade Liberalization</td>
<td>-18.1 **</td>
<td>-35.2 ***</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td>7.3</td>
<td>12.8</td>
<td>11.8</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is noted in the columns. The sum of the coefficients from the markup and marginal costs regression equals their respective coefficient in the price regression. The regressions exclude outliers in the top and bottom 3rd percent of the markup distribution, and include firm-product fixed effects and sector-year fixed effects. The final row uses the average 62% and 24% declines in output and input tariffs from 1989-1997, respectively, to compute the mean and standard error of the impact of trade liberalization on each performance measure. That is, for each column the mean impact is equal to the -0.62*100*(coefficient on output tariff) + -0.24*100*(coefficient on input tariff). The regressions use data from 1989-1997. The table reports the bootstrapped standard errors that are clustered at the industry level. Significance: * 10 percent, ** 5 percent, *** 1 percent.
Table 10: Pro-Competitive Effects of Output Tariffs

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output Tariff(_{it})</td>
<td>0.125 **</td>
<td>0.127 **</td>
<td>0.104 **</td>
<td>0.117 *</td>
</tr>
<tr>
<td></td>
<td>0.049</td>
<td>0.061</td>
<td>0.051</td>
<td>0.061</td>
</tr>
<tr>
<td>Output Tariff(<em>{it}) x Top(</em>{fp})</td>
<td></td>
<td></td>
<td>0.440 ***</td>
<td>0.213</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.142</td>
<td>0.156</td>
</tr>
<tr>
<td>Within R-squared</td>
<td>0.58</td>
<td>0.57</td>
<td>0.58</td>
<td>0.57</td>
</tr>
<tr>
<td>Observations</td>
<td>21,122</td>
<td>15,887</td>
<td>21,122</td>
<td>15,887</td>
</tr>
<tr>
<td>2nd-Order Marginal Cost Polynomial</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Firm-Product FEs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Sector-Year FEs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Instruments</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>First-stage F-test</td>
<td>-</td>
<td>8.1</td>
<td>-</td>
<td>8.1</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is (log) markup. All regressions include firm-product fixed effects, sector-year fixed effects and a second-order polynomial of marginal costs (these coefficients are suppressed and available upon request). Columns 2 and 4 instrument the second-order polynomial of marginal costs with second-order polynomial of lag marginal costs and input tariffs. Columns 3 interacts output tariffs and the second-order marginal cost polynomial with an indicator if a firm-product observation was in the top 10 percent of its sector’s markup distribution when it first appears in the sample. The regressions exclude outliers in the top and bottom 3rd percent of the markup distribution. The table reports the bootstrapped standard errors that are clustered at the industry level. Significance: * 10 percent, ** 5 percent, *** 1 percent.
Figure 1: Marginal Costs and Quantities

Variables demeaned by product-year FEs. Markups, cost and quantity outliers are trimmed below and above 3rd and 97th percentiles.

Figure 2: Markups, Costs and Product Sales Share

Markups and marginal costs are demeaned by product-year and firm-year FEs. Markup and marginal cost outliers are trimmed below and above 3rd and 97th percentiles.
Appendix

A  A Formal Model of Input Price Variation

This appendix provides a formal economic model that rationalizes the use of a flexible polynomial in output price, market share and product dummies to control for input prices. The model is a more general version of the models considered in Kremer (1993) and Verhoogen (2008).

We proceed in the following steps. We first show that under the assumptions of the model, the
quality of every input is an increasing function of output quality. Next, we show that this implies that the price of every input will be an increasing function of output quality. In the final step, we show that output quality can be expressed as a flexible function of output price, market share and a set of product dummies. Having established a monotone relationship between input prices and output quality, this implies that the price of every input can also be expressed as a function of the above variables.

A.1 Production Function for Output Quality

In order to proceed, we must specify the production function for quality. Let \( v_j \) indicate quality of product \( j \) and \( \psi_i \) indicate the quality of input \( i \) used to produce product \( j \).\(^{54}\) The production function for output quality is given by:

\[
v_j = \prod_{i=1}^{n} [\psi_i]^{\kappa_i} \omega_j \quad \text{with} \quad \sum \kappa_i < 1 \quad (A.1)
\]

For example, with three inputs, the above production function takes the form:

\[
v_j = \psi^K \psi^K \psi^K \omega_j
\]

This function belongs to the class of ‘O-Ring’ production functions discussed in Kremer (1993) and Verhoogen (2008). The particular (multiplicative) functional form is not important; the important feature is that \( \frac{\partial v_j}{\partial \psi_i \partial \psi_k} > 0 \), \( \forall i, k \) and \( i \neq k \). This cross-derivative implies complementarity in the quality of inputs. A direct consequence is that higher output quality requires high quality of all inputs (e.g., high quality material inputs are used by high-skill workers operating high-end machinery).

In addition to the production function for quality, we assume that higher quality inputs are associated with higher input prices. Let \( \overline{W}_i \) denote the sectoral average of the price of input \( i \) (e.g., sectoral wage) and \( W_i(\psi_i) \) the price of a specific quality \( \psi \) of input \( i \). Then,

\[
W_i(\psi_i) - \overline{W}_i = z_i \psi_i \quad \text{and} \quad z_i > 0. \quad (A.2)
\]

The equation above says that in order to use higher quality inputs, a firm needs to pay higher input prices. There are many ways to justify this relationship. For example, if input markets are competitive but have vertical differentiation, firms must pay higher prices for higher quality inputs. So while high quality inputs are expensive, all firms pay the same input prices conditional on input quality.

A.2 Demand

We close the model by specifying the demand and firms’ behavioral assumptions.

\(^{54}\)Here, the subscript \( j \) denotes a particular product produced by a firm.
The indirect utility $V_{nj}$ that consumer $n$ derives from consuming one unit of product $j$ can be written in general form as:

$$V_{nj} = \theta_n v_j - \alpha p_j + \varepsilon_{nj}$$

(A.3)

where $p_j$ is output price, $\theta_n$ denotes the willingness to pay for quality and $\varepsilon_{nj}$ denotes an idiosyncratic preference shock. This specification is general and encompasses all demand models commonly used in the literature. In its most general formulation, the specification above corresponds to the random coefficients model. In models of pure vertical differentiation, the utility will be given by the above expression with $\varepsilon_{nj} = 0$. A simple logit sets $\theta_n = \theta = 1$ (i.e., no observable consumer heterogeneity) and $\varepsilon_{nj}$ is assumed to follow the extreme value distribution. In the nested logit, $\theta_n = \theta = 1$ and $\varepsilon_{nj}$ follows the generalized extreme value distribution. Following the Industrial Organization literature, it is convenient to define the mean utility $\delta_j$ of product $j$ as $\delta_j = v_j - \alpha p_j$. The output quality $v_j$ is typically modeled as a function of product characteristics.

We now show how to control for quality variation across firms using observable characteristics using the specification in (A.3). Berry (1994) shows that the actual market share of a product $(ms_j)$ is a function of product characteristics and output price:

$$ms_j = s_j(\delta, \sigma) = s_j(v, p, \vartheta)$$

(A.4)

where $\sigma$ denotes a vector of density parameters of consumer characteristics and $\vartheta$ denotes a parameter vector. While the exact functional form is determined by choice of a particular demand structure, the general insight is that market shares are a function of product characteristics (i.e., quality) and prices. Berry (1994) shows that equation (A.4) can be inverted to obtain the mean utilities $\delta$ as a function of the observed market shares and the density parameters to be estimated.\textsuperscript{55}

With the $\delta$'s in hand, quality is function of output price and the mean utility. This insight is exploited by Khandelwal (2010) who uses a nested logit model to express quality as a function of output price and conditional and unconditional market shares. In a simple logit model, quality is a function of only output prices and unconditional market shares. Here, we use a general formulation that specifies quality as a function of output price, a vector of (conditional and unconditional) market shares and a set of product dummies:

$$v_j = v(p_j, ms_j, D)$$

(A.5)

The product dummies are used in lieu of product characteristics (which are not available in our data) and can accommodate more general demand specifications such as the nested logit and random coefficients model.

\textsuperscript{55}In the random coefficients model, the $\delta$'s are solved numerically. In simpler models, one can solve for the parameters analytically.
A.3 The Firm’s Maximization Problem

Without loss of generality, we assume that firms use prices and quality as strategic variables to maximize profits. Conditional on exogenous (to the firm) input prices that are determined in competitive input markets, firms choose input qualities. These choices determine the output quality according to the quality production function in (A.1). Let \( mc_j \) denote the marginal cost of producing a product \( j \) of quality \( v_j \). The marginal cost can be written as a function of quantity produced \( q_j \), quality \( v_j \), a parameter vector \( \gamma \) and productivity \( \omega_j \), \( mc_j(q_j, v_j, \gamma, \omega_j) \).

The profit function for a firm producing product \( j \) is:

\[
\pi_j = N \cdot s_j[p - mc_j(q_j, v_j(\psi, \omega_j), \gamma, \omega_j)]
\]  
(A.6)

where \( N \) denotes the market size (number of potential consumers). Output quality \( v_j \) is now explicitly written as a function of a vector of input qualities \( \psi \) and productivity \( \omega_j \) using the production function for quality in (A.1).

The first order condition with respect to price is

\[
p_j = mc_j(q_j, v_j, \gamma, \omega_j) + \frac{s_j}{\partial s_j/\partial p_j}.
\]  
(A.7)

The term \( s_j/[\partial s_j/\partial p_j] \) represents the markup, and as shown in Berry (1994), p. 254) it equals \( \frac{1}{\alpha}[s_j/(\partial s_j/\partial \delta_j)] \).

The first order condition with respect to the quality of each input \( i, \psi_i \), is:

\[
(p_j - mc_j) \cdot \frac{\partial s_j}{\partial \psi_i} - s_j \frac{\partial mc_j}{\partial \psi_i} = 0
\]  
(A.8)

From the first order condition with respect to price, we have

\[
(p_j - mc_j) = \frac{s_j}{\partial s_j/\partial p_j} = \frac{1}{\alpha} \frac{s_j}{\partial s_j/\partial \delta_j}.
\]  
(A.9)

Substituting this latter expression for the markup into the first order condition for input quality, we obtain:

\[
s_j \frac{1}{\alpha} \frac{1}{(\partial s_j/\partial \delta_j)} \frac{\partial s_i}{\partial \psi_i} - s_j \frac{\partial mc_j}{\partial \psi_i} = 0
\]  
(A.10)

or

\[
\frac{1}{\alpha} \frac{1}{(\partial s_j/\partial \delta_j)} \left[ \frac{\partial s_j}{\partial v_j} \frac{\partial v_j}{\partial \psi_i} \right] = \frac{\partial mc_j}{\partial \psi_i}
\]  
(A.11)

From \( \delta_j = v_j - \alpha p_j \) follows that \( \frac{\partial s_j}{\partial v_j} = \frac{\partial s_j}{\partial \delta_j} \), and the above first order condition simplifies to:

\[
\frac{1}{\alpha} \frac{\partial v_j}{\partial \psi_i} = \frac{\partial mc_j}{\partial \psi_i}
\]  
(A.12)
Using the production function for quality to obtain the derivative $\frac{\partial v_j}{\partial \psi_i}$ and substituting into (A.12), we obtain

$$\psi_i = \frac{1}{\alpha} \kappa_i v_j \left[ 1/\frac{\partial mc_j}{\partial \psi_i} \right] \forall i$$  \hspace{1cm} (A.13)

This expression is similar to the one derived in Verhoogen (2008), but with two differences. First, as we have shown above, the above expression can be derived from a very general demand system and market structure. Second, we did not assume a Leontief production technology. The last feature of the model complicates the analysis slightly. With a Leontief production technology, the derivative $\frac{\partial mc_j}{\partial \psi_i}$ is constant, and it will be positive given the assumption that higher quality inputs demand higher prices. However, with more general production technologies, this derivative will itself depend on quality. We therefore need to show explicitly that $\psi_i$ is an increasing function of $v_j$. The latter can be established using the second order conditions associated with profit maximization:

$$\frac{1}{\alpha} \kappa_i \frac{\partial v_j}{\partial \psi_i} \frac{1}{\psi_i} - \frac{1}{\alpha} \kappa_i v_j \frac{1}{(\psi_i)^2} - \frac{\partial^2 mc_j}{\partial \psi_i^2} < 0$$  \hspace{1cm} (A.14)

$$\frac{1}{\alpha} \kappa_i^2 v_j \frac{1}{(\psi_i)^2} - \frac{1}{\alpha} \kappa_i v_j \frac{1}{(\psi_i)^2} - \frac{\partial^2 mc_j}{\partial \psi_i^2} < 0$$

Let us define function $F \equiv \psi_i \left( \frac{\partial mc_j}{\partial \psi_i} \right) - \frac{1}{\alpha} \kappa_i v_j$. From the implicit function theorem, $\frac{\partial \psi_i}{\partial v_j} = -\frac{F_j}{F_i}$ where

$$F_j = -\frac{1}{\alpha} \kappa_i < 0$$  \hspace{1cm} (A.15)

and by virtue of the second order condition,

$$F_i = \frac{\partial mc_j}{\partial \psi_i} + \psi_i \frac{\partial^2 mc_j}{\partial \psi_i^2} - \frac{1}{\alpha} \kappa_i^2 v_j \frac{1}{\psi_i} + \psi_i \frac{\partial^2 mc_j}{\partial \psi_i^2} - \frac{1}{\alpha} \kappa_i^2 v_j > 0$$  \hspace{1cm} (A.16)

It follows that $\frac{\partial \psi_i}{\partial v_j} = -\frac{F_j}{F_i} > 0$. That is, input quality is an increasing function of output quality for every input.

Given the assumption that higher input quality demands a higher input price, it immediately follows that input prices will also be an increasing function of output quality for all inputs. From equation (A.2):

$$W_i(\psi_i) = W_i + z_i \psi_i = W_i + z_i \frac{1}{\alpha} \kappa_i v_j \left[ 1/\frac{\partial mc_j}{\partial \psi_i} \right]$$

In light of the above discussion, each input price facing a particular firm can be expressed as a function of the firm’s output quality, $W_i = g(v_j)$. Moreover, given that output quality is a function of output price, market share and product dummies, we have: $W_i = w(p_j, ms_j, D)$. 

55
B Estimation Procedure under a Special Case: Cobb-Douglas Production Function

We present our estimation procedure under the predominantly used production function specification in applied work: the Cobb-Douglas (CD) production function. While restrictive on the input-substitution patterns and the output elasticities, it greatly simplifies the estimation routine and the recovery of the input allocation terms ($\rho$). In addition, it helps to highlight the fundamental identification forces as the input price correction term does not include (interactions of) deflated expenditures.

We follow the structure of the main text (Section 3) and impose the CD functional form:

$$f(x_{fjt}) = \beta_l l_{fjt} + \beta_m m_{fjt} + \beta_k k_{fjt}.$$ \hspace{1cm} (B.1)

Following the same steps as in the main text we get the following estimating equation for the single-product firms corresponding to equation (10). We omit the product subscript $j$ given that the firms used in the estimation produce a single product:

$$q_{ft} = \beta_l \tilde{l}_{ft} + \beta_m \tilde{m}_{ft} + \beta_k \tilde{k}_{ft} - \Gamma w_{ft} + \omega_{ft} + \epsilon_{ft},$$ \hspace{1cm} (B.2)

where $\Gamma w(.)$ is a special case of the function $B(.)$ in the main text, $\Gamma = \beta_l + \beta_m + \beta_k$ is the returns to scale parameter, and as before $w_{ft} = \tilde{x}_{ft} - x_{ft}$ $\forall x = \{l, m, k\}$.

After running the first stage

$$q_{ft} = \phi_t(\tilde{x}_{ft}, z_{ft}) + \epsilon_{ft},$$ \hspace{1cm} (B.3)

with $\tilde{x}_{ft} = \{\tilde{l}_{ft}, \tilde{m}_{ft}, \tilde{k}_{ft}\}$, we have an estimate of predicted output $(\hat{\phi}_{ft})$. It is then immediate that the input price correction term $B(.)$ enters in equation (20) in a separate and additive fashion:

$$\omega_{ft}(\beta, \delta) = \hat{\phi}_{ft} - \beta_l \tilde{l}_{ft} - \beta_m \tilde{m}_{ft} - \beta_k \tilde{k}_{ft} - \Gamma w(p_{ft}, ms_{ft}, D, G_{ft}; \delta),$$ \hspace{1cm} (B.4)

where $-\Gamma w(.)$ is a special case of the function $B(.)$ in the main text. If one assumes a vertical differentiation model of demand, then the input price control function $w(.)$ will take only output price as its argument, and the last term in (B.4) becomes $\Gamma w(p_{ft}, \delta)$. We form moments on $\xi_{ft}(\beta, \delta)$ by exploiting the same law of motion of productivity in equation (18), and the same timing assumptions as in the main text.

To estimate markups and marginal costs we need the input allocation terms $\rho_{fjt}$. In the case of the CD, their derivation is simplified to solving the system of equations given by:

$$\omega_{ft} + \Gamma \rho_{fjt} \hat{w}_{fjt} = \hat{\phi}_{fjt} - \beta_l \tilde{l}_{fjt} - \beta_m \tilde{m}_{fjt} - \beta_k \tilde{k}_{fjt}$$ \hspace{1cm} (B.5)

where $\hat{w}_{fjt}$ is the input price term that we compute based on the estimated function $w(.)$ and $\Gamma$ is defined as above. Taking into account that $\sum_j \exp(\rho_{fjt}) = 1$, this results in a system of $J_{ft} + 1$
equations (one for each product $j$ produced by firm $f$ at time $t$, plus the summing up constraint for the input allocations) in $J_{ft} + 1$ unknowns (the $J_{ft}$ input allocations for each firm-year pair and firm productivity) and we can solve for $\rho_{fjt}$ and $\omega_{ft}$.

We now have all we need to compute markups and marginal costs. The major difference is that $\theta_{fjt}^M = \beta_m$, so that all the variation in markups (and marginal costs) comes from the expenditure share $\alpha_{fjt}$.

C Data Appendix

We use the Prowess data, compiled by the Centre for Monitoring the Indian Economy (CMIE), that spans the period from 1989 to 2003. In addition to standard firm-level variables, the data include annual sales and quantity information on firms’ product mix. Although Prowess uses an internal product classification that is based on the Harmonized System (HS) and National Industry Classification (NIC) schedules, our version of Prowess did not explicitly link the product names reported by the firms to this classification. We hired two research assistants, working independently, to map the codes to the product names reported by firms. The research assistants assigned product codes with identical NIC codes in 80% of the cases, representing 91% of output. A third research assistant resolved the differences between the mappings done by the first two research assistants by again manually checking the classifications.

To estimate the production function, we need firm-level labor, capital and materials. Prowess does not have reliable employment information, so we use the total wage bill (which includes bonuses and contributions to employees’ provident funds) as our measure for labor. Materials are defined as the consumption of commodities by an enterprise in the process of manufacturing or transformation into product. It includes raw material expenses and consumption of stores and spares. Capital is measured by gross fixed assets, which includes movable and immovable assets. These variables are deflated by two-digit NIC wholesale price indexes.

We match the firm variables to tariff data. The tariff data are reported at the six-digit HS level and were compiled by Topalova (2010). We pass the tariff data through India’s input-output matrix for 1993-94 to construct input tariffs. We concord the tariffs to India’s NIC schedule developed by Debrov and Santhanam (1993). Formally, input tariffs are defined as $\tau_{it}^\text{input} = \sum_k a_{ki} \tau_{kt}^\text{output}$, where $\tau_{kt}^\text{output}$ is the tariff on industry $k$ at time $t$, and $a_{ki}$ is the share of industry $k$ in the value of industry $i$.

D Markups and Monopsony Power

If firms have monopsony power, this would alter the first order conditions in Section 3.1 (equations 3-5). We briefly discuss under which conditions our main results, relating markups to tariff changes, are not affected.

Consider a firm that produces just one product, and suppose production requires just one flexible
input \( V_{ft}^v \). The Lagrangian in this case would be:

\[
L = W_{ft}^v V_{ft}^v + \lambda_{ft} \left( Q_{ft} - Q_{ft} \left( V_{ft}^v, \omega_{ft} \right) \right).
\]  

(D.1)

Taking first order conditions and allowing for monopsony power gives:

\[
\frac{\partial L}{\partial V_{ft}^v} = W_{ft}^v + \frac{\partial W_{ft}^v}{\partial V_{ft}^v} V_{ft}^v - \lambda_{ft} \frac{\partial Q(.)}{\partial V_{ft}^v} = 0.
\]  

(D.2)

If a firm has no monopsony power, \( \frac{\partial W_{ft}^v}{\partial V_{ft}^v} = 0 \). For firms with monopsony power, \( \frac{\partial W_{ft}^v}{\partial V_{ft}^v} < 0 \): the more the firm buys, the lower the price of the input. We can rearrange the FOC as:

\[
W_{ft}^v + \frac{\partial W_{ft}^v}{\partial V_{ft}^v} V_{ft}^v = \lambda_{ft} \frac{\partial Q(.)}{\partial V_{ft}^v}.
\]  

(D.3)

The Lagrange multiplier remains: \( \lambda_{ft} = P_{ft}/\mu_{ft} \), we get

\[
\mu_{ft} \left( W_{ft}^v + \frac{\partial W_{ft}^v}{\partial V_{ft}^v} V_{ft}^v \right) = P_{ft} \frac{\partial Q(.)}{\partial V_{ft}^v}.
\]  

(D.4)

If we now compare a firm with and without monopsony power, ceteris paribus, the markup for the firm with monopsony power will be larger. This implies that we may be under-estimating the markup by ignoring potential monopsony power.

However, even if our estimates of the markup levels were biased due to the existence of monopsony power, it is still unlikely that our conclusions regarding the effects of tariffs on markups and costs would be affected. To see this, note that the above expression can be simplified to

\[
\mu_{ft} = (\theta_{ft} + 1)/(1 + v_{ft}).
\]  

(D.5)

\(^{56}\)Dividing through by \( W^v \), and dividing and multiplying the right-hand-side by \( (V^v/Q) \), and rearranging terms.

\(^{56}\)Dividing through by \( W^v \), and dividing and multiplying the right-hand-side by \( (V^v/Q) \), and rearranging terms.
business group.\textsuperscript{57} This leads us to believe that monopsony power is not a first order concern in our setting.

\textsuperscript{57}Results are available upon request.
Firm Performance in a Global Market

Jan De Loecker\textsuperscript{1,3,4} and Pinelopi Koujianou Goldberg\textsuperscript{2,4}

\textsuperscript{1}Department of Economics, Princeton University, Princeton, New Jersey 08544; email: jdeloeck@princeton.edu
\textsuperscript{2}Department of Economics, Yale University, New Haven, Connecticut 06520
\textsuperscript{3}Centre for Economic and Policy Research, London EC1V 3PZ, United Kingdom
\textsuperscript{4}National Bureau of Economic Research, Cambridge, Massachusetts 02138

Abstract

In this article, we introduce an empirical framework to analyze how firm performance is affected by increased globalization. Using this framework, we discuss recent work on measuring the impact of various shocks firms face in the global marketplace, such as reductions in trade costs (through lowering tariffs and abolishing quotas). Our analytical framework nests most empirical approaches to estimating the impact of trade and industrial policies on firms active in international markets. We identify outstanding issues surrounding the identification of the underlying mechanisms and conclude with suggestions for future research.
1. INTRODUCTION

Few topics have been researched as extensively as the relationship between trade openness and firm, or industry, performance. A Google Scholar search on the keywords “trade and productivity” returns 1,920,000 papers! The purpose of this article is twofold: First, we develop a general framework for discussing the large empirical literature on the topic. Second, we summarize the main insights from work to date. Although much of the empirical research we draw upon rests on theoretical models, we do not review theoretical work directly, except to the extent that this theoretical work informs empirical specifications and helps identify underlying mechanisms. Furthermore, this article is not intended to be a comprehensive literature review. Instead, we discuss a selected set of papers from the perspective of our empirical framework that nests the productivity and pass-through literatures spanning industrial organization, international trade, and international macro.

Our point of departure is the voluminous empirical literature on the effects of trade on firm productivity. Economists have always postulated that one of the main benefits of opening up markets to foreign competition is to make firms more efficient and have proceeded to estimate the effects of trade liberalizations, or globalization more generally, on efficiency. Early empirical attempts predate theoretical work in this area and for the most part lack a sound theoretical justification for why one would expect efficiency to increase with exposure to foreign markets. With the development of models of firm heterogeneity, however, this literature has experienced a renaissance. There are two main strands. The first one investigates the performance of exporters and relies primarily on cross-sectional comparisons to assess whether exporters are more efficient than firms that sell to domestic markets only. The second strand focuses on trade liberalization episodes and investigates whether industry productivity increases in the postreform period, in which the productivity gains can arise either because of within-firm improvements or because of reallocation of economic activity toward more efficient firms.

Although this literature has established some interesting patterns, we show that with few exceptions, it has been loose in its use of the term productivity. What it actually delivers is a measure of firm performance or profitability. The distinction between productivity and profitability is important; the latter depends not only on physical efficiency, but also on prices, which reflect product differentiation and markups in addition to costs. The framework we develop allows one to explicitly trace these components. This decomposition offers the advantage that one can link improvements in firm performance to specific mechanisms through which globalization affects firms. Understanding these mechanisms is important for assessing the welfare and distributional effects of trade openness; for example, a trade liberalization that improves firm performance by inducing improvements in physical efficiency has different implications than a liberalization that makes firms better off by increasing their profits.

The realization that measured firm performance captures markups as well as physical efficiency naturally leads to two other literatures that were developed in different contexts: the large industrial organization literature on imperfect competition and the international literature on incomplete (exchange rate) pass-through. The first explicitly investigates the measurement and determinants of markups (e.g., the role of market structure, product differentiation, and demand elasticities); the second focuses on how a certain type of cost shock (i.e., exchange rate changes) is passed through to prices. Firms competing in global markets are faced with a different market (and likely also demand) structure and may as a result change their prices and markups or their product characteristics. A tariff reduction, for example, intensifies the import competition domestic firms face, and we would expect them to reduce their prices and markups in response. This is the usual procompetitive effect one associates with trade liberalizations. At the same time, a reduction in
tariffs allows firms to purchase imported intermediates that they use as inputs in their production at lower prices; these reductions in input prices are tantamount to a cost shock facing the firm, similar to the exchange rate changes considered in the international literature. Whether this cost shock will be passed through to prices depends on many factors that reflect demand and market structure conditions.

Our review concludes that there is one robust finding that emerges from this literature: Globalization improves industry performance. However, there is less consensus on how these improvements arise. Many studies find that improvements are generated through the reallocation of market shares toward better-performing firms, whereas others document significant within-firm improvements. Several papers find that the effects of input tariff liberalization dominate those of output tariff liberalization in some developing countries. Perhaps more importantly, the majority of studies do not distinguish between physical efficiency and price/markup effects. Hence, it is not clear that the aforementioned improvements represent true productivity gains as opposed to increases in market power. A couple of recent papers emphasizing this distinction show that the effects on markups and prices are significant. However, it is too early to know whether the results from these studies generalize to other contexts. This is an exciting area for future research, and we hope that the current article helps guide future endeavors in this direction.

The remainder of the article is organized as follows. Section 2 introduces a general framework that can be used to discuss existing work and clarify what economic theory suggests should be measured, and what is measured in practice. We use this framework to decompose firm performance into its components. In Section 3, we discuss the particular mechanisms through which trade openness is expected to affect each of these components. Section 4 briefly summarizes the insights obtained from work to date, and Section 5 concludes.

2. A FRAMEWORK FOR MEASURING PERFORMANCE

Performance at the firm level is measured in many different ways. Such ways include accounting measures of profitability, the Lerner index, sales per input, and total factor productivity. Although correlated, the various measures capture different aspects of firm performance, and exposure to a global market is not expected to affect these aspects in the same way. In this section, we first use a simple regression framework to summarize the large empirical literature on the topic. Next we introduce an empirical framework, based on a production function, that allows us to reinterpret the commonly used measures of performance and decompose them into their underlying determinants: demand and cost primitives, as well as the market structure of the industry under study.

2.1. Measuring Performance

We characterize a large literature—spanning industrial organization, international economics, and international macroeconomics—by considering firm performance \( (\pi_{it}) \) as the residual in a regression of sales \( (s_{it}) \) on input expenditures \( (e_{it}) \).\(^1\) Applied researchers typically consider a log-linear relationship between sales and expenditures:

\[
s_{it} = e_{it}' \beta + \pi_{it}, \tag{1}
\]

\(^1\)The vector \( e_{it} \) typically includes expenditures on labor, intermediate inputs (materials), and capital. Unless noted otherwise, lowercase letters denote the log of the corresponding variable throughout this article.
where $\mathbf{b}$ is the vector of coefficients. The data we have in mind track firms indexed by $i$ for which we observe sales and input use, both expressed in monetary terms (i.e., sales and expenditures), over time $t$. The residual $\pi_{it}$ as a measure of performance is closely related to profits because we subtract expenditures from sales.

The fundamental research question economists have been trying to answer over the past few decades is: How do changes in international competition affect the performance of firms ($\pi_{it}$) and industries and ultimately the overall welfare of countries or regions? Trade liberalization episodes, preferably exogenous to firms in an industry, are the natural place to turn to analyze this question empirically by relating performance measures to changes in trade protection, such as tariff declines, lifting of quota restrictions, or removal of antidumping duties.

Equation 1 is a point of departure for the literature that typically utilizes firm- or plant-level data across many different sectors of one or more economies. Such data tend to be readily available for a large set of countries and time periods. In contrast, the case study approach commonly employed in modern industrial organization makes use of much more detailed information on product-level prices and product- or firm-specific cost variables to compute performance measures, at the cost of the approach being feasible only for the specific industries and countries for which this information is available. Another approach would be to use accounting data to measure performance as operating profits. However, this article is concerned with recovering measures of firm performance that accurately reflect economic costs and benefits, and it is well known that accounting profits provide poor measures of economic profits (see Schmalensee 1989 for a detailed discussion of the problems of using accounting profits as a measure of firm performance).

With few exceptions, the existing literature has viewed Equation 1 as the empirical analog of a production function and interpreted the residual $\pi_{it}$ as a measure of total factor productivity. As we argue below, under a set of restrictive assumptions, Equation 1 is indeed a production function, with $\mathbf{b}$ the estimated production function coefficients. One possible set of such assumptions features firms that produce a homogeneous product and are active in an industry characterized by perfectly competitive output and input markets. In this case, the relationship between sales and expenditures is equivalent to the relationship between physical output and inputs, and consequently, $\pi_{it}$ is an estimate of (physical) productivity. However, in the general case, the performance residual $\pi_{it}$ will capture much more than productivity.

When we try to understand how increased international competition impacts firm performance, recovering the residual of Equation 1 is only the first step; to identify the underlying mechanism(s), we need to understand the underlying components of $\pi_{it}$. To this end, we need to be more precise about what $\pi_{it}$ actually measures. We turn to this issue in the next section and use a production function approach to interpret the performance residual ($\pi_{it}$).

### 2.2. Firm Performance: A View from the Production Side

Using a basic production function\(^2\) that relates quantities produced ($q_{it}$) to a vector of physical input usage ($\mathbf{x}_{it}$) allows us to express sales in a (perhaps) more familiar fashion as

$$
s_{it} = \mathbf{x}_{it} \mathbf{\alpha} + \omega_{it} + p_{it} = \mathbf{z}_{it} \mathbf{\alpha} + \omega_{it} + p_{it} - \mathbf{z}_{it} \mathbf{\alpha},
$$

where

\(^2\)To simplify notation, we base our discussion on a Cobb-Douglas production function, but our framework generalizes to any other functional form.
where we rely on the definition of sales, \( s_{it} = q_{it} + p_{it} \); a standard Hicks-neutral production function, \( q_{it} = x_{it} \alpha + \omega_{it} \), with \( \alpha \) and \( \omega_{it} \) the vector of production function coefficients and productivity, respectively\(^3\); and the definition of input expenditures, \( e_{it} = x_{it} + z_{it} \), with \( z_{it} \) the vector of input prices.

This simple framework allows us to immediately connect our structural equation (Equation 2) to the profitability residual in Equation 1 by grouping the last three terms in the equation above to \( \pi_{it} = \omega_{it} + p_{it} - z_{it} \alpha \). Hence, we can rewrite Equation 2 as an equation relating sales to expenditures:

\[
 s_{it} = e_{it} \alpha + \pi_{it}. 
\]

Equations 1 and 3 are almost identical up to the coefficients \( \beta \) and \( \alpha \). We distinguish between the structural coefficients of the production function, \( \alpha \), and the coefficients obtained after estimating an equation such as Equation 1. It is only under a set of very restrictive assumptions that these two vectors will be identical.

Consider Equation 1 and assume that the researcher is willing to make the economic assumption that there is neither output nor input price variation across firms.\(^4\) Even in this case, ordinary least squares (OLS) estimation of Equation 1 on a panel data set of firms is not expected to yield the vector of coefficients (\( \alpha \)) due to the well-known simultaneity and selection biases. Both biases arise from the likely correlation between inputs and unobserved productivity (\( \omega_{it} \)). This is a well-known problem when estimating production functions and has been addressed in the literature; we refer the reader to a recent overview by Ackerberg et al. (2007) for a detailed discussion. For the remainder of this article, we assume that these biases can be appropriately dealt with when estimating the underlying structural parameters. Of course, the treatment of the unobserved productivity shock is not independent of the unobserved output and input price problem. In fact, the framework we put forward treats both unobservables, price and productivity, jointly in one consistent framework.\(^5\) However, our point is that if output and input prices do not vary across firms, it is possible in principle to recover both the production function coefficients \( \alpha \) and firm productivity \( \omega_{it} \) by applying the techniques suggested in the recent literature. In practice, both price and productivity variation will most likely play an important role, and as a consequence, ignoring (unobserved) prices in the estimation of production functions will lead to biased results, and vice versa.

Consider Equation 1 once more, but let us now allow for both output and input price variation. In this case, the structural error \( \pi_{it} \) contains two more components in addition to productivity: the output price \( (p_{it}) \) and the vector of input prices \( (z_{it}) \). Relying on sales and expenditure data will clearly not deliver an estimate of productivity, nor will it deliver the vector of production function coefficients. In this case, \( \beta \) is a vector of coefficients describing the mapping from expenditures to sales.

To see why estimation of Equation 1 will lead to biased coefficients in the presence of output and/or input price variation, let us explicitly introduce deflators and rewrite Equation 2 to reflect

\[^3\text{For the well-known Cobb-Douglas production function in three inputs \([\text{labor} (l), \text{materials} (m), \text{and capital} (k)]\), we get} \quad q_{it} = \alpha_l l_{it} + \alpha_m m_{it} + \alpha_k k_{it} + \omega_{it}.\]

\[^4\text{Time variation in both output and input prices can be accommodated either by deflating the variables appropriately using industry-wide deflators or by including a full set of year fixed effects. In any case, the price variables would not be indexed by} \ i.\]

\[^5\text{De Loecker et al. (2012) provide a framework that considers the joint estimation of productivity, marginal costs, and markups.}\]
the usual practice in empirical work in this area. Let \( \bar{p}_I^t \) be the price deflator for industry \( I \); \( z_I^t \) be a vector of industry-specific input price deflators; and \( p^*_ix^t = p^*_i - \bar{p}_I^t, z^*_ix^t = z_I^t - \bar{z}_I^t, s^*_ix^t = s^*_i - \bar{p}_I^t \) and \( e^*_ix^t = e^*_i - \bar{z}_I^t \) denote deflated output price, input prices, sales, and expenditures, respectively. Researchers typically take the following version of Equation 2 to the data:

\[
\begin{align*}
\hat{s}_{it} & = c_{it}^o \alpha + \omega_{it} + p^*_ix^t - z^*_ix^t \\
& = (x_{it}^o + z^*_ix^t) \alpha + \omega_{it} + p^*_ix^t - z^*_ix^t \alpha.
\end{align*}
\]

First, consider the case in which there is output price but no input price variation across firms. In this case, \( z^*_{it} = 0 \), and the terms related to input price variation drop out from Equation 4. This is the typical case considered in industrial organization studies that assume that input prices are equalized across firms (once regional differences are controlled for). Output price variation captured by \( p^*_it \) will generally be correlated with \( e^*_it \) (which is equal to \( x_{it}^o \) in this scenario): Ceteris paribus, we expect firms that charge higher prices to sell lower quantities, which in turn implies lower input quantities. Hence, the correlation between \( p^*_it \) and \( x_{it}^o \) is likely to be negative, leading to a downward bias in the estimates of the coefficients \( \alpha \) as well as the returns to scale. This bias is the focus of the work of Klette & Griliches (1996), as well as De Loecker (2011). We refer to it as the output price bias. A closely related point is made by Katayama et al. (2009), who allow for input prices to vary across firms but presume that this input price variation can be completely controlled for when one observes firm-specific wages (so that \( z^*_{it} \) can be treated as an observable). As we argue below, this assumption is strong—even when wages are observed, the prices of other inputs (e.g., materials) are typically not observed, while a firm-specific price for capital is never available.

Second, consider the case in which there is no output price variation, but input prices vary across firms. This case may seem irrelevant in practice, but as we show below, it corresponds to the case in which we observe firm-specific output prices and hence can control for output price variation, while input price variation remains uncontrolled for. In this case, it is evident from Equation 4 that input price variation will lead to a strong negative bias in the estimated coefficients; a firm that faces higher input prices will have higher input expenditures that will not lead to higher physical output. We refer to this bias as the input price bias. In contrast to the output price bias, the input price bias has received no attention in the literature so far; the only study we are aware of that has attempted to address it is De Loecker et al. (2012).

Realistically, the data will be characterized by both output and input price variation so that estimates of Equation 1 will suffer from both output and input price biases. Unless these two biases interact in a way so as to offset each other, they will lead to biased estimates of the production function coefficients. Furthermore, even conditional on the parameter vector that the estimation delivers in this case, one will be able to recover only a (biased) estimate of the composite residual \( \pi_{it} \).

The message so far is that with sales and expenditure data alone, one cannot generally recover the underlying components of firm performance or identify productivity. As always in empirical work, there are two ways out of such a situation: Either one collects more data or one makes additional assumptions that will allow identification.

### 2.2.1. Additional data

One might be tempted to conclude at this point that the solution to all issues discussed in this section would be to collect more data on firms’ output and input prices. Obviously, more data are always preferable (at a minimum, one can ignore them). However, the introduction of additional data creates its own challenges; although more data may help alleviate some of the problems discussed above, they are not a panacea.
Output prices tend to be more readily available than input prices in firm- or plant-level surveys. However, these data are often unit values (derived by dividing revenues by quantities over a period of time) and suffer from the well-documented problems associated with them. Furthermore, their use in the context of production function estimation poses additional challenges associated with differences in the units in which quantities are recorded across firms and products. Finally, the attempt to exploit output prices forces the researcher to explicitly confront issues that are specific to multiproduct firms; output prices are recorded at the product level, which calls for estimation of production functions at the product level. However, input expenditures are recorded only at the firm level. Hence, even when output prices are available, estimation of production functions for multiproduct firms is not possible unless one adopts one of three approaches: (a) eliminate multiproduct firms from the sample and focus on single-product firms only, (b) aggregate product prices to the firm level and conduct the analysis at the firm level, or (c) devise a mechanism for allocating firm input expenditures to individual products and conduct the analysis at the product level.

Each of these approaches has its drawbacks. Given that multiproduct firms account for a significant fraction of output in the manufacturing sector, eliminating them from the analysis is hard to defend. Approach (b) requires one to specify a demand system that allows aggregation in a consistent manner and creates the need for additional assumptions. Similarly, approach (c) requires spelling out the assumptions needed to allocate input expenditures across products.

That the use of output price data requires assumptions is not necessarily a weakness, especially because the assumptions one needs in this context either are already made (albeit implicitly) in the existing literature or are weaker than the ones required under alternative approaches. Nevertheless, our point is that more data do not eliminate the need for assumptions and structure.

Assuming the challenges outlined above can be dealt with, the use of output prices should allow one to eliminate the output price bias discussed above. Syverson (2004) and Foster et al. (2008) provide examples of studies that have accomplished this successfully. They rely on a selected set of plausibly homogeneous good industries (e.g., ready-mixed concrete) and exploit output price data to separate out price variation from productivity. An implicit assumption in their framework is that input prices do not vary across firms. This assumption is indeed plausible in the context of the homogeneous product industries they consider; for example, it is plausible to assume that (conditional on region) the input prices ready-mixed concrete producers face are the same. In this setting, the only bias present in Equation 4 is the output price bias, and this bias can be successfully dealt with when one observes output prices.

However, the focus on homogeneous product industries leaves us with the bulk of economic activity unaccounted for. The set of industries characterized by substantial product differentiation comprises a large share of economic activity. In such industries, controlling for output prices alone is insufficient; differentiated products require differentiated inputs, so that we would expect input prices to vary across firms, even when these firms are located in the same region and even when input markets are perfectly competitive. Input prices are typically unobserved. Controlling for output prices in this case will eliminate the output price bias, but will leave the input price bias intact, and in fact will make its consequences for estimation more salient. In this case, Equation 4 becomes

\[ q_{it} = (x_{it} + z_{it}^e)\alpha + \omega_{it} - z_{it}^e\alpha. \quad (5) \]

The problem with trying to estimate this version of Equation 4 is immediately apparent: Without controlling for the (unobserved) variation in input prices, the coefficients \( \alpha \) will be biased. The seriousness of this problem is demonstrated in De Loecker et al. (2012): The authors estimate
a physical production function for Indian manufacturing that relates physical output to (deflated) input expenditures. When input price variation is not controlled for, the coefficients $\alpha$ often seem nonsensical and have the wrong sign. Yet these apparently nonsensical results make a lot of sense in the presence of input price bias. Consider, for example, two firms that produce shirts and use the same technology. However, one firm uses silk as an input to produce silk shirts, while the other firm uses less expensive cotton to produce (less expensive) cotton shirts. Suppose that both firms produce the same number of shirts in a period. If we relate the number of shirts produced to (deflated) expenditures on materials, we would find that the firm that uses higher expenditures (silk) produces as much as the firm that uses lower expenditures (cotton). Hence, the coefficient on materials will be negative. Note that if one had not corrected for the output price bias in this case and had used deflated revenues instead of quantities as the left-hand side variable, the problem would have been less transparent as silk shirts would be associated with higher revenues so that higher expenditures would have led to higher revenues. Of course, that the problem is in this case less transparent does not mean that the problem does not exist.

Ultimately, the source of the problem is that products and inputs are differentiated. De Loecker et al. (2012) address this issue by introducing a control function for the (unobserved) input prices. This is based on the premise that conditional on regional variation, input price differences reflect quality differences in the inputs across firms. The authors go on to argue that the assumptions underlying this control function are weaker than the ones required under alternative approaches that develop full-fledged models of product differentiation (with assumptions on particular demand functions and market structure). Nevertheless, their approach once again demonstrates the need for assumptions; data alone do not eliminate the need for economic structure.

Finally, one might argue that if we were able to observe, in addition to output, input prices for every single input, then we would be able to estimate the structural equation (Equation 2) with only a limited set of assumptions. Although in theory this is an appealing prospect, in practice it is unlikely that we will ever be able to control for input price variation using data. To our knowledge, the only survey that contains information on prices of materials purchased is the one for Colombian manufacturing firms, but even there, the firm-specific price of capital is not observed, and utilizing the detailed input price information for materials is a challenge itself because of a variety of measurement issues. We conclude that although the use of additional data on prices can improve on certain aspects of estimation and identification, the need to introduce assumptions remains. The question is not whether one needs assumptions, but which set of assumptions is less restrictive.

### 2.2.2. Additional assumptions

Let us return to Equation 3 and consider the typical case in which we have a standard data set containing (deflated) sales and expenditures $(s_{it}, e_{it})$ for a panel of firms in a given industry. We now ask, Under which assumptions can we recover the structural parameters and the components of firm performance when estimating Equation 3?

One case that allows us to recover productivity $(\omega_t)$ has already been discussed: the case in which there is neither output nor input price variation across firms. The assumptions required to produce this lack of price variation are strong. Therefore, we ask whether it is possible to achieve the same result with a set of less restrictive assumptions; specifically, we consider a setting in which there is both output and input price variation and ask under which conditions these two types of variation will interact in such a way that the output price bias exactly offsets the input price bias, that is, $p^*_i - z^*_i \alpha = 0$. We show that the following assumptions are required for the price bias to be completely eliminated:

1. The industry is characterized by monopolistic competition.
2. Firms produce a horizontally differentiated product and face the same constant elasticity of substitution (CES) demand system.
3. Production is characterized by constant returns to scale (CRS).
4. Input price variation (across firms and time) is input neutral, such that \( z_{lt}^b = \lambda_{lt} \forall b = \{1, 2, \ldots, H\} \), with \( H \) the total number of inputs.

The first two assumptions imply that firms will pass through costs completely to prices; therefore, any input price variation, both in the cross section and in the time series, will be completely reflected in output price variation so that \( p_{lt}^* = z_{lt}^* \). The fourth assumption is required to make sure that in any cross section of firms, the input price variation is restricted to a scalar. The CRS assumption is important to guarantee that \( \sum_b \alpha_b = 1 \), and consequently the input price variation exactly offsets the output price variation.\(^6\) To see the impact of this assumption, let \( \sum_b \alpha_b < 1 \). This would lead to a price error of \( p_{lt}^* - \sum_b \alpha_b \lambda_{lt}^* \), which, even under complete pass-through, would leave \( 1 - \sum_b \alpha_h \lambda_{lt}^* \) in the error term, biasing the estimates of interest.

Although the above assumptions are somewhat weaker than the assumption of no output or input price heterogeneity, they are still restrictive and inconsistent with a substantial body of work in the international and macroeconomics literature that has documented incomplete pass-through of cost shocks to prices (see Goldberg & Knetter 1997 for an overview of this literature). Hence, the question arises whether there are alternative assumptions that would allow for a more satisfying treatment of unobserved prices. We briefly discuss the set of assumptions underlying two recent papers that explicitly address these biases: De Loecker (2011) and De Loecker et al. (2012).

**Output price heterogeneity.** De Loecker (2011) works with standard data in which output and input prices are unobserved. He assumes away input price variation across firms\(^7\) and focuses on addressing the output price bias. To this end, he explicitly introduces a demand system in his empirical model; in the particular application he considers, he works with CES, but his approach works also for alternative demand systems as long as one can relate log quantity to log prices and additional demand shifters; such demand systems include the nested logit and the random coefficient demand model as in Berry et al. (1995), in which log prices enter the indirect utility function.

The basic idea behind De Loecker’s approach is to use the demand structure to express the (unobserved) price variation \( p_{lt}^* \) as a function of observables, in his case, an industry quantity index \( q_{lt} \) and a set of product dummies \( D_i \). De Loecker (2011) shows that this gives rise to the following version of Equation 4:

\[
\begin{align*}
  s_{lt}^* &= x_{lt}' \beta + \beta_d q_{lt} + D_i \delta + \omega_{lt},
\end{align*}
\]

where \( \beta_d = 1/|\eta_I| \), with \( \eta_I \) the elasticity of demand of industry \( I \); \( \beta = ((\eta_I + 1)/\eta_I) \alpha \), with \( \alpha \) denoting the production function coefficients; and \( \omega_{lt} = ((\eta_I + 1)/\eta_I) \omega_u \). Estimation of the above equation delivers the demand elasticity \( \eta_I \), along with the true production coefficients \( \alpha \), and allows one to separate the physical productivity \( \omega_u \) from the output price variation (\( \beta_d q_{lt} + D_i \delta \)).

The results are consistent with one’s priors; the production function coefficients obtained after correcting for output price bias are larger in magnitude (consistent with the presence of a

---

\(^6\)Consider a two-input production function with labor and capital \((l, k)\). Under assumption 4, the structural error term can be written as \( p_{lt}^* - a_l w_{lt} - a_k r_{lt}^* = p_{lt}^* - a_l \lambda_{lt}^* - a_k r_{lt}^* = a_l (\lambda_{lt}^* - \lambda_{lt}) \), where \( w \) and \( r \) denote the (log) prices of labor and capital, respectively; asterisks denote deviations from industry averages; and \( \lambda_{lt} \) is a scalar that captures the input price–neutral variation of input prices relative to the input price indexes \( (\lambda_{lt} = w_{lt}^* = r_{lt}^*) \). CRS guarantees that \( (\alpha_l + \alpha_k) = 1 \), and with complete pass-through, \( p_{lt}^* = \lambda_{lt} \).

\(^7\)This is plausible in the context of Belgian textile producers who are geographically concentrated.
downward bias when output price bias is not controlled for) and suggest increasing returns to scale. The approach rests crucially on assuming a demand system, but assumptions on the demand system and cost pass-through are implicit whenever one estimates Equation 4 without controlling for the output price bias.

**Output and input price heterogeneity.** De Loecker et al. (2012) work with a different data set from India that contains information on product-specific prices (i.e., unit values). As noted above, the availability of output price data is not a silver bullet, at least not when one considers a large set of differentiated product industries. Although the output price data allow De Loecker et al. (2012) to eliminate the output price bias without resorting to any assumptions, they still need to address the input price bias. To do so, they assume that the only source of input price variation across firms (apart from regional differences) is quality differentiation; this assumption rules out imperfect competition in input markets. Furthermore, they assume an output quality production function that displays complementarities in the qualities of inputs and output: Higher-quality output demands higher-quality inputs, and high-quality inputs are complements to each other. Under these assumptions, the authors show that conditional on regional variation, input prices will be a function of output quality, which can be proxied through a flexible polynomial in output prices, market shares, product dummies, and interactions thereof. This polynomial represents a control function for input prices and is consistent with a large set of alternative demand and market structures and the main models used in industrial organization and international trade. Given that output prices are observed, the output price variation is eliminated from both the left- and right-hand sides of Equation 4, which becomes

\[ q_{it} = e_i'\alpha + \omega_{it} + z_t(p_{it}, ms_{it}, D_i, G_i), \]  

(7)

where \( p_{it} \) is the output price of the firm, \( ms_{it} \) is a vector of market share variables (including unconditional and conditional market shares), \( D_i \) captures product dummies, and \( G_i \) denotes firm location. The function \( z_t(\cdot) \) serves here as a control for the unobserved input price variation \( z_{it}'\alpha \). Conditional on productivity and input price variation captured by \( z_t(\cdot) \), we obtain the correct structural production function parameters by considering how variation in physical input use maps into variation in physical output.

In sum, whereas the bulk of empirical work has focused on estimating the sales generating function (Equation 1), recent papers have focused on estimating the structural equation (Equation 2) by correcting for the output and/or input price biases. In all cases, the correction involves not only additional data, but also explicit statement of the assumptions under which the correction is valid.

### 2.2.3. A classification of existing work.

In Table 1, we classify existing papers on the broad subject of globalization and firm performance based on how they deal with unobserved

<table>
<thead>
<tr>
<th>Output price (( p_{it} ))</th>
<th>Input price (( z_{it} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{it}^* )</td>
<td>( z_{it}^* )</td>
</tr>
<tr>
<td>Case A (standard framework)</td>
<td>Case C (observing output prices)</td>
</tr>
</tbody>
</table>

\( \overline{p}_{il}^f \) is the price deflator for industry \( l \); \( z_{il}^f \) is a vector of industry-specific input price deflators; and \( p_{it}^* = p_{it} - \overline{p}_{il}^f \) and \( z_{it}^* = z_{it} - z_{il}^f \) denote deflated output prices and input prices, respectively.
output and input prices. The columns and rows indicate which kind of price variation is controlled for.

**Case A: standard framework.** Most existing work uses output and input price deflators common across firms that capture industry-wide movements in output and input prices. The only price variation occurs in the time series, and any variation away from these industry-wide deflators will introduce the output and input price biases discussed above.

**Case B: De Loecker (2011).** In the setting considered in De Loecker (2011), price variation across firms is controlled for by explicitly introducing a demand system. On the input side, it is assumed that firms face common input prices.

**Case C: observing output prices.** In this case, input prices vary across firms, but output prices do not. As noted above, this case corresponds to a specification estimated in De Loecker et al. (2012) for expositional purposes, in which physical quantities are regressed against deflated input expenditures, as in Equation 5.

**Case D: pass-through literature and De Loecker et al. (2012).** Perhaps the only literature that has allowed for both output and input price variation to characterize firms active in international markets is the pass-through literature. (Representative papers include Goldberg & Verboven 2001, Nakamura & Zerom 2010, Berman et al. 2012, Goldberg & Hellerstein 2013, and Amiti et al. 2014.) This literature recognizes both that prices vary across firms and that firms respond incompletely and differentially to cost shocks; this incomplete response generates not only price dispersion across firms, but also, for a particular firm, price dispersion across destinations. The cost shocks considered in the pass-through literature are exchange rate changes, but the insights of this literature are equally applicable to other cost shocks, including input price shocks or changes in tariff and other trade policies.

In contrast to the literature described above that takes a production function approach, the pass-through literature relies on a demand-based approach in which assumptions on the demand side, market structure, and firm behavior are combined to derive measures of firm performance. Although this literature offers the advantage of a much richer treatment of heterogeneity and product differentiation, it comes at the cost of an extensive set of assumptions. These assumptions seem defendable in the context of case studies of particular industries, for which knowledge of the institutional details can guide the choice of the appropriate structure, but are more controversial when applied to a large cross section of industries. We therefore do not devote any space to discussing this literature here but focus on the production function–based approach instead. From a production function perspective, to our knowledge, only De Loecker et al. (2012) develop a framework that accounts for both output and input price variation and estimate Equation 7. Interestingly, the insights obtained using their approach turn out to be consistent with the main insights of the pass-through literature; not only do the findings indicate substantial heterogeneity across firms, but the authors find evidence consistent with incomplete pass-through of cost shocks to prices.

**2.2.4. An example.** We conclude this subsection with an example. We have a sample of 318 single-product Indian textile producers over the period 1989–2003 for which we observe (deflated)
We consider a standard Cobb-Douglas production function and highlight the output and input price biases using four distinct specifications:

1. An OLS regression of sales against expenditures
2. An OLS regression of quantity against expenditures
3. An OLS regression of quantity against expenditures and a control function for input prices that includes only the output price (this control function is a special case of the control function used in De Loecker et al. 2012; it rests on a vertical differentiation model of consumer demand in which the output price is a sufficient statistic for quality and hence input price variation)
4. A special case of De Loecker et al. (2012) that assumes an AR(1) process for productivity and, as in specification 3, a control function for input prices that depends only on output price

Table 2 lists the estimated coefficients.

Specification 1 generates perhaps the most familiar numbers for the various coefficients. This OLS regression is a useful way to describe the underlying data and check that the estimates are within the range of existing studies; there is a long list of papers that estimate production functions using different data sets. Therefore, the literature in this area has settled on what are reasonable-looking estimates.

In specification 2, we consider the case represented in Equation 5: Physical quantity is projected onto deflated expenditures, leaving input price variation uncontrolled for. We get results that are hard to interpret at first, such as negative labor and capital coefficients. But our framework actually predicts, or at least is consistent with, negative coefficients. Just as in our shirt production example, we find that firms that spend more on labor produce less output but generate more sales.

In specification 3, we stick to a simple OLS regression but add a third-order polynomial in the firm’s output price. We recover coefficients similar to specification 1. This OLS regression is a very useful diagnostic check for the problem at hand: By merely including output prices, while ignoring sales and expenditures on labor, intermediate inputs, and capital; in addition, we observe product-level prices.8

Table 2 Estimated coefficients: evaluating the price biases

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Revenue</th>
<th>Physical output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Specification 1</td>
<td>Specification 2</td>
</tr>
<tr>
<td>Labor</td>
<td>0.162</td>
<td>−0.029</td>
</tr>
<tr>
<td>Materials</td>
<td>0.812</td>
<td>0.576</td>
</tr>
<tr>
<td>Capital</td>
<td>0.035</td>
<td>−0.514</td>
</tr>
</tbody>
</table>

| The data were constructed by De Loecker et al. (2012). We observe 318 producers of textiles (industry code PNIC 17). We omit the standard errors, but all coefficients are significant at the 1% level, as is common in the estimation of production functions. Specifications 1, 2, and 3 are ordinary least squares regressions of the relevant dependent variables (revenue in 1 and physical output in 2 and 3) on deflated expenditures, plus a polynomial in output price in 3. Specification 4 estimates the coefficients using generalized method of moments, in which we rely on lagged inputs as instruments exploiting the variation in adjustment costs in labor and capital, while allowing current productivity shocks to affect current material choices. Readers are referred to De Loecker et al. (2012) for further discussion.

8 We consider only single-product producers to demonstrate the price bias in isolation. Using multiproduct firms requires a treatment of the unobserved input allocations, which is not the focus of this article [see the discussion under Section 2.2.1, option (a)].

9 We obtain almost identical coefficients if we include output price in a linear fashion.
all other well-known identification problems, we generate plausible production function coefficients.\textsuperscript{10} Of course, the proper benchmark should by no means be specification 1. However, it is reassuring that we get positive output elasticities for all three inputs and that the returns to scale now look sensible ($\sum \hat{a}_b = 1.015$).\textsuperscript{11}

Finally, in specification 4, we allow for both unobserved productivity shocks to affect input choices and for heterogeneous input prices. This specification can be thought of as a special case of De Loecker et al. (2012): We control for serially correlated productivity using an AR(1) process for productivity, and for unobserved input price variation by including a control function in output prices as in specification 3. The coefficients are similar to those in specification 3, but there are a few differences, as expected. By controlling for unobserved productivity differences, we undo the negative correlation between capital and productivity and recover a substantially higher capital coefficient; furthermore, we find a lower coefficient on materials owing to the positive correlation between input use and productivity. Table 2 does not list the remaining structural parameters describing the relationship between input prices and output price—i.e., $z_t(p_{du})$. But, interestingly, we find that $(\partial z_t(\cdot))/\partial p_{du} > 0$ for all firms and time periods, confirming that output prices are positively correlated with input prices.\textsuperscript{12}

### 2.3. Interpreting the Performance Residual ($\pi_{it}$)

Equation 2 is useful for the remainder of this article for two reasons. First, it allows us to highlight the potential channels through which globalization can affect firm performance: Firms competing in international markets are likely to adjust their scale of production ($x_{it}$ and hence $e_{it}$), productive efficiency ($\omega_{it}$), and prices and associated markups ($p_{it}$), as well as product and input quality (reflected in both $p_{it}$ and $z_{it}$). Second, we can classify almost all studies on the subject based on which component(s) of profitability they have focused on in each instance.

Existing work has, in one form or another, studied how episodes of trade reform affect productivity, prices, and quality across a wide range of countries, industries, and time periods. Below we argue that a large part of the literature has recovered only the impact on profits $\pi_{it}$, without decomposing it into the separate effects on the underlying factors. But in most instances, we care about the exact mechanisms through which trade affects firm performance, as these have different distributional and potentially also aggregate welfare implications.

Our production function framework suggests ways for identifying the components of the structural error $\pi_{it}$. After estimating the structural production function and demand parameters, De Loecker (2011) recovers separate estimates of productivity and markups, and consequently estimates of (average firm-level) prices. This is accomplished by committing to a particular demand system and market structure (i.e., CES demand paired with monopolistic competition). However, even when we are not willing to make assumptions on the underlying demand system and market structure, we may still be able to recover, if not the levels, at least the changes in markups using the

\textsuperscript{10}We have found output price to have a first-order effect when we correct for input price variation. The use of output price as the sole control can be justified on the basis of a vertical differentiation model, in which price is a proxy for output quality, and hence also input quality and input prices. Additional variables, such as market shares and product dummies, make the control function more general and consistent with a larger set of demand models but lead to very similar results. Readers are referred to De Loecker et al. (2012) for more discussion on the estimation and identification of this control function, $z_t(\cdot)$.

\textsuperscript{11}The OLS regressions do not control for the standard simultaneity bias, so even if we observed $q_{it}$ and $x_{it}$, we would not recover the production function coefficients $\alpha$.

\textsuperscript{12}This positive correlation is also documented by Kugler & Verhoogen (2012), who have the advantage of directly observing some input prices (e.g., wages and the prices of materials) as well as output prices in their Colombian data.
production function framework. If researchers observe prices, they can further obtain estimates of marginal costs.

De Loecker & Warzynski (2012) show how to recover markups from production data. The essential insight is that for any variable input free of adjustment costs, the markup drives a wedge between the input’s output elasticity and the input’s revenue share. The latter is directly observed in the data; the former is not but can be estimated. In the context of our Cobb-Douglas production function, the markup $\mu_i$ for firm $i$ at time $t$ is given by

$$
\mu_i = \frac{\alpha_v}{E^m_{it}},
$$

where $\nu$ stands for variable. Depending on the application, variable inputs can include labor, electricity, or any other intermediate input.

To illustrate this approach, let us consider the same data on Indian textile producers. Let materials be a variable input in production. Using Equation 8, we compute the markup as $0.634 \times (S^m_{it}/E^m_{it})$ and obtain an average markup of 1.30 with a standard deviation of 0.65. The markup distribution suggests considerable variation, with the 25th percentile firm breaking even with a markup of approximately 1, while the 75th percentile firm makes substantial profits (excluding fixed costs) with a markup of approximately 1.42.

The markup calculation relies on estimates of the production function, which in turn deliver estimates of productivity. The specifics depend on the data at hand, and we refer readers to De Loecker & Warzynski (2012) and De Loecker et al. (2012) for detailed discussions of the issues that arise in the context of different data sets.

In sum, the production function framework has the potential to generate separate measures of productivity and markups without commitment to specific demand and market structure assumptions, as is common in the demand-oriented industrial organization literature on imperfect competition. Once the components of profitability are identified, one can examine how globalization affects each of these components. This is the subject of the next section.

3. MECHANISMS

We use the framework introduced above to discuss the main mechanisms through which participation in international markets (and, in particular, trade reforms) affects performance. We denote trade reforms by $T_{it}$ and allow them to affect firms differently over time. This specification allows for firms in an industry to produce different products, for instance, so that they end up facing different rates of protection. Whether there is variation across firms is crucial for any identification strategy that tries to recover the causal effect of $T_{it}$ on firm performance and its components. The mechanisms we discuss below can be broadly classified into two categories: mechanisms that induce changes within firms and hence affect firm-level components of profitability and mechanisms that induce the reallocation of economic activity across firms in an industry. In the latter case, firm-level profitability may be unaffected by trade, but trade-induced reallocation of resources from less to more profitable firms can still lead to better performance at the industry level.

3.1. Within-Firm Changes

Firms participate in international markets both as producers/sellers of goods and as buyers of intermediate inputs used in the production of these goods. Trade policies may affect both aspects of firm activity. Specifically, changes in the protection of final products (e.g., reductions in output
tariffs) will affect the competition domestic producers face, and changes in the protection of intermediate inputs (e.g., input tariffs) will affect the costs of production. The channels through which trade reforms affect firms will accordingly depend on the specific nature of the trade policy changes, and in particular on whether these affect output versus input markets. Therefore, in the course of our discussion, we often find it necessary to make a distinction between output- and input-oriented trade policies (for expositional purposes, we base our discussion on tariffs given that these are easily measured, but in principle the arguments apply to any other trade policy).

We use Equation 2 as the basis of our discussion to differentiate between mechanisms that affect (a) the firm-level productivity \( v_{it} \), (b) the expenditures \( e_{it} \) and their components (input quantities \( x_{it} \) and prices \( z_{it} \)), (c) output prices \( p_{it} \) and markups, and (d) none of the above, but induce within-firm reallocation in multiproduct firms.

3.1.1. Reduction of X-inefficiencies and management practices. The perhaps most advocated argument for opening up a country to foreign markets is that exposure to international competition increases the efficiency of the previously protected domestic producers. In terms of our framework, this channel would lead to an increase in the physical efficiency \( v_{it} \). But why would the efficiency of these producers increase? A popular argument is that intensified competition will reduce X-inefficiencies at the firm level. Although intuitive, this argument has little theoretical appeal in its simplest form; why were firms willing to leave money on the table prior to the trade reforms? A potential answer is that in practice, the reduction of X-inefficiencies is costly, and therefore it takes an increase in competition for firms to find it profitable to undertake the actions necessary to become more efficient. For example, in the face of intensified competition, firms might find it necessary to replace old, inefficient managers by more competent ones, or adopt better management practices. Although these considerations feature prominently in casual discussions of trade and productivity, we are aware of only a handful of papers that formalize these arguments. (Readers are referred to Schmidt 1997, and more recently Bloom et al. 2013, for an explicitly theoretical framework that relates competition to managerial incentives.)

From an empirical point of view, this mechanism suggests a reduced form way of introducing trade policy by making physical efficiency a function of trade policy:

\[
\omega_{it} = \omega(T_{it}).
\]  

(9)

Note that the above mechanism suggests that the relevant policy is one that affects the output side of the firm, as the motivation for reducing X-inefficiencies arises from the exposure to intensified competition; in the case of tariffs, for example, the relevant measure of trade policy would be output tariffs.\(^{13}\)

3.1.2. Feedback effects. The reason we expect firms to increase their productivity in response to a trade shock is that we expect them to undertake actions to become more efficient. Some of these actions may be unobservable to the researcher, in which case they will be subsumed in the residual \( \omega_{it} \) (e.g., this is the case if firms replace current managers by better ones or adopt better management practices and these actions are not reflected in changes in expenditures). But in many instances, improvements in productivity will be associated with actions that are observable (e.g., investment in new technologies, R&D, and entry in export markets). In these cases, the law of motion of productivity should be modified so as to explicitly allow for these actions to affect productivity:

\(^{13}\)Below we discuss the importance of including both contemporaneous and lagged trade policy variables to accommodate the role of expectations and dynamics.
\[
\omega_{it} = g(\omega_{it-1}, A_{it-1}) + \xi_{it}.
\] (10)

The term \(A_{it-1}\), denoting any action undertaken by the firm to increase its productivity, is lagged given that it likely takes time for actions to take effect. Of course, that these actions (e.g., investment, R&D, exporting) are allowed to affect productivity does not mean that they will in fact do so. The above law of motion is entirely consistent with a finding that the action undertaken by the firm did not have an effect on productivity ultimately; hence, it does not assume the result. Nevertheless, if one believes that a certain action is likely to affect productivity, it is imperative to include it in the law of motion.

Productivity-enhancing actions will typically be correlated with the inputs in the production function so that their omission from the law of motion will generate an omitted variable bias. This is the main point of De Loecker (2013), and it can be made clear using the example of a productivity-enhancing investment. The investment will affect not only productivity, but also the capital stock. Suppose that we do not allow investment (the action) to affect productivity through the law of motion. Then the estimation of the production function will suffer from an omitted variable bias that will generate a biased capital coefficient; specifically, given that higher investment will likely be associated with higher capital, we would expect an upward bias in the capital coefficient estimate. Moreover, a second-stage regression relating productivity to investment would tend to underestimate the role of investment for the same reason; given that investment was not included in the first stage that estimated the production function, the improvement in firm performance will be attributed to the higher capital and not to the productivity improvement.

In conclusion, recognizing that productivity evolves endogenously in response to firms’ actions calls for a modified law of motion for productivity at a minimum. Ideally, one would like to supplement this law of motion with an explicit model of how the actions \(A\) are determined. This is done, for example, in recent papers by Bustos (2011) and Aw et al. (2011). The relevant actions are exporting and adoption of new technology in the first paper and exporting and R&D in the second, and in both cases, the authors use structural models to show how these actions are determined and how they respond to trade liberalization.

3.1.3. Input side. In Equation 2, we explicitly distinguish between the two components of expenditures \((e_{it})\): physical input use \((x_{it})\) and input prices \((z_{it})\). Both components will typically vary across firms and will be affected by globalization; for example, a reduction of input tariffs will have a direct effect on the prices of imported inputs, and hence \(z_{it}\). There are three distinct reasons for firm-specific input prices \((z_{it})\), even for firms active in narrowly defined industries: (a) pure geographical variation in input prices (e.g., local labor markets and constrained labor mobility imply regional differences in wages), (b) variation in input quality leading to differences in input prices, and (c) firm-specific input prices due to monopsony power in input markets. The literature on production function estimation has typically ignored or assumed away heterogeneity in input prices due to quality differences across producers or to imperfect competition in input markets.

To highlight the forces generating variation in expenditures across firms, let us write \(e_{it}^{b} = x_{it}^{b} + z_{it}^{b}(\nu_{it}^{b}, G_{it})\) for a given input \(b\). The term \(\nu_{it}^{b}\) refers to the quality of input \(b\).\(^{14}\) We collect all other firm-specific factors determining input prices, including the firm’s geographic location, 

\(^{14}\)We use the term quality to capture differences in observed and unobserved attributes of a given input (e.g., skill differences across workers). Whether one can measure input quality will depend on the data at hand. Standard firm-level data usually provide us with the total use of intermediate inputs in dollars, and sometimes even with physical units, but will typically not record input characteristics or direct measures of quality.
in $G_{it}$, which will capture, among other things, the (re)location of economic activity (e.g., plant closings, offshoring) induced by a firm’s exposure to global markets or other shocks.

Four distinct and largely disconnected literatures have focused on how globalization affects the various components of input expenditures. First, the trade and productivity literature has set out to measure how the transformation of physical inputs ($x$) to output ($q$) changes with increased foreign competition; indirectly, this literature also deals with the scale effects that operate through inputs ($x$). Second, the trade and quality literature has focused on whether producers upgrade (or downgrade) the quality of their products and inputs in response to increased exposure to international trade.$^{15}$ Third, the trade and labor literature has focused on how globalization affects workers of different skills. Finally, the literature on multinationals and offshoring has investigated how globalization affects firms’ locational choices (part of $G_{it}$). In all these cases, exposure to global markets affects $e_{it}$ directly.

A different mechanism through which globalization can affect firm performance is highlighted by Halpern et al. (2011). So far, we have abstracted from the fact that the materials used in the production, $m$, are a composite of many different domestic and imported intermediate inputs. Acknowledging the aggregation underlying $m$ suggests an additional channel for performance improvements: If the reduction in trade costs leads to the import of new intermediate inputs, then we would expect an increase in production beyond the one predicted by the increase in expenditures. This increase will be more pronounced if the new inputs are of higher quality compared to the ones previously used, but the argument does not rest on quality improvements: As long as the production technology exhibits a taste for variety, a larger number of imported inputs will imply higher output.

The simplest way to make this point is to consider (as in Halpern et al. 2011 or Goldberg et al. 2010) a standard Cobb-Douglas production function in capital, labor, a set of intermediate inputs $M$, and productivity $\Omega$. To keep notation manageable, we abstract from quality differences between imported and domestic products and suppress the firm and time subscripts. Each intermediate input $M_j$ is assembled from a combination of a domestic and imported variety:

$$Q = \Omega L^\alpha L K^\alpha K \prod_{j=1}^{J} M_j^{\alpha_j},$$

$$M_j = \left[ M_{jF}^{\theta_j} + M_{jD}^{\theta_j} \right]^{\theta_j},$$

where $M_{jF}$ and $M_{jD}$ denote the quantities of the foreign and domestic inputs, respectively, and $\theta$ is the elasticity of substitution. Let us abstract from the output and input price biases. One can show that in this setting, Equation 2 takes the form

$$s_{it}^* = d_{it} = x_{it}^* \alpha + F_i(n_{it}) + \omega_{it} = e_{it}^* \alpha + F_i(n_{it}) + \omega_{it}.$$  

The term $F_i(n_{it})$ is a function increasing in the number of imported inputs $n_{it}$. Hence, if a trade liberalization increases the number of imported intermediates, we will expect to see a rise in

---

$^{15}$Schott (2004) and Khandelwal (2010) are classic references on the relationship between trade and product quality, and Verhoogen (2008) is a classic reference on the effects of globalization on both product and input quality. The main message of these papers is that producers often need to change/upgrade the quality of their products to enter foreign markets. These changes in product quality induce changes in input quality. Through the link between output and input quality, shocks in output markets affect factor markets.
measured productivity, $F(n_{it}) + \omega_{it}$. Why would a reduction in trade costs lead to the import of new intermediates? When deciding whether to import intermediates, firms balance the marginal cost savings associated with the new inputs against the fixed costs of importing. A reduction in the tariffs on inputs increases the cost savings associated with importing, leading to a larger number of imported intermediates. We note that this mechanism does not suggest any improvements in the physical productivity $\omega_{it}$. The issue here is analogous to the well-known gains from the introduction of new products; because the standard deflators used to obtain deflated expenditures $e_{it}$ do not account for new imported inputs, these new inputs will ultimately show an increase in measured productivity. This mechanism likely underlies the large within-firm productivity gains found in studies that examine the effects of input tariff liberalization, such as Amiti & Konings (2007) for Indonesia and Khandelwal & Topalova (2011) for India. In fact, Goldberg et al. (2009, 2010) explicitly show that the input tariff liberalization in India led to a large increase in the number of imported inputs.

3.1.4. Price and markup changes. It is natural to expect that firms will adjust prices and markups when faced with a trade shock. Many trade models assume CES preferences with monopolistic competition. Under these assumptions, markups are constant; hence, trade shocks do not affect markups. Recent theoretical work has moved away from these restrictive assumptions and considered alternative demand systems (while maintaining, for the most part, the assumption of monopolistic competition) to investigate how markups and prices respond to trade liberalization (see, e.g., Melitz & Ottaviano 2008, Feenstra & Weinstein 2010, Mayer et al. 2011, Arkolakis et al. 2012).

If we observe prices and have a plausibly exogenous source of variation for the trade policy $T_{it}$, it is conceptually straightforward to evaluate the price effects of trade reforms. We would start with the reduced form:

$$p_{it} = p(T_{it}) + \epsilon_{it},$$

where $\epsilon_{it}$ is a standard independently and identically distributed error term. A tougher task is to identify the specific channels leading to price changes (i.e., the cost and markup responses to changes in international competition). To highlight these channels, let us write the price as the sum of marginal cost $mc$ and markup $\mu$:

$$p_{it} = mc_{it}(q_{it}, z_{it}, \omega_{it}) + \mu_{it}(D, M).$$

The marginal cost $mc$ is a function of the quantity produced $q$, the input prices the firm faces $z$, and firm productivity $\omega$. The markup $\mu$ will be a function of the demand structure $D$ and the market structure and firm behavior, which we summarize in $M$. Trade shocks are expected to affect each of these components, with the specific effects depending on the particular nature of the trade shock. As noted above, CES preferences and monopolistic competition imply constant markups (in log terms); unless trade liberalization affects the costs $mc$, it will have no effect on prices. In more general setups, markups will be variable and respond to the trade reform.

To illustrate the mechanisms at work, let us consider a simple example that features linear demand and monopolistic competition. When assessing the effects of trade reforms, one must distinguish conceptually between reforms that affect the input markets (e.g., input tariff liberalization) and reforms that affect the output markets (e.g., output tariff liberalization). Both types of reforms are likely to affect both marginal costs and markups, albeit through different channels.

Consider a unilateral output tariff liberalization first. The decrease in tariffs exposes domestic producers to increased import competition. In the context of Equation 15, this translates to a change
in market structure and, consequently, the residual demand curve facing the firm. Figure 1a plots the initial equilibrium in this market, which occurs at the point at which the original marginal revenue curve intersects the marginal cost. For ease of exposition, we assume that marginal cost is constant and that it is not affected by trade liberalization. Trade liberalization implies intensified competition, so the residual demand curve will shift inward and become flatter. The new equilibrium occurs at a point at which both the price and markup are lower. This case corresponds to the standard intuition that trade liberalization, by intensifying competition, leads to lower prices and lower markups. Figure 1 allows one to trace the particular forces that shift price from its prereform level $p_0$ to its postreform level $p_1$.

Next consider a unilateral input tariff liberalization. Input tariff declines will have direct effects on the marginal cost; they will reduce the input prices $z_{it}$ and may further lead to improvements in productivity $\omega_{it}$ through the import of new intermediates discussed above. In the context of Figure 1b, this implies a downward shift in the marginal cost curve. The decline in input tariffs does not affect competition; hence, we would not expect any effects on markups arising from changes in the residual demand facing the firm. However, as long as the underlying structure does not imply constant markups (as is the case with the CES), markups will change as a result of the incomplete pass-through of the marginal cost change to price. This is shown explicitly in Figure 1b. The marginal revenue curve is not affected by the trade reform, while the marginal cost curve shifts downward. The postreform equilibrium is associated with a higher markup, although the price is lower than before. The reason is not that the environment has become less competitive. The higher markup arises as a result of the incomplete response of the price to the marginal cost change.

This apparently counterintuitive effect to trade and industrial organization economists is completely intuitive to international macro economists who have studied the incomplete response of prices to exchange rates. Just like input tariff shocks, exchange rate changes represent cost shocks to firms. It is well documented that prices respond incompletely to exchange rate shocks, a phenomenon known as incomplete exchange rate pass-through. Tariff and exchange rates have a symmetric effect on firms’ profits; applying the insights of the exchange rate pass-through

![Figure 1](https://www.annualreviews.org/static html/annualreviews.org/1462014.6:image/figures/F1.png)

**Figure 1**

Price, quantity, and markup response to trade liberalization. The solid lines represent the initial demand and cost conditions, and dashed lines indicate the new demand and cost conditions (i.e., post-trade liberalization).
literature to input tariffs immediately yields the result that with variable markups, input tariff reductions will lead to markup increases. To our knowledge, this is an insight that the bulk of the trade literature has missed; only De Loecker et al. (2012) address this issue explicitly and show that incomplete pass-through of input prices led to markup increases in the case of the Indian trade liberalization. Prices did decline as a result of the trade reform, but the price reductions were only a small proportion of the cost declines; the bulk of the cost reductions benefited firms in the form of higher markups.

In reality, most trade reforms, especially those implemented in developing countries, combine input with output tariff liberalization (i.e., the real-world trade liberalizations are characterized by a combination of the channels in Figure 1). In such instances, it is particularly important to assess not only the reform’s impact on prices, but also its effects on the price determinants. The reason is that the effects on input prices as well as markups can have important distributional implications.

3.1.5. Within-firm reallocation. A different mechanism that can lead to within-firm performance improvements is highlighted by Bernard et al. (2010) and is specific to multiproduct firms: Firms can improve revenue productivity, that is, \( \omega_{jt} + \frac{p_{jt}}{C_{jt}} \) in the context of Equation 2, by reallocating within-firm resources from the production of less profitable products to the production of more profitable products, where \( j \) denotes products. This mechanism is similar in spirit to the one discussed in Section 3.2 on the role of reallocation in increasing aggregate industry performance. The difference is that here this mechanism generates performance improvements at the firm level. Importantly, this mechanism does not hinge on any improvements on the physical firm productivity, \( \omega_{it} \), which remains unaffected. The distinction between physical and revenue productivity is therefore important here: It is only revenue productivity that increases, and this increase is brought about entirely through the reshuffling of resources across products with different profitability.

3.2. Reallocation: Aggregate Effects

Above we focus on the potential effects of trade liberalization on individual producers. However, at the end of the day, what we care about is how an industry, country, or group of countries is affected by trade. Reallocation of economic resources from less toward more profitable producers is one way in which industry (or country) performance can increase, even in the absence of any effects on individual firms. There is by now a large theoretical and empirical literature highlighting the aggregate productivity gains arising from such reallocation.\(^{16}\)

Collard-Wexler & De Loecker (2013) discuss some of the recent findings and point out that although we know by now that this reallocation process plays a substantial role in the data, it has been hard to identify specific forces that induce reallocation. Empirical work in international trade is perhaps the one big exception. The main advantage of studying large and arguably exogenous (at least from the perspective of an individual producer or industry) trade reforms is that they present us with exogenous shocks to the residual demand curves (in the case of output tariffs) and/or costs (in the case of input tariffs) facing domestic producers. We can then trace how the allocation of economic activity, usually measured by the market share in a particular market/industry, changes with the change in trade policy. The reshuffling of market shares toward the more productive/profitable firms has the potential to raise aggregate

\(^{16}\)For example, this reallocation mechanism is central to the Melitz (2003) model and many other follow-up papers that feature firm heterogeneity.
performance beyond the potential individual firms’ improvements discussed in the previous section. The extent to which the reallocation process is important depends of course on how dispersed profitability was initially, prior to the reforms. A well-known and influential study in this line of work is by Pavcnik (2002), who investigates the reallocation effects in the aftermath of the Chilean trade reforms.

Recently, Hsieh & Klenow (2009) put forward a simple theoretical framework to highlight this mechanism by focusing on wedges in marginal revenue products and pointing out distortions that can give rise to such wedges. Changes in both output and input tariffs fall nicely into this framework: Output tariffs present standard distortions in output markets as they constrain competition, whereas input tariffs generate distortions in capital and intermediate input markets. The reduction of these distortions through trade reforms should lead to a more efficient allocation of resources across firms.

This reallocation mechanism is in principle simple to measure in the data: calculate the covariance of productivity and market share for each time period and see how it reacts to the trade liberalization episode. However, the above discussion clearly demonstrates that the interpretation of the results will depend greatly on whether we rely on actual productivity \( v_{it} \) or a performance measure (such as \( \pi_{it} \)) that contains cost, demand, and market structure components. For example, if we relied on \( \pi_{it} \) and found that the market share of more profitable firms increased post–trade liberalization, this increase might not be desirable from an aggregate welfare point of view if it represented a shift toward firms with more market power rather than higher efficiency.

### 3.3. Static Versus Dynamic Effects

So far we have not made a distinction between trade shocks that affect producers instantaneously and shocks that affect producers with a lag. Many theoretical models are static in nature or focus on steady-state predictions that blur the distinction. However, for empirical models that wish to separately identify the impact of trade liberalization on the various components of performance, the difference between static and dynamic effects can be important.

Let us focus on the productivity channel \([\omega(T_{it})]\). It is reasonable to ask whether producers can immediately adjust their productive efficiency when faced with a change in protection. This question boils down to whether the trade reforms were expected, and hence whether producers had time to adjust prior to the actual change.

The standard working assumption is to assume that firm-level productivity moves over time according to a first-order Markov process: \( \omega_{it} = g(\omega_{it-1}) + \xi_{it} \). This specification is based on the observation that firm-level productivity (independently of how exactly it is measured) is highly persistent over time. However, the question remains whether we can think of trade liberalization as a shock that immediately affects productive efficiency. Take the example of a trade-induced reorganization of production. It arguably takes time for a firm to change the organization of production. Therefore, we would expect productivity to not react immediately or even possibly drop temporarily during the reorganization.

Therefore, it is important to let the function \( \omega(\cdot) \) be sufficiently flexible in how trade policy shocks enter, by including both contemporaneous trade policy changes and lags. For example, in the context of analyzing the impact of trade reforms on firm performance in Indian manufacturing, De Loecker et al. (2012) consider a law of motion that, in addition to lagged productivity, also includes lagged output and input tariffs and the firm’s lagged export status.
4. EVIDENCE

4.1. Profitability and Feedback Effects

If there is one robust finding that the literature has delivered to date, it is that industry profitability increases with exposure to foreign competition. This relationship is documented both in the time series, in studies that exploit trade liberalization episodes to identify the effects of trade openness on firm performance, and in the cross section, in comparisons of the performances of exporters and nonexporters. Most papers that exploit trade liberalizations focus on changes in output tariffs (Pavcnik 2002 is a classic reference), but recently, starting with the work of Amiti & Konings (2007), the focus has shifted toward input tariff liberalizations. Indeed, several studies on developing countries find the effects of input tariffs (representing direct cost shocks to firms) to be larger than those of output tariffs (which operate through changes in the competition facing firms). Furthermore, the literature finds evidence of both within-firm performance improvements and reallocation, with the relative importance of each channel depending on the particular setting. These findings are well documented in the literature, and we refer the reader to recent surveys by Melitz & Trefler (2012) and Melitz & Redding (2014) for a more extensive discussion.

At the cost of repeating ourselves, we emphasize once again that these findings of improved performance, although robust across countries and time, refer to profitability only and are therefore not particularly illuminating regarding the mechanisms at work. With this caveat in mind, it is worth pointing out that studies allowing for endogenous productivity evolution along the lines suggested in Section 3.1.2 find significant evidence of feedback effects. These feedback effects have two implications. First, in studies of trade liberalizations, they suggest that performance improvements are heterogeneous across firms, as firms with different characteristics (e.g., initial profitability levels, R&D expenditures, capital intensity) optimally choose different actions in response to trade shocks, which in turn affect their profitability. Bustos (2011), Aw et al. (2011), and Lileeva & Trefler (2010) find evidence of such heterogeneity. Second, in studies that compare exporters to nonexporters, feedback effects can lead to different conclusions regarding the relative importance of the effects of selection versus learning by exporting. De Loecker (2013), for example, demonstrates that in the case of Slovenia, learning by exporting seems to play an important role once one explicitly controls for the fact that entering export markets is associated with higher investment. These findings again point to the importance of employing empirical specifications that are consistent with the underlying mechanisms one has in mind when analyzing the data.

4.2. Mechanisms Underlying Profitability

It is only recently that research has focused on unpacking the mechanisms that generate the aforementioned performance improvements. Although it is still too early to draw general conclusions based on the findings of the few studies that explicitly distinguish between physical productivity and price effects, the evidence to date suggests that demand side and price effects are important and may be the primary factors generating the documented profitability increases.

To demonstrate the significance of these effects, consider De Loecker’s (2011) study of the Belgian textile market. As explained above, De Loecker does not have price data but introduces a demand system to separate productivity from price effects. Hence, one can distinguish in his framework between aggregate profitability ($\Pi_t$) and aggregate productivity ($\Omega_t$). Table 3 illustrates the standard decomposition of aggregate (i.e., industry-level) productivity changes. Columns 2–4 refer to revenue productivity, and columns 5–7 refer to physical productivity. Columns 2
and 5 show the aggregate changes at the industry level, columns 3 and 6 show the within-firm component, and columns 4 and 7 capture the reallocation.17

There are two interesting features of this table. First, the aggregate physical productivity change between 1994 and 2002, a period that spans the removal of major trade restrictions (i.e., quotas) in textiles, displayed in column 5 appears significantly lower than the change in revenue productivity in column 2. Hence, it seems that the usual productivity improvement shown in column 2 primarily reflects changes in prices. Second, column 4 suggests significant reallocation effects, from less profitable toward more profitable firms. The literature has often interpreted these effects as reallocation from less efficient toward more efficient firms. However, column 7 suggests that this interpretation is misguided: The reallocation effects computed using physical productivity are substantially smaller—almost nonexistent. Hence, it appears that the reallocation documented using revenue productivity as a measure of firm performance was reallocation toward higher price and higher markup firms, not toward firms with higher efficiency.

Because De Loecker (2011) works with a CES demand system, albeit one that allows for different demand elasticities across products within the textile industry, markups and prices at the product/firm level are not affected by the trade liberalization; the price and markup effects in his framework are the result of reallocation (across firms, or across products within a firm) toward firms/products with higher markups. However, we would expect trade liberalizations to also affect prices and markups at the firm/product level through both the residual demand (i.e., intensified

and 5 show the aggregate changes at the industry level, columns 3 and 6 show the within-firm component, and columns 4 and 7 capture the reallocation.17

There are two interesting features of this table. First, the aggregate physical productivity change between 1994 and 2002, a period that spans the removal of major trade restrictions (i.e., quotas) in textiles, displayed in column 5 appears significantly lower than the change in revenue productivity in column 2. Hence, it seems that the usual productivity improvement shown in column 2 primarily reflects changes in prices. Second, column 4 suggests significant reallocation effects, from less profitable toward more profitable firms. The literature has often interpreted these effects as reallocation from less efficient toward more efficient firms. However, column 7 suggests that this interpretation is misguided: The reallocation effects computed using physical productivity are substantially smaller—almost nonexistent. Hence, it appears that the reallocation documented using revenue productivity as a measure of firm performance was reallocation toward higher price and higher markup firms, not toward firms with higher efficiency.

Because De Loecker (2011) works with a CES demand system, albeit one that allows for different demand elasticities across products within the textile industry, markups and prices at the product/firm level are not affected by the trade liberalization; the price and markup effects in his framework are the result of reallocation (across firms, or across products within a firm) toward firms/products with higher markups. However, we would expect trade liberalizations to also affect prices and markups at the firm/product level through both the residual demand (i.e., intensified

17Levinsohn & Petrin (2012) criticize this market share–based measure of reallocation as being uninformative in welfare calculations. Although we are sympathetic to their criticism, we simply want to compare our physical productivity–based results to those one would obtain using the standard approach in the literature, and to this end, we adopt the standard practices of this literature throughout the calculations.
competition) and the cost (i.e., lower prices of imported intermediates) channels. To this end, one needs to consider more general demand systems that allow for variable markups and incomplete pass-through of cost changes to prices. Such a framework is considered in De Loecker et al. (2012). As noted above, these authors do not commit to a particular demand or market structure but adopt an empirical specification that nests the main models used in trade and industrial organization, including those that generate variable markups. They use the Indian trade liberalization to separately identify the effects of the residual demand (reductions in output tariffs) versus cost (reductions in input tariffs) channels. The authors find that output tariff declines have the expected procompetitive effects; they lead to lower prices and lower markups. But the striking result is that in the end, the net effect of the Indian liberalization was to increase markups. This increase, which at first seems at odds with the standard intuition that trade has procompetitive effects and hence reduces markups, does not come about because of firm collusion, or any other attenuation of competition. It is the result of the incomplete pass-through of input tariff declines to prices.

These results, which may seem surprising in the context of trade, are consistent with the findings of the exchange rate literature as well as the substantial macro literature on price rigidities. Overall, studies that have attempted to explicitly address the price and markup effects associated with trade openness suggest that the demand side of the market is as important as the cost side; trade liberalizations lead not only to (physical) productivity improvements, but also to changes in prices and markups that need to be modeled explicitly.18

5. CONCLUSIONS AND FUTURE WORK

We conclude with some final thoughts on the state of the literature and future work. There are several strands within the trade literature that deal broadly with firm performance and globalization, each employing different assumptions and approaches. Unfortunately, there has been minimal cross-fertilization of ideas across these literatures up to now.

The empirical productivity literature has focused on estimating the effects of trade on firm performance, without distinguishing between physical efficiency and price/markup effects. Mainstream theoretical models in the trade literature often employ assumptions that imply constant markups (e.g., CES preferences with monopolistic competition) and hence abstract from the potential of trade to affect markups. Models that do allow for markup effects have typically focused on the procompetitive effects of trade, paying little attention to the markup effects that arise as a result of incomplete pass-through of trade-induced cost reductions to prices. But this type of incomplete pass-through has been precisely the focus of the large literature on exchange rate pass-through, which tries to understand how prices and markups respond to exchange rate shocks. Its insights have never been applied to trade liberalizations, despite that (input) tariff reductions and exchange rates have similar effects on firm profits. Finally, there have been case studies of the effects of trade liberalization on firm performance in particular industries, for example, automobiles (see Goldberg 1995, Berry et al. 1999), that rest on estimation of structural industry models in the industrial organization tradition, but the results of these studies do not readily generalize to the economy at large.

We believe that the time is ripe for the methods and insights of these separate literatures to be combined in theoretical and empirical work in this area and hope that this article represents

---

18A similar point emphasizing the role of the demand side is made by Foster et al. (2008, 2012) but in the context of the domestic market.
a small step in this direction. Indeed, there are encouraging signs. Recent empirical work on the
effects of trade liberalizations has tried to distinguish between efficiency and markup effects,
yielding novel insights. Current work on the effects of trade on markups makes an explicit distinc-
tion between the competition and the pass-through channels. But much more research, on
different countries and different time periods, is needed before we will be able to draw general
conclusions. Furthermore, the past few years have seen the emergence of an exciting new literature
on assessing the aggregate gains from trade under alternative modeling assumptions. This liter-
ature that is primarily, although not exclusively, theoretical has emphasized the role that functional
form assumptions, especially ones with implications for markup adjustment, play in evaluating the
gains from trade. Careful empirical work that is motivated by and consistent with theoretical models,
but does not depend heavily on restrictive functional form assumptions, can play an important role
in informing the assumptions of the models used to assess the welfare gains from trade.

DISCLOSURE STATEMENT
The authors are not aware of any affiliations, memberships, funding, or financial holdings that
might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS
We would like to thank Tim Bresnahan for early conversations on the topic and especially Pol
Antràs for valuable comments. Channing Verbeck Jr provided excellent research assistance.

LITERATURE CITED
Ackerberg D, Benkard LC, Berry S, Pakes A. 2007. Econometric tools for analyzing market outcomes. In
In press
Amiti M, Konings J. 2007. Trade liberalization, intermediate inputs, and productivity: evidence from
Indonesia. Am. Econ. Rev. 97:1611–38
Arkolakis C, Costinot A, Donaldson D, Rodríguez-Clare A. 2012. The elusive pro-competitive effects of trade.
101:1312–44
127:437–92
100:70–97
policy. Am. Econ. Rev. 89:400–30
103:208–13
Bustos P. 2011. Trade liberalization, exports, and technology upgrading: evidence on the impact of
NBER Work. Pap. 18739
De Loecker J. 2011. Product differentiation, multiproduct firms, and estimating the impact of trade liber-
alization on productivity. Econometrica 79:1407–51


Contents

Probabilistic Expectations in Developing Countries
   Adeline Delavande ........................................ 1

Ill-Posed Inverse Problems in Economics
   Joel L. Horowitz .......................................... 21

Financing Old Age Dependency
   Shinichi Nishiyama and Kent Smetters .................. 53

Recent Developments in Empirical Likelihood and Related Methods
   Paulo M.D.C. Parente and Richard J. Smith ............ 77

Belief Elicitation in the Laboratory
   Andrew Schotter and Isabel Trevino ..................... 103

Models of Caring, or Acting as if One Cared, About the Welfare of Others
   Julio J. Rotemberg ..................................... 129

Exchange Rate Stabilization and Welfare
   Charles Engel ........................................... 155

Copulas in Econometrics
   Yanqin Fan and Andrew J. Patton ..................... 179

Firm Performance in a Global Market
   Jan De Loecker and Pinelopi Koujianou Goldberg .... 201

Applications of Random Set Theory in Econometrics
   Ilya Molchanov and Francesca Molinari ............... 229

Experimental and Quasi-Experimental Analysis of Peer Effects: Two Steps Forward?
   Bruce Sacerdote ........................................ 253

Coordination of Expectations: The Eductive Stability Viewpoint
   Gabriel Desgranges .................................... 273
From Sudden Stops to Fisherian Deflation: Quantitative Theory and Policy
Anton Korinek and Enrique G. Mendoza ........................... 299

China’s Great Convergence and Beyond
Kjetil Storesletten and Fabrizio Zilibotti .......................... 333

Precocious Albion: A New Interpretation of the British Industrial Revolution
Morgan Kelly, Joel Mokyr, and Cormac Ó Gráda .................... 363

Disclosure: Psychology Changes Everything
George Loewenstein, Cass R. Sunstein, and Russell Golman ........ 391

Expectations in Experiments
Florian Wagener .................................................. 421

Optimal Design of Funded Pension Schemes
Lans Bovenberg and Roel Mehlkopf .............................. 445

The Measurement of Household Consumption Expenditures
Martin Browning, Thomas F. Crossley, and Joachim Winter .... 475

Empirical Revealed Preference
Ian Crawford and Bram De Rock ................................ 503

Quality of Primary Care in Low-Income Countries: Facts and Economics
Jishnu Das and Jeffrey Hammer ................................. 525

The Endowment Effect
Keith M. Marzilli Ericson and Andreas Fuster ..................... 555

Decentralization in Developing Economies
Lucie Gadenne and Monica Singhal ............................. 581

Local Labor Markets and the Evolution of Inequality
Dan A. Black, Natalia Kolesnikova, and Lowell J. Taylor ....... 605

People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs
Patrick Kline and Enrico Moretti ................................ 629

Applying Insights from Behavioral Economics to Policy Design
Brigitte C. Madrian ............................................. 663

The Economics of Human Development and Social Mobility
James J. Heckman and Stefano Mosso ........................... 689

Firms, Misallocation, and Aggregate Productivity: A Review
Hugo A. Hopenhayn ............................................. 735

Endogenous Collateral Constraints and the Leverage Cycle
Ana Fostel and John Geanakoplos .............................. 771
Teacher Effects and Teacher-Related Policies  
C. Kirabo Jackson, Jonah E. Rockoff, and Douglas O. Staiger ........... 801

Social Learning in Economics  
Markus Mobius and Tanya Rosenblat ............................................ 827

Rethinking Reciprocity  
Ulrike Malmendier, Vera L. te Velde, and Roberto A. Weber ............ 849

Symposium: The Institutional Underpinnings of Long-Run Income Differences  
Institutions, Human Capital, and Development  
Daron Acemoglu, Francisco A. Gallego, and James A. Robinson ........ 875

Growth and the Smart State  
Philippe Aghion and Alexandra Roulet ........................................... 913

The Causes and Consequences of Development Clusters: State Capacity,  
Peace, and Income  
Timothy Besley and Torsten Persson ............................................. 927

Under the Thumb of History? Political Institutions and the Scope for Action  
Abhijit V. Banerjee and Esther Duflo ............................................ 951

Indexes  
Cumulative Index of Contributing Authors, Volumes 2–6 973  
Cumulative Index of Article Titles, Volumes 2–6 976

Errata  
An online log of corrections to Annual Review of Economics articles may be found at http://www.annualreviews.org/errata/economics
New From Annual Reviews:

**Annual Review of Statistics and Its Application**

Volume 1 • Online January 2014 • http://statistics.annualreviews.org

Editor: **Stephen E. Fienberg**, Carnegie Mellon University

Associate Editors: **Nancy Reid**, University of Toronto

**Stephen M. Stigler**, University of Chicago

The *Annual Review of Statistics and Its Application* aims to inform statisticians and quantitative methodologists, as well as all scientists and users of statistics about major methodological advances and the computational tools that allow for their implementation. It will include developments in the field of statistics, including theoretical statistical underpinnings of new methodology, as well as developments in specific application domains such as biostatistics and bioinformatics, economics, machine learning, psychology, sociology, and aspects of the physical sciences.

**Complimentary online access to the first volume will be available until January 2015.**

**TABLE OF CONTENTS:**

- A Systematic Statistical Approach to Evaluating Evidence from Observational Studies, David Madigan, Paul E. Stang, Jesse A. Berlin, Martijn Schuemie, J. Marc Overhage, Marc A. Suchard, Bill Dumouchel, Abraham G. Hartzema, Patrick B. Ryan
- The Role of Statistics in the Discovery of a Higgs Boson, David A. van Dyk
- Brain Imaging Analysis, F. DuBois Bowman
- Statistics and Climate, Peter Guttorp
- Climate Simulators and Climate Projections, Jonathan Rougier, Michael Goldstein
- Probabilistic Forecasting, Tilmann Gneiting, Matthias Katzfuss
- Bayesian Computational Tools, Christian P. Robert
- Bayesian Computation Via Markov Chain Monte Carlo, Radu V. Craiu, Jeffrey S. Rosenthal
- Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models, David M. Blei
- Structured Regularizers for High-Dimensional Problems: Statistical and Computational Issues, Martin J. Wainwright
- High-Dimensional Statistics with a View Toward Applications in Biology, Peter Bühlmann, Markus Kalisch, Lukas Meier
- Next-Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data, Kenneth Lange, Jeannette C. Papp, Janet S. Sinsheimer, Eric M. Sobel
- Breaking Bad: Two Decades of Life-Course Data Analysis in Criminology, Developmental Psychology, and Beyond, Elena A. Erosheva, Ross L. Matsueda, Donatello Telesca
- Event History Analysis, Niels Keiding
- Statistical Evaluation of Forensic DNA Profile Evidence, Christopher D. Steele, David J. Balding
- Using League Table Rankings in Public Policy Formation: Statistical Issues, Harvey Goldstein
- Statistical Ecology, Ruth King
- Estimating the Number of Species in Microbial Diversity Studies, John Bunge, Amy Willis, Fiona Walsh
- Dynamic Treatment Regimes, Bibhas Chakraborty, Susan A. Murphy
- Statistics and Related Topics in Single-Molecule Biophysics, Hong Qian, S.C. Kou
- Statistics and Quantitative Risk Management for Banking and Insurance, Paul Embrechts, Marius Hofert

Access this and all other Annual Reviews journals via your institution at www.annualreviews.org.