Board of Governors of the Federal Reserve System

International Finance Discussion Papers

ISSN 1073-2500 (Print)
ISSN 2767-4509 (Online)

Number 1417

August 2025

Why are Manufacturing Plants Smaller in Developing Countries? Theory and Evidence from India

Anil K. Jain; Siddharth Kothari

Please cite this paper as:
Jain, Anil K., and Siddharth Kothari (2025). "Why are Manufacturing Plants Smaller in Developing Countries? Theory and Evidence from India," International Finance Discussion Papers 1417. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/IFDP.2025.1417.

Why are Manufacturing Plants Smaller in Developing Countries? Theory and Evidence from India[*]

Anil K. Jain [†]
Board of Governors of
the Federal Reserve System

Siddharth Kothari
International Monetary
Fund

August 25, 2025

Abstract

Poorer countries (and poorer states within India) have a larger share of manufacturing employment in small plants. This paper presents empirical evidence and a theoretical model to show that this relationship is driven by greater demand for lower quality goods in poorer regions, which can be produced efficiently in small plants. First, using data for India, we show that richer households buy higher price goods and larger plants produce higher price products. Second, we develop a model that matches these facts. Finally, we find that our model explains about forty percent of the cross-state variation in the size distribution of manufacturing plants in India.

JEL codes: O11, O16, O17; **Keywords:** Firm size distribution; Product quality; India

[*] A previous version of this paper was circulated as "The Size Distribution of Manufacturing Plants and Development."

This paper benefited significantly from in-depth discussions with Manuel Amador, Nicholas Bloom, Pascaline Dupas, Robert Hall, Chad Jones, Pablo Kurlat, Kalina Manova, Monika Piazzesi, Martin Schneider, and Christopher Tonetti. Siddharth gratefully acknowledges support from the Leonard W. Ely and Shirley R. Ely Graduate Student Fund Fellowship (SIEPR), B.F. Haley and E.S. Shaw Fellowship for Economics (SIEPR), and the SEED Fellowship from the Stanford Institute for Innovation in Developing Economies. The findings and conclusions in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System, the views of any other person associated with the Federal Reserve System, or the International Monetary Fund.

[†] Corresponding author; email: anil.k.jain@frb.gov.

1. Introduction

The typical size distribution of manufacturing establishments in developing countries has a thick left tail compared to developed countries. For instance, as shown in figure (1) in India, six out of ten manufacturing workers are employed in establishments with fewer than five people—a share thirty times larger than in the United States (2005-2006).[1]

This size-income relationship also holds across Indian states—poorer states have a larger fraction of employment in the smallest of firms. Figure 2 plots the share of employment in establishments of size five or less in 2005-06 for different Indian states against the per-capita net domestic product (NDP) of the state relative to the poorest state, Bihar.[2] The richest Indian states have about four times the per-capita NDP of the poorest states. While the poorest states have almost 90 percent of their manufacturing workforce employed in establishments of size five or less, the richer states have only about 40 percent of their workforce working in small establishments.[3]

What explains this negative correlation between income levels and the share of employment in small establishments? A leading view, starting from De Soto [1989], blames size-dependent regulation—such as licensing rules or small-scale reservations—for the prevalence of small plants. Such policies can distort resource allocation, depress incomes, and, ultimately, keep establishments small.

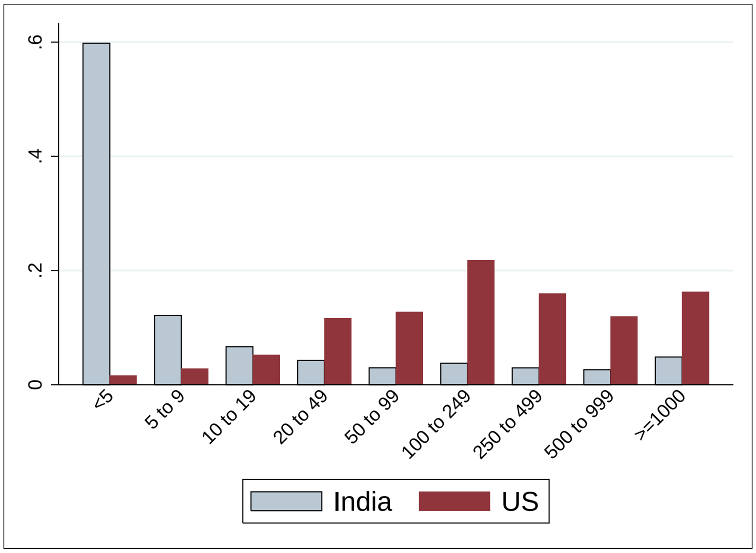
This paper explores an alternative (though potentially complementary) explanation for this size-income relationship that is driven by consumer preferences and technology rather than distortions.

[1] The US data is retrieved from the US Business County Patterns Database maintained by the US Census Bureau. The Indian data combines two surveys, the Annual Survey of Industries (ASI) and the Survey of Unorganized Manufacturing (SUM). Appendix Sections A.1, A.2, and A.5 give more details regarding these datasets.

[2] For ease of exposition Figure 2 plots only the 15 largest states that cover over 96 percent of the manufacturing workforce. The negative relationship between share of employment in small plants and per-capita state NDP is robust to including all the states. In a regression of share of employment in plants of size five or less on log of per-capita state NDP, the coefficient (standard error) on log state NDP is -0.320 (0.0553) when restricting to 15 states and -0.319 (0.0568) when including all the states. A possible concern with the relationship seen in Figure 2 is that it might be driven by differences in industry composition across states. However, a large part of the differences in share of employment in small plants across states is actually driven by within industry differences in size. Controlling for industry composition (weighting the size distribution in every state by the all India industry composition instead of the state specific industry composition) at the 2-digit level causes the slope coefficient on log of state per-capita NDP to fall from -0.320 (0.0553) to -0.274 (0.0417).

[3] The differences in share of employment in small plants also reflects in differences in average plant size across states. The average plant size in the richest states is about double the average in the poorest states.

Figure 1: Share of Employment by Size Category: India vs. US

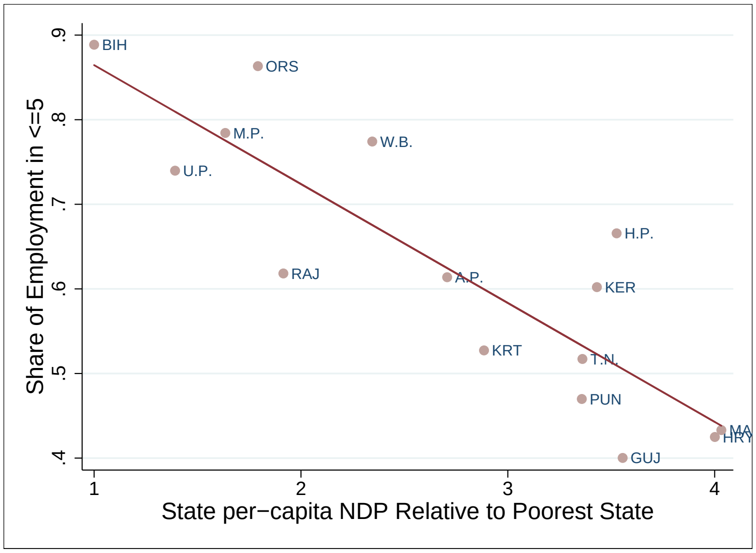


Notes: The graph plots the share of total employment in establishments of different size categories for India and the US. The data for India combines two sources, the Annual Survey of Industries (ASI) and the Survey of Unorganized Manufacturing (SUM) for 2005-06. The data for the US is taken from the County Business Patterns Database for 2006.

Our hypothesis is that poor households have high demand for low quality products, which can be produced efficiently in small establishments as they require small fixed investments (no research and development expenditure, or no need for large investments in fixed capital). On the other hand, richer households tend to demand higher quality goods, whose production requires a larger scale due to the need for larger fixed investments. In effect, we argue that higher quality goods require higher sunk and fixed costs, and consequently larger firms, consistent with the seminal industrial organization papers by John Sutton (such as Shaked and Sutton [1987] and Sutton [1991]).

This relationship between income levels and demand for quality implies that poor countries or states have demand skewed towards goods which require a small scale of production, which in turn causes the size distribution to be dominated by small plants. As a region develops and income levels increase, demand shifts towards high quality products, which in turn leads to a shift on the production side towards higher quality goods. This shift in production causes the share of employment in small plants to decrease, and thus can generate the negative relationship between the share of employment in small plants and income levels seen in the data.

Figure 2: Size Distribution of Manufacturing Establishments: Across Indian States



Notes: The graph plots the share of employment in plants of size five or less in a state against per-capita NDP of the state relative to the poorest state. The data for the states combines two sources, the Annual Survey of Industries (ASI) and the Survey of Unorganized Manufacturing (SUM). Only the 15 largest states are included to keep the graph readable.

We provide empirical evidence in support of this hypothesis using Indian data from consumer and producer surveys. First, consistent with the hypothesis that richer households buy higher quality products we show that in the cross-section richer households pay a higher unit price for similar goods (using data from consumer expenditure surveys). Second, we show substantial evidence that larger plants produce higher quality goods than smaller firms. We start by showing that larger plants charge a higher unit price for a similar good than smaller plants—a relationship that holds across formal and informal plants. Moreover, consistent with higher quality, we show that larger plants use more expensive material inputs, use more capital per unit of output, invest more in capital per unit of output, and hire more skilled workers than smaller firms making similar goods.

We develop a general equilibrium model that matches these cross-sectional facts. Households choose from a finite number of quality levels. The choice over quality levels is modeled as a discrete-choice problem with households choosing to consume one quality level out of those available in the economy. Their preferences exhibit non-homotheticity with respect to quality: richer households are more likely to choose higher quality levels. The non-homotheticity arises because the utility function features complementarity between quality and quantity consumed (the marginal increase in utility from a given increase in quantity consumed is larger for higher quality goods) and richer households can consume more quantity of whichever quality level they choose.

On the producer side, production of high quality goods uses skilled labor more intensively. Also, starting a higher quality plant requires higher fixed costs, which combined with a free entry condition implies that producers of high quality goods will be larger on average (in order to recover their larger fixed costs).

The model parameters are chosen to match the micro-facts documented on the consumer and producer side. The quality-size relationship on the producer side is matched to the relationship between prices and plant size from the producer surveys, while the degree of non-homotheticity is chosen to match the price-income relationship seen in the consumer surveys.

The empirical findings and our model raise the question: How much of the cross-state variation in the size distribution seen in Figure 2 can be explained by the model? We answer this question through a counterfactual exercise in which we simulate changes in per-capita income levels in the model (by varying productivity and the skill level of the population) and see what is the effect on the size distribution.

In our model, as income levels rise, demand shifts to high quality goods due to the non-homotheticity of preferences. This shift in demand towards higher quality leads to a shift on the production side, with a fall in the number of low quality producers and an increase in the number of high quality producers. As high quality producers are larger on average compared to low quality producers, there is also a shift in the size distribution towards larger plants. We find that the share of employment in plants of size five or less goes down by nearly 20 percentage points (which is about 43 percent of the difference seen across Indian states) when income in the model varies by the same extent as it does across Indian states. Further, we document that the share of employment in plants of size five or less has gone down by about 20 percentage points in India between 1989 and 2009, and show that the model can explain about 65 percent of this change. That said, the model abstracts from several factors that may impact the quantitative result. For example, we assume perfectly competitive final goods markets with zero economic profits, and monopolistically competitive intermediate producers with constant markups, thus not allowing for possible markup variation across quality levels. If consumer valuation of quality allows higher quality firms to charge larger markups, this could weaken the size-quality relation for a given fixed cost. We also do not allow for standard supply side frictions (for example, state-level entry barriers, size-dependent labor laws, and credit-market imperfections), which could potentially interact with firms' decisions on quality choice. As such, our quantitative exercises should be interpreted as upper-bound estimates of the impact of quality-driven demand.

As robustness, Section D shows evidence that inter-state trade is not driving our model results. Our model and the counterfactual exercises assume that each state can be treated as a closed economy in which local demand is met by local production. A potential confounding effect of inter-state trade could come through the location choice of large plants. For example, if the richer states are more suited for operating large plants (for example, due to availability of skilled labor or less stringent labor laws), then larger plants might choose to locate in these states (and ship their goods to the poor states) and this might be driving the negative relationship between income and firm size. If inter-state trade was an important force, then we would expect the more tradable industries within manufacturing to have a stronger negative relationship between size and income levels across states. To test this, we construct two measures of tradability at the 3-digit level of industrial classification. We find that the size-income relationship across states is not stronger for tradables as compared to non-tradables (for one of the measures, the non-tradables actually have a stronger negative relationship as compared to tradables) indicating that inter-state trade is unlikely to be an important force driving larger share of employment in small plants in poorer states.

Our paper contributes to several literatures.

A large literature has studied the question of why the size distribution differs markedly across countries. The role of distortionary policies and the regulatory environment in determining the size distribution of plants (and the extent of informality) has extensively been studied in Little et al. [1987], De Soto [1989], Loayza [1996], Djankov et al. [2002], Loayza et al. [2005], Loayza et al. [2009], Garicano et al. [2016], and Ulyssea [2018]. While size-dependent policies are potentially an important determinant of the size distribution, these policies are unlikely to explain all the differences in size distribution seen between developing and developed countries. Tybout [2000] notes that developing countries tend to have a large share of their population in small plants, irrespective of whether they have policies which discriminate against large plants or not. This finding suggests that these policies cannot be the only factor driving plant size.

Consistent with regulatory policies not fully explaining the puzzle of why there are so many small firms in developing countries, Gollin [1995] and Hsieh and Klenow [2014] conduct quantitative exercises in which they find that size-dependent policies leave a large part of the differences in size across countries unexplained. Moreover, Hsieh and Olken [2014] document that the "missing middle" in the size distribution in developing countries actually does not exist and that regulatory obstacles which become binding at particular threshold levels do not seem to lead to discontinuities in the size distribution in developing countries. [4]

Complementary to this argument, recent quantitative and theoretical work has emphasized that deeper structural factors—such as technological design, entrepreneurial skill distribution, and human capital may also account for cross-country differences in firm size. Bento and Restuccia [2017, forthcoming] show that models featuring technology that favors small-scale production, or frictions that prevent firm growth, can rationalize observed variation in both the firm size distribution and aggregate productivity. Poschke [2018] builds on this insight by highlighting that countries with lower skill availability and more rigid entrepreneurial technologies naturally exhibit a thinner right tail of the firm size distribution, even in the absence of distortive policies.

Our paper suggests that a large part of the differences in size distribution that we see across countries and states is a natural consequence of the low levels of income in developing countries and is not necessarily caused by policies which discriminate against large productive plants in favor of small unproductive plants. The hypothesis considered in the paper is closer to the dual-sector view of the informal sector in La Porta and Shleifer [2008] according to which the informal sector does not compete directly with the formal sector. For instance, Bloom and Van Reenen [2007] and Scur et al. [2024] document that smaller firms in developing countries tend to have significantly weaker management practices, and that this management gap is a key driver of lower productivity and scale.

A growing body of research describes obstacles to firm growth in low-income countries such as

[4] There is also a quantitative literature which looks at the role of distortionary policies in explaining cross-country differences in Total Factor Productivity. See Guner et al. [2008], Alfaro et al. [2009], García-Santana and Pijoan-Mas [2010], Barseghyan and DiCecio [2011], Hsieh and Klenow [2014], and Restuccia and Rogerson [2013].

managerial, technological, credit, and informational frictions. Akcigit et al. [2021] show that weak selection pressures and limited managerial delegation in developing economies hinder the growth and upgrading of high-potential firms. Bassi et al. [2023] highlight that many small firms operate informally and depend on self-employed workers with minimal access to capital or training, restricting their capacity to improve product quality. Human capital limitations further constrain innovation: Cox (2025) documents that deficits in tertiary education reduce firms' ability to adopt new technologies and upgrade product standards. Similarly, Cirera et al. [2022] attribute lagging adoption of quality-enhancing technologies to poor managerial practices, limited digital infrastructure, and information constraints. Evidence from a field experiment by Atkin et al. [2017] shows that targeted support can enable quality upgrading among small Egyptian firms seeking to enter export markets. Taken together, these studies suggest that small firms in low-income contexts face a web of supply-side constraints and limited market incentives that give rise to a comparative advantage in the production of lower-quality goods.

Our paper's results are similar to the evidence in Lagakos [2016] that provides cross-country evidence for retail trade showing that poor countries exhibit lower TFP because they rationally choice technology with low measured labor productivity due to high costs of transportation and poor household wealth. We focus on the heterogeneity of quality levels being produced by plants of different sizes and how the demand for low quality falls with development.[5]

Some of the empirical results documented here have been studied in different contexts. Deaton and Dupriez [2011] and Dikhanov [2010] document that richer Indian households buy higher price goods. However, these papers focus on spatial differences in prices within India and not the price income relationship itself and its implication for the size distribution. Bils and Klenow [2001] show that richer households in the US also buy higher priced durable products.

Our paper is related to Kugler and Verhoogen [2012] finding that larger plants produce higher price goods and use higher price inputs in Colombia. Similar, to our paper, they interpret these price differences as representing quality differences and develop a model in which more productive firms choose to produce higher quality goods at a higher unit cost. Our paper extends this result

[5] The idea of quality dualism between the formal and the informal sector has been looked at by Banerji and Jain [2007], who develop a partial equilibrium model in which formal sector establishments have a comparative advantage in producing higher quality goods due to differences in factor prices across the two sectors. However, their partial equilibrium model does not have implications for the size distribution of firms and its relationship to income levels.

for India and, by combining data from the formal and informal sector, show that the price size relationship also holds when we include very small plants in the sample (the Colombian data only has plants of size ten or more).[6]

A number of papers, especially related to international trade, have developed models of non-homothetic preferences with respect to quality. These include Flam and Helpman [1987], Mitra and Trindade [2005], Dalgin et al. [2008], and Choi et al. [2009]. The model we develop is most closely related to the model in Fajgelbaum et al. [2011]. Their model features non-homothetic preferences with respect to quality where the non-homotheticity arises due to complementarity between the homogenous good and quality. The non-homotheticity with respect to quality in our model arises due to complementarity between the quantity of the good consumed and quality.

The rest of the paper is structured as follows: Section 2 documents that richer households buy higher price goods and that larger plants produce higher price goods and use higher price inputs. Section 3 presents the model and Section 4 discusses the calibration. Section 5 presents the results for the counterfactual exercises and explores the sensitivity of the results to some key parameters. Section D considers the role of inter-state trade in explaining the cross-state relationship seen in Figure 2 and Section 6 concludes.

2. Empirical Results

In this section, we provide empirical evidence which is consistent with our hypothesis of richer households consuming higher quality products which are produced by larger plants. In particular we show the following facts:

1. Richer households buy higher price goods
2. On average, larger plants produce higher price goods
3. Larger plants use higher price material inputs and hire more skilled labor

[6] There is a large international trade literature which documents heterogeneity in prices either at the product or the firm level for exports and imports and interprets these price differences as quality differences. Some papers in this literature include Schott [2004], Hummels and Klenow [2005], Hallak [2006], Mandel [2010], Manova and Zhang [2012], Iacovone and Javorcik [2012], and Hallak and Sivadasan [2013].

The facts are documented using four Indian surveys. We give a brief description of each survey along with the main results in the sections that follow.

2.1. Households: Richer Households Buy Higher Price Goods

This sections shows that richer households buy higher price goods, which is consistent with them consuming higher quality products. We use data from the Consumer Expenditure Survey of 2004-05 conducted by the National Sample Survey Office (NSS) of India. About 125,000 households from all Indian states and union-territories were interviewed for the survey. The survey asks households to report the value of consumption for 339 different goods. Households report quantities and rupee values separately for 209 goods, which can be used to compute prices for these goods. More details about the survey can be found in Appendix A.3.

We run regressions of the form

$$\ln\left(P_{h,g}\right) = \alpha_{g,state,rural} + \beta \ln\left(c_h\right) + \varepsilon_{h,g},$$

where $P_{h,g}$ is the price paid by household h for good g, c_h is per-capita expenditure of the household excluding durables, and $\alpha_{g,state,rural}$ represents fixed effects for each product, state, and urban-rural cell. c_h is a proxy for the income level of the household, adjusting for household size.[7] $\alpha_{g,state,rural}$ controls for the fact that different goods have different average price levels and that these price levels can vary across rural and urban areas and across states. For example, real estate prices might differ across rural and urban areas or across states with different levels of per-capita income and this can drive differences in cost of living and all prices. The fixed effects ensure that the price-income relationship is not identified out of differences in average price levels across states of different income levels or across rural-urban area. Intuitively, the coefficient β is the elasticity of price with respect to per-capita consumption level and is identified out of variation in prices paid for the same good by households of different income levels within each state's urban or rural sector.

[7] Purchase of durables is excluded as these are lumpy, infrequent purchases. Two households with the same level of permanent income might have very different levels of durable expenditure in any particular year simply because of differences in timing of durable purchases.

Table 1: Household Regressions: Richer Households Buy Higher Price Goods

	(1) log(price)	(2) log(price)	(3) log(price)	(4) log(price)	(5) log(price)
log(per-capita expenditure)	0.17^{***}	0.12^{***}	0.11^{***}		
	(0.00071)	(0.00059)	(0.00056)		
log(per-capita expenditure): winsored				0.11^{***}	
				(0.00056)	
log(per-capita expenditure): exclude own-product					0.11^{***}
					(0.00056)
Adjusted R^2	0.967	0.969	0.976	0.976	0.976
Price Ratio (75th to 25th percentile)	1.14	1.1	1.09	1.09	1.09
Price Ratio (95th to 5th percentile)	1.39	1.25	1.23	1.24	1.23
Winsor	Yes	Yes	Yes	Yes	Yes
Observations	5348463	5348463	5348463	5348463	5348463
Block FE	N/A	N/A	N/A	N/A	N/A
Product FE	Yes	Yes	N/A	N/A	N/A
State x Rural FE	No	Yes	N/A	N/A	N/A
State x Rural x Product FE	No	No	Yes	Yes	Yes
Number of Products	188	188	188	188	188
SE clusters:	Household	Household	Household	Household	Household
Number of Clusters	124635	124635	124635	124635	124635

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The data is from the Consumer Expenditure Survey of 2004-05. This table examines whether richer households purchase more expensive goods controlling for different definitions of products and inclusion of fixed effects. Column 1 only includes product fixed effects. Column 2 includes product fixed effects and state interacted with rural fixed effects. Column 2 winsorizes 1 percent tails of per-capita expenditure and goods prices. Column 3 excludes the expenditure on the good itself from the independent variable. The price ratio implied by the coefficient estimates for different percentiles of per-capita expenditure are reported in the rows called "Price Ratio". Standard errors are clustered at the household level.

The results of Table 1 suggest that richer households pay more for the same product than other households. Columns 1-5 of Table 1 reports the estimate of β, the elasticity of price with respect to per-capita consumption, with slightly different specifications based on 188 goods.[8] The point estimate for β varies between 11 and 17 percent which implies that the average price paid by the 95th percentile household in terms of per-capita expenditure is 23 to 39 percent more than the

[8] Although prices can be computed for 209 goods, only 188 were included in the regression. The goods excluded were heavy durables and all goods with the word “other” mentioned in the description. The results do not change substantially if these goods are included.

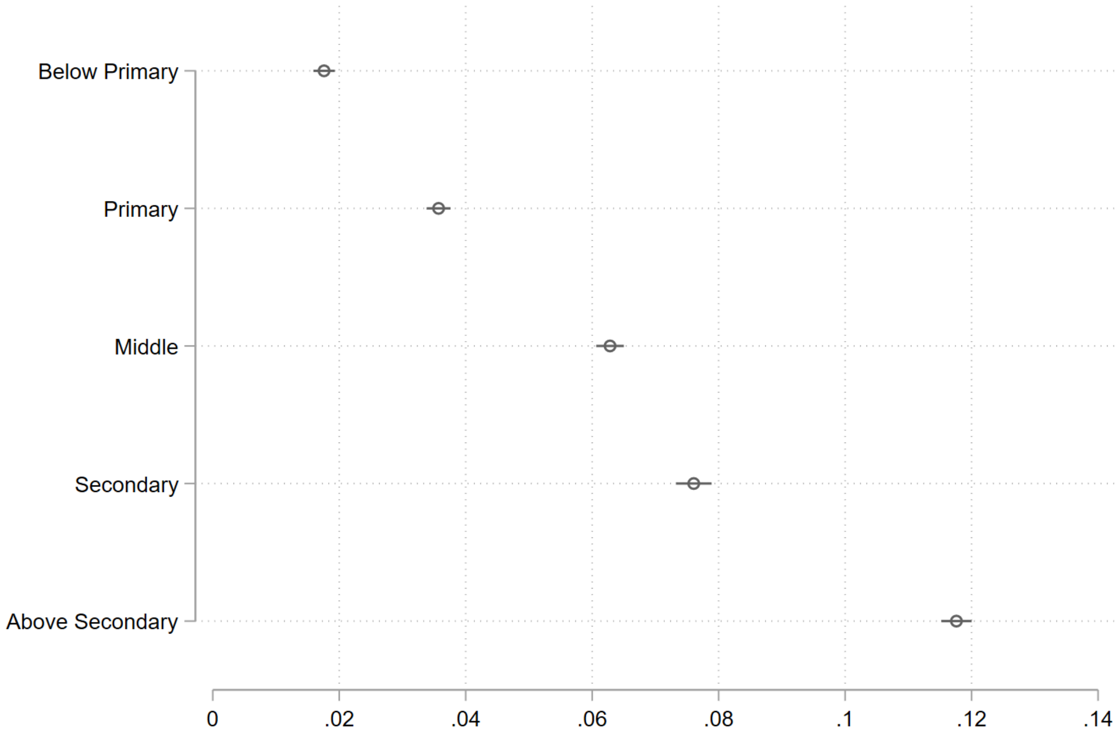
price paid by the 5th percentile household (in contrast, the 95th percentile household's per-capita expenditure is about seven times that of the 5th percentile household). Columns 1 through 3 use tighter fixed effects and Column 4 shows that winsorizing 1 percent tails for per-capita expenditure and prices (for a good within a state and urban-rural cell) does not materially change the results. Column 5 uses the households' expenditure after removing the consumer's expenditure on that product.

A possible concern with the results in table (1) is that the independent variable is itself a function of the dependent variable as per-capita expenditure sums the expenditure of the household across all goods, i.e., $c_h = \frac{\sum_g P_{h,g} Q_{h,g}}{\text{household size}}$ where $Q_{h,g}$ is the quantity consumed by household h of good g. This can give rise to a mechanical correlation and also cause a bias if the variables are measured with error. Therefore, to overcome this possible bias, we regress the price paid on a set of education dummies and other controls. In figure (4), from this regression, we plot the estimated coefficients (the circle) and the 95 percent confidence intervals (the line) for each of the education dummies with "illiterate" being the omitted education category. The key inference is that households with more education—which is likely strongly correlated with income—spend more on the same good than other households. Moreover, this result strongly increases for higher levels of education.

To provide further evidence that richer households pay more for the same product relative to other households, we compare households that differ in the number of dependents in the household. Specifically, we examine whether households with similar levels of expenditure but more dependents in the household spend less than other households. Our logic is that households with similar income but with more dependents will have less disposable income, therefore, in effect will be poorer. Therefore, we test whether households with more dependents will pay a lower price relative to other households. Our results in table (2) show that households with more dependents spend less per unit of product than other households after controlling for household income and other controls.

One final robustness concern is that richer households pay higher prices for the same good because they have higher opportunity cost of time and consequently search less for lower prices. To provide evidence that rules out this concern, we exploit data on whether there are adults that not working in the household—the underlying assumption is that these households should have a

Figure 3: The relationship between education and the price paid for each product



This figure plots the estimated coefficients from the regression of log(price) on a set of education dummies and additional controls. These controls are the same as those used in table 1 column 4 and is the triple interaction of state, rural, and product fixed effects. The omitted education category is "illiterate." The 95% confidence intervals are denoted by the lines. Standard errors are clustered at the household level.

low opportunity cost of time. We find that even after controlling for households with non-working adults, we still observe a strong relationship (and a similar magnitude) between household income and prices paid. More details on these regressions and results are in Appendix (C).

The results in this section has shown strong evidence that, on average, richer households pay more for the same product than other households, which is consistent with the hypothesis that they are consuming higher quality products. For robustness, we also investigate variation in the price elasticity relative to household income by product. Figure (4) plots the frequency histogram of the product-specific price elasticity to household expenditure, while including state interacted with

Table 2: Households with more dependents pay a lower price than other households with similar income for the same product

	(1) log(price)	(2) log(price)
log (household expenditure)	0.091^{***}	0.079^{***}
	(0.00053)	(0.00052)
Number of dependents	-0.016^{***}	
	(0.00018)	
Share of dependents in household		-0.067^{***}
		(0.0012)
Adjusted R^2	0.976	0.976
Winsor	Yes	Yes
Observations	5348463	5348463
State x Rural x Product FE	Yes	Yes
Number of Products	188	188
SE clusters:	Household	Household
Number of Clusters	124635	124635

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

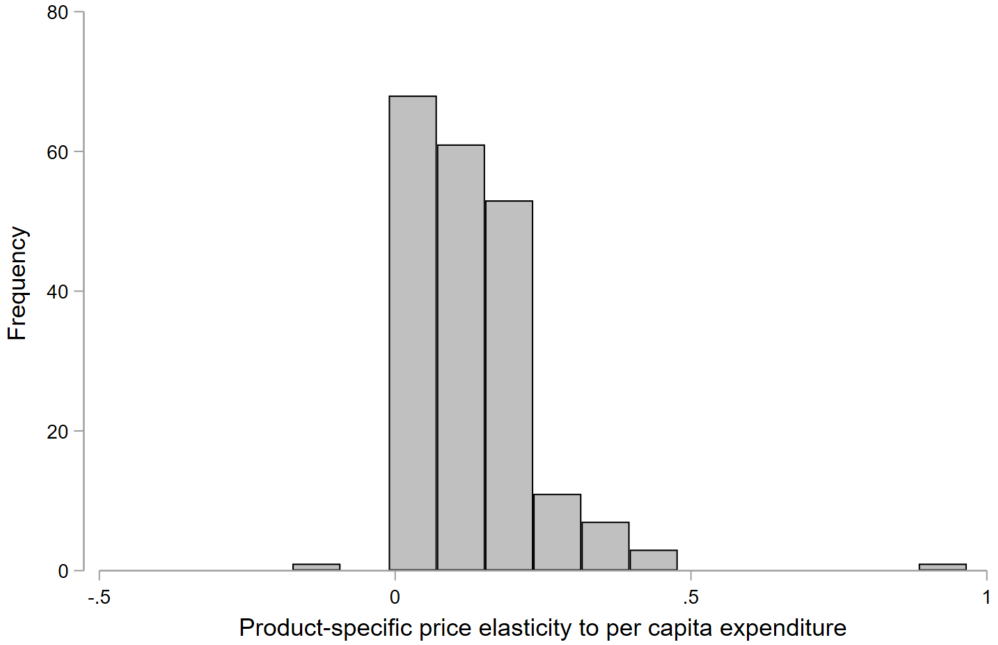
This table examines whether households with more dependents spend less than other households with similar income for the same product. We define a dependent as a person in the household who is younger than 16 and older than 70. In column 1 (2), we include the number (share) of dependents in the household as a regressor.

rural fixed effects ($\alpha_{state,rural}$). Specifically, we do following regressions:

$$log(price)_{h,g} = \alpha_{rural,state} + \beta_p log(\text{per capital expenditure})_p + \varepsilon_{h,g} \tag{1}$$

for each product and collect each β_p (so 188 regressions since there is 188 individual consumer products in our dataset). We then plot the frequency histogram for these β_p. These coefficients represent the estimated increase in price paid for each product as the household expenditure (which proxies household income) rises. The key interpretation of figure (4) is that for vast majority of products (more than 80 percent), the estimated elasticity is between 0 and 20 percent. This suggests that the result that households with higher income pay more for the same product is a general result and not driven by outlier products. Out of the 188 products in our regression, we only find 2 products where higher income households pay less on average for that product than other households.

Figure 4: Frequency Histogram of Product-Specific Price Elasticity to Expenditure



This figure plots the frequency distribution of the coefficient on "log(per-capital expenditure)" when regressing 'log(price)" on "'log(per-capital expenditure)" for each individual product while including rural interacted with state fixed effects. Specifically, we do the following regression log(price)$_{p,h} = \beta_p$log(per-capita expenditure)$_{p,h} + \alpha_{r,s} + \varepsilon_{p,h}$ for each product and collect each β_p (so 188 regressions since there is 188 individual consumer products in our dataset). We then plot the frequency histogram for these β_p.

2.2. Firms

Starting from Shaked and Sutton [1987], there's a long history of industrial organization papers arguing that markets with vertical product differentiation can exhibit firms of different scales due to the presence of fixed costs. Specifically, firms that produce higher quality goods will be larger and invest more in fixed costs. This section starts by showing strong evidence that larger firms produce higher quality goods—through showing that larger firms charge a higher price for the same product and other more direct measures of quality (such as international product certifications). Second, consistent with the higher quality goods requiring more expensive inputs and investment, we show three pieces of compelling evidence: larger firms have both a higher capital stock (per unit of output) and a higher capital investment flow (per unit of output); firms with a higher capital to labor ratio charge higher prices; larger firms hire more educated workers.

To build intuition for why larger firms produce higher quality products, it is helpful to consider a single product. Specifically consider firms that produce the product "finished cotton cloth." As described in substantial detail in Appendix (B), this industry exhibits large dispersion in firm size and final prices, with a strong positive correlation between firm size and prices. Consistent with our theory, we find the largest firm with public data in this industry uses expensive and high-end imported machinery, skilled labor, multiple product certifications, and sells at significantly higher prices than its competitors.

To show these results, we predominantly rely on combining data from the Annual Survey of Industries (ASI) of 2005-06 and the Survey of Unorganized Manufacturing (SUM) of 2005-06. The ASI covers all manufacturing plants registered under the Factories Act, 1948. This includes manufacturing plants employing twenty or more workers and not using electricity or employing ten or more workers and using electricity. The SUM on the other hand covers the smaller manufacturing plants not covered by the ASI. The two surveys together should provide a representative sample of the manufacturing sector as a whole. [9]

Both the surveys ask manufacturing establishments detailed questions about the products they

[9] A number of recent papers have combined these two surveys to construct a dataset which is representative of the manufacturing sector as a whole. These include Hasan and Jandoc [2010], Nataraj [2011], Hsieh and Klenow [2014], and Ghani et al. [2012].

produce and inputs they use. Each establishment reports the quantity of the product it produces (for a 5-digit product classification, which has about 5,500 possible products) and its value (before taxes and distribution expenses) which can be used to compute prices. For the ASI, each products quantity is supposed to be reported for a standardized unit (kilograms, numbers, etc). In the SUM, different plants can report the same products price in different units. We concord units across the two survey so that the price of the same product is not getting compared for different units.[10]

[10] In the ASI all plants reporting a certain product are supposed to report quantities in the same units. However, there are clear cases in which plants are misreporting quantity units. For example, all plant which produce milk are supposed to report quantities in terms of kiloliters which means that the price computed by dividing the rupee value by the quantity should yield prices per kiloliter. However, there is a group of plants whose prices are approximately 1000 times lower than others. This is clearly a case of some plants reporting quantities in liters instead of kiloliters. We have manually gone through all product categories and identified products with this problem and split these into two separate categories based on a sensible price cutoff. In addition to this manual check, we have also implemented an algorithm to identify these problem products and used the algorithm generated cutoff's to split problematic products. The results are similar to the ones reported here. Appendix F gives more details regarding this problem and how it is being tackled.

2.2.1. On average, larger plants produce higher quality goods goods

To show evidence that, on average, larger plants produce higher quality goods than other firms, we start by showing that larger plants charge higher prices relative to other firms. Second, consistent with Verhoogen [2008], we show that larger firms produce higher quality goods by using more direct measures of product quality (specifically, showing that larger firms are more likely to have international certifications and be an export than other firms producing the same good).

To show that, on average, larger plants produce higher quality goods than other firms, while controlling for the triple interaction of product, state, and urban fixed effects, we run regressions of the form:

$$\ln\left(P_{f,g}\right) = \alpha_g + \alpha_{state,rural} + \gamma \ln\left(L_f\right) + \varepsilon_{f,g},$$

where $P_{f,g}$ is the price charged by plant f for product g, L_f is the number of workers employed by plant f, α_g is a product fixed effect, and $\alpha_{state,rural}$ is a state times urban-rural fixed effect. Intuitively, the coefficient γ is the elasticity of the price of output produced with respect to plant size and it is identified out of variation in prices charged by plants of different sizes producing the same product (reported in the same units) and allowing for differences in average price levels across states and urban and rural areas.[11]

Column 1 of Table 3 reports results when the sample is restricted to the ASI only. The estimate for the elasticity of price with respect to size, γ, is 0.096 and is statistically significant at the 1 percent level. The point estimate implies that a plant which employs 500 people on average charges a price which is 55.6 percent more than a plant employing 5 workers.[12]

Column 2 report results when the sample is restricted to the SUM only. The point estimate for the coefficient γ (elasticity of price with respect to size) is still positive but smaller. This is not surprising as the variation in employment levels within the SUM is small with 95 percent of the plants employing 16 workers or less.

Column 3 reports results when the two surveys are combined. The estimate for the elasticity of

[11] Note that the definition of a product differs between the consumer and firm datasets; therefore, even though we include product fixed effects in both sets of regressions, they are fixed effects for a different set of products.

[12] Note that the formal plants surveyed in the ASI report the value of output before taxes and distribution costs. Therefore, the price-size relation documented here is not driven mechanically by the fact that larger plants might be paying taxes while the smaller plants are not.

Table 3: Plant Regressions: Larger Plants Produce Higher Price Goods

	(1) log(output price)	(2) log(output price)	(3) log(output price)
log(labor)	0.10***	0.055***	0.11***
	(0.010)	(0.018)	(0.014)
Adjusted R^2	0.883	0.919	0.883
Price Ratio (Size 50 to 5)	1.26	1.14	1.28
Price Ratio (Size 500 to 5)	1.60	1.29	1.64
Sample	ASI	SUM	Both
Winsor	Yes	Yes	Yes
Observations	46704	28457	75161
State x Rural x Product FE	Yes	Yes	Yes
Number of Products	1218	2740	3182
SE clusters:	Product	Product	Product

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The data is from the ASI and SUM for 2005-06. All columns report results for regressions of log price charged by plants for their products on log of number of employees hired by the plant. Column 1 restricts the sample to the ASI, Column 2 restricts the sample to the SUM, while column 3 combines the two. One percent tails of prices (within a product) and plant size are winsorized. Regressions include the triple interaction of product, state, and rural fixed effects. Standard errors are clustered at the product level. The price ratio for different sized plants implied by the coefficient estimates are reported in the rows called "Price Ratio".

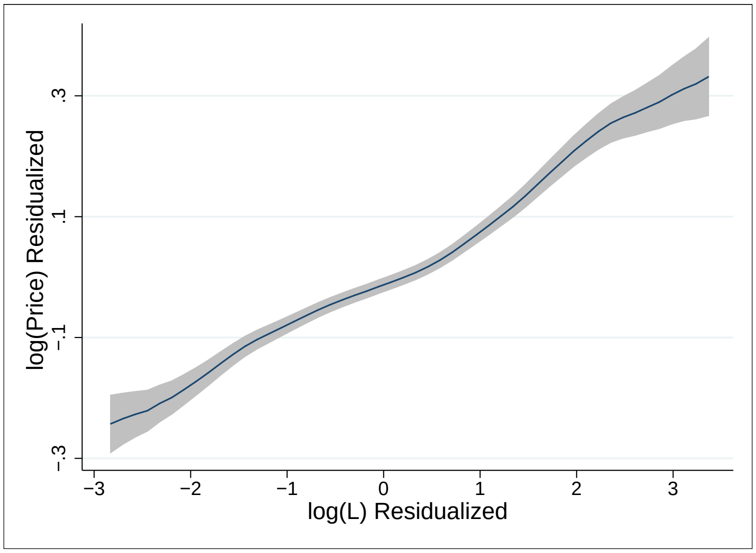
price with respect to size implies that a plant which employs 500 people on average charges a price which is 62.9 percent more than a plant employing 5 workers.

Figure 5 plots the non-parametric equivalent of the the regression in column 3 of Table 3. In particular, it estimates a kernel-smoothed local linear regression of residualized log prices (after removing product fixed effects and state times urban-rural fixed effects) on residualized log of plant size.[13] Again, the non-parametric estimates suggest that the price size relation across plants is close to log-linear.

The fact that larger plants produce goods which they sell at a higher price is consistent with the

[13] Log price and log of employment of each plant is regressed on product and state times urban rural fixed effects. The residuals from this procedure are used to run a kernel-smoothed local linear regression with an Epanechnikov kernel and a bandwidth of 0.502. The top and bottom 1 percent of residualized log of employment are excluded.

Figure 5: Non-parametric Estimate: Larger Plants Produce Higher Price Goods



Notes: The data is from the ASI and the SUM of 2005-06. The graph plots the kernel-smoothed local linear regression of residualized log prices charged by a plant for its products on residualized log employment of that plant (removes product fixed effects and the interaction of state and urban-rural fixed effects). Products which have the units problem discussed in footnote 10 and in Appendix F are split into two product categories. 1 percent tails of residualized log employment are excluded. An Epanechnikov kernel with a bandwidth of 0.502 used. The grey regions is the 95 percent confidence interval for the non-parametric estimate.

hypothesis that larger plants produce higher quality products.

To provide more direct evidence that larger firms produce higher quality products, we supplement our analysis by examining two more direct measure of quality: whether a firm has an international certification and whether it is an exporter. Both measures have been used as measures of quality in the literature (Verhoogen [2008, 2023]).[14] To show this evidence, we use a more recent Indian manufacturing dataset (ASI 2009-2010) because this dataset contains information on International Organization for Standardization certification (commonly referred to as ISO certification), a datapoint that is not available in earlier surveys. Specifically, this survey asks whether the firm has ISO 14000 series certification, which focuses on environmental practices. Consistent with higher prices being higher quality, table (4) column 1 (column 2) shows that firms with ISO certification (firms that are exporters) charge, on average, 12 percent (15 percent) more than other firms even after controlling for the triple interaction of state, rural, and product fixed effects.

Table 4: Firms with ISO certification or firms that export charge higher prices for their goods

	(1) log(price)	(2) log(price)
ISO certified	0.12^{***}	
	(0.038)	
Exporter		0.15^{***}
		(0.045)
Observations	42507	42549
Adjusted R^2	0.764	0.764
State x Rural x Product FE	Yes	Yes
SE clusters:	Product	Product

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table examines whether firms with ISO certification or firms that export produce higher price goods. Column 1 (2) regresses ISO certification dummy (export status dummy) on the natural logarithm of price charged while including the triple interaction of state, rural and product fixed effects. Since ISO certification is not available in the 2005-2006 ASI survey, we create this table using the 2009-2010 ASI survey. This survey includes a question on whether the firm has the ISO 14000 series certification, which focuses on environmental practices.

[14] Many other papers have theoretically and empirically argued that firms that export produce higher quality products such as Kugler and Verhoogen [2012], Manova and Zhang [2012], Iacovone and Javorcik [2012], Hallak and Sivadasan [2013].

To supplement this evidence, we also show that larger firms are more likely to be ISO certified and be an exporter (table 5). Specifically, this table shows that firms with more employees are more likely to be ISO certified and be an exporter than other firms that produce the same product, even after controlling for the triple interaction of state, rural, and product fixed effects.

Table 5: Larger plants are more likely to be ISO certified and be an exporter

	(1) ISO certified	(2) Exporter
log(Labor)	0.065***	0.045***
	(0.0023)	(0.0017)
Observations	93895	93984
Adjusted R^2	0.214	0.229
State x Rural x Product FE	Yes	Yes
SE clusters:	Product	Product

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table examines whether larger plants are more likely to be ISO certified and an exporter. Column 1 (2) regresses firm size—proxied by the natural logarithm of the number of employees—on an ISO certification dummy (export status dummy) while including the triple interaction of state, rural and product fixed effects. Since ISO certification is not available in the 2005-2006 ASI survey, we create this table using the 2009-2010 ASI survey. This survey includes a question on whether the firm has the ISO 14000 series certification, which focuses on environmental practices.

2.2.2. Larger firms invest more and use higher price inputs

To provide more evidence that larger firms produce higher quality outputs, we examine the investment and inputs use. First we show that larger plants pay a higher price for the same material input as compared to smaller plants. This is consistent with the idea that larger plants produce higher quality products which require higher quality inputs. We then show that larger plants hire more educated workers as compared to small plants.

Table 6: Plant Regressions: Larger Plants Use Higher Price Inputs

	(1) Log(input price)	(2) Log(input price)	(3) Log(input price)
Log(labor)	0.065***	0.042**	0.053***
	(0.0065)	(0.017)	(0.010)
Adjusted R^2	0.893	0.929	0.902
Price Ratio (Size 50 to 5)	1.16	1.1	1.13
Price Ratio (Size 500 to 5)	1.35	1.21	1.28
Sample	ASI	SUM	Both
Winsor	Yes	Yes	Yes
Observations	107325	105422	212747
State x Rural x Product FE	Yes	Yes	Yes
Number of Products	1218	2740	3182
SE clusters:	Product	Product	Product

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The data is from the ASI and SUM for 2005-06. All columns report results for regressions of log of price paid by establishments for material inputs used on log of number of employees hired by the establishment. Column 1 restricts the sample to the ASI only. Column 2 restricts the sample to the SUM only while column 3 combines the ASI and the SUM. One percent tails of prices (within a product) and plant size are winsorized. Regressions include the triple interaction of product, state, and rural fixed effects. Standard errors are clustered at the product level. The price ratio for different sized plants implied by the coefficient estimates are reported in the rows called "Price Ratio".

We use the ASI and SUM are used to show that larger plants use higher price material inputs. Each establishment reports the material inputs it uses (for a 5-digit product classification, which has about 5,500 possible products) and the price it pays for the input. The units between the surveys are again concorded.[15]

We run a regression of the form

$$\ln\left(P_{f,i}\right) = \alpha_i + \alpha_{state,rural} + \gamma \ln\left(L_f\right) + \varepsilon_{f,i},$$

where $P_{f,i}$ is the price paid by plant f for input i, L_f is the number of workers employed by plant f, α_i is a product fixed effect, and $\alpha_{state,rural}$ is a state times urban-rural fixed effect. Intuitively, the coefficient γ is the elasticity of the price paid for inputs with respect to plant size and it is identified out of variation in prices paid by plants of different sizes for the same inputs (reported in the same units), controlling for differences in average prices across states and urban-rural sectors.

Column 1 of Table 6 reports results when the sample is restricted to the ASI only. The estimate for the elasticity of input prices with respect to plant size, γ, is 0.077 and is statistically significant at the 1 percent level. The point estimate implies that a plant which employs 500 people on average pays prices for inputs which are 42.6 percent more than a plant employing 5 workers. Column 2 reports results when the sample is restricted to the SUM only. The coefficient γ is positive but smaller.

Column 3 reports results when the two surveys are combined. When combining the two surveys, the estimate for the elasticity of input prices with respect to size implies that a plant which employs 500 people on average pays a price for inputs which is 25.9 percent more than a plant employing 5 workers.

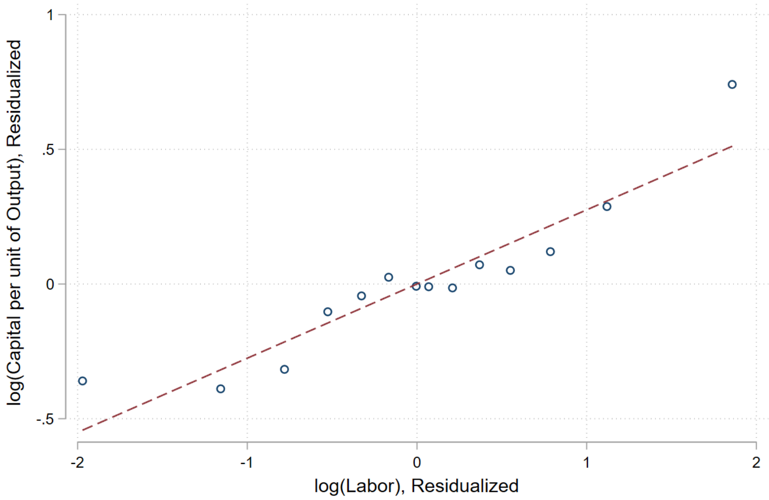
Not only do larger plants use higher price inputs, they also use more capital and invest more. Specifically, they have both a higher capital stock to output ratio and a higher capital investment flow to output ratio.

Figure (6) shows the strong relationship between firm size and firm capital using a binned scatterplot. A binned scatterplot is a convenient way of visualizing relationships when working with large datasets. Specifically, we plot the residuals for the natural logarithm of employees in a firm (x-axis) and for the natural logarithm of amount of capital in the firm relative to the firm's output (y-axis) after controlling for the triple interaction of product, state, and rural fixed effects. The

[15] The same problem of unit misreporting in the ASI discussed in footnote 10 is also present for inputs. We perform the same correction for this problem as we did in the previous section. The data appendix provides more details.

key inference from this figure is that larger firms use relatively more capital, even after controlling for the fact that larger firms produce more output. To reinforce this result, in table (7), we show that this result also holds for the level of firm investment (column 3) and different functional forms (columns 1 and 2). In all these regressions we include the triple interaction of state, rural, and product fixed effects. The results in columns 2 and 4 in table (7) suggests that a firm that employs 10 percent more people would have a capital stock to output ratio that is roughly 3 percent higher and capital investment to output ratio that is roughly 4 percent higher.

Figure 6: Relationship between firm size and capital



This figure shows the binned scatterplot and line of best fit for the relationship between the size of the firm and the amount of capital used in the firm. Specifically, we plot the residuals for the natural logarithm of employees in a firm (x-axis) and for the natural logarithm of amount of capital in the firm relative to the firm's output (y-axis) after controlling for the triple interaction of product, state, and rural fixed effects. To be precise, a binned scatterplot is a non-parametric method of plotting the conditional expectation function (which describes the average y-value for each x-value).

Consistent with large firms making higher quality goods, we find that firms with a higher capital to labor ratio charge higher prices for their goods and they use more expensive inputs. Specifically, in figure (7) we plot the residuals for the natural logarithm of the capital to labor ratio (x-axis) and for the natural logarithm of the product price (y-axis) after controlling for the triple interaction of product, state, and rural fixed effects. The key inference from this figure is that firms that use relatively more capital to labor charge higher prices for their product than other firms.

Table 7: Relationship between firm size, capital, and investment

	(1) Capital/Output	(2) log (Capital/Output)	(3) Investment/Output	(4) log (Investment/Output)
log(labor)	322.1***	0.28***	83.6***	0.38***
	(65.5)	(0.028)	(15.6)	(0.032)
Adjusted R^2	0.576	0.701	0.547	0.655
Winsor	Yes	No	Yes	No
Observations	23155	23155	23155	21604
State x Rural x Product FE	Yes	Yes	Yes	Yes
SE clusters:	Product	Product	Product	Product
Number of Clusters	945	945	945	915
Sample	ASI	ASI	ASI	ASI

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table examines whether larger plants (measured by number of employees) have a larger stock of capital and a higher level of investment relative to final output. Column 1 (2) regresses the stock of total capital (natural logarithm of total capital) relative to final output on the natural logarithm of employees. Column 2 (4) regresses total investment (natural logarithm of total investment) relative to final output on the natural logarithm of employees. To prevent our results being distorted by outliers, the ratios in columns 1 and 3 have been winsorized at the 5 percent level. Each regression includes the triple interaction of state, rural, and input product fixed effects. We only include firms that report capital levels. We cluster our standard errors for each product.

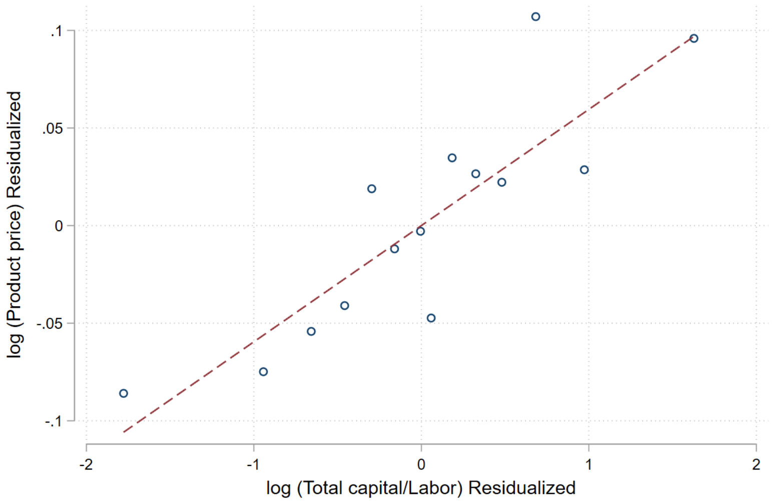
To further explore why firms with more capital charge higher prices, we examine different forms of capital. In table (8), we examine the relative stock of machinery capital and total capital (columns 1 and 3), and the relative level of machinery investment and total investment to labor on product prices. Across these regressions we find that firms with higher relative capital stock and capital investment to labor charge higher prices. The only regression coefficient that is not statistically significant is machinery investment to labor, but even that coefficient is close to significant at the 10 percent level and shows results consistent with the other regressions.[16]

Larger firms also employ more skilled labor. To show this we use the Employment-Unemployment Survey of 2004-05 conducted by the National Sample Survey Office (NSS) of India.[17] The

[16] We also find that firms that have a higher capital to labor ratio, also use more expensive inputs. Specifically, in table (16) in Appendix (H), we do regressions analogous to those in table (8) but use a firm's input prices rather than output prices. The results in table (16) are consistent with firms that use more capital require higher quality inputs. Therefore, this result contradicts the potential concern that firms with more capital are using this capital to reduce their input costs rather than producing higher quality outputs.

[17] We use the NSS because plants in the ASI and SUM do not report the education level of their workers.

Figure 7: Relationship between capital and prices



This figure shows the binned scatterplot and line of best fit for the relationship firm capital and the price of a product. Specifically, we plot the residuals for the natural logarithm of total capital to employee ratio (x-axis) and for the natural logarithm of price charged (y-axis) after controlling for the triple interaction of product, state, and rural fixed effects. To be precise, a binned scatterplot is a non-parametric method of plotting the conditional expectation function (which describes the average y-value for each x-value).

Employment-Unemployment Survey records demographic information (including education levels) for about 600,000 individuals. It also asks individuals to report the size of the establishment in which they work, with five permissible values: less than six workers; between six and nine workers, between ten and nineteen workers, twenty or greater workers, and unknown size. Table 9 reports the skill composition of workers for the different size categories. Out of the workers in establishments of size less than six workers, 43 percent have never attended school while only 3 percent have graduated from high school. On the other hand, out of workers in establishments of size more than 20 workers, only 23 percent have never attended school while 22 percent percent have graduated high school. As can be seen, a larger share of workers in big establishments have high levels of education.

Table 8: Plants with more capital charge higher prices

	(1) log(price)	(2) log(price)	(3) log(price)	(4) log(price)
Machinery Capital/ Labor ratio	0.024^{***} (0.0063)			
Machinery Investment/ Labor ratio		0.031 (0.024)		
Total Capital/ Labor ratio			0.018^{***} (0.0057)	
Total Capital Investment/ Labor ratio				0.027^{*} (0.014)
Adjusted R^2	0.878	0.878	0.881	0.881
Winsor	Yes	Yes	Yes	Yes
Observations	45303	45303	46143	46143
State x Rural x Product FE	Yes	Yes	Yes	Yes
Number of Products	1218	1218	1218	1218
SE clusters:	Product	Product	Product	Product
Number of Clusters	1178	1178	1178	1178
Sample	ASI	ASI	ASI	ASI

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table examines whether plants with a higher capital to labor ratio use charge higher prices. Column 1 (2) regresses the stock of machinery capital (total capital) to labor on a product's price. Column 2 (4) regresses machinery investment (total investment) to labor on input prices. To prevent our results being distorted by outliers, the prices are winsorized at the 1 percent level. Each regression includes the triple interaction of state, rural, and input product fixed effects. We cluster our standard errors for each input product.

Table 9: Larger Plants Hire More Educated Workers

	No School	Grade 1 to 9	Grade 10 to 12	> Grade 12
$L <= 5$	0.43	0.41	0.13	0.03
$5 < L <= 10$	0.34	0.41	0.17	0.08
$10 < L <= 20$	0.33	0.41	0.16	0.10
$L > 20$	0.23	0.32	0.22	0.22

Notes: The data is from the Employment-Unemployment Survey of 2004-05. The rows of the table represent the size category of the establishment in which an individual works while the columns represent the education level. Each number represents the share of individuals in the given size category who have attained the level of education given by the column.

2.2.3. Substantial variation in the elasticity of product price to firm size

This section shows there is substantial variation in the elasticity of product price to firm size. Specifically even though, on average, larger firms charge higher price–there's many products with the opposite relationship.

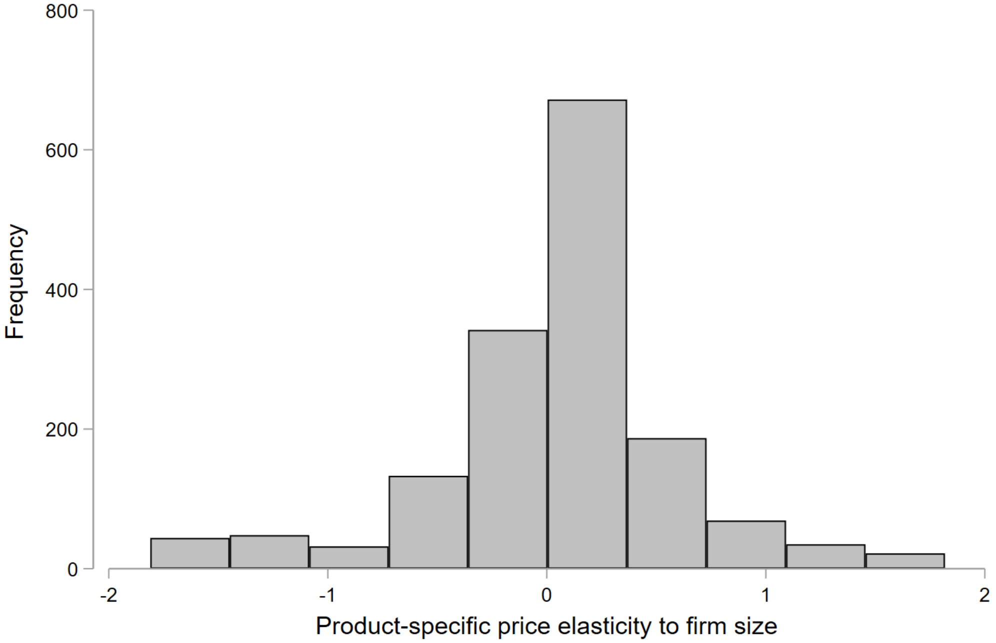
To examine variation in the price elasticity relative to firm size income, we start by plotting the frequency histogram of the product-specific price elasticity to firm size (proxied by number of employees), while including state interacted with rural fixed effects ($\alpha_{state,rural}$). Specifically, we do the following regressions:

$$\ln\left(P_{f,g}\right) = \alpha_{state,rural} + \gamma_g \ln\left(L_f\right) + \varepsilon_{f,g},$$

for each product and collect each γ_g (so 1217 regressions since there is 1217 individual products in the ASI dataset), where the f and g, refer to firm and good, respectively. We then plot the frequency histogram for these γ_g. These coefficients represent the estimated increase in price charged for each good as the firm size increases. Figure (8) shows two key results. First, there's substantial variation across products. Second, even though most products exhibit a positive elasticity (nearly two-thirds of products) and the median good having a positive elasticity of nearly 10 percent, there remains a sizable minority of products with a negative elasticity. In effect, this figure corroborates the anecdote that you may expect *some* industries to be characterized by smaller firms producing higher quality outputs (for example, tailored suits compared to mass-market men's formal wear), but, on average, we find that larger firms produce higher quality goods.

To complement this analysis, we can examine variation in prices charged within product and across similar products. Specifically, we can estimate the differences in prices using a less narrow definition for a "product." Table (10) examines the effect of using less narrow fixed effects for estimating the price elasticity; therefore, examining if larger firms produce slightly different and more expensive goods. Using the ASI dataset, for convenience, column 1 repeats the regressions Table (3 that uses the triple interaction of product, state, and rural fixed effects. In columns 2 and 3, we use increasingly broader product categories. Specifically, column 2 (3) interacts the state x rural fixed effect with a fixed effect for the 4-digit (2-digit) NIC code. Moreover, since

Figure 8: Frequency Histogram of Product-Specific Price Elasticity to Firm Size



Using the ASI dataset, this figure plots the frequency distribution of the coefficient on "log(number of employees)" when regressing 'log(price)" on "'log(number of employees)" for each individual product while including rural interacted with state fixed effects. Specifically, we do the following regression log(price)$_{f,g}$ $= \gamma_p$log(number of employees)$_f + \alpha_{r,s} + \varepsilon_{f,g}$ for each product and collect each γ_g (so 1217 regressions since there is 1217 individual products in our ASI dataset), where f, g, r and s subscripts refer to firm, good, rural and state, respectively. We then plot the frequency histogram for these γ_g. To ensure the figure is not distorted by outliers, we omit the extreme 2 percents of the distribution.

the units for different products within the same 4-digit NIC code may differ (for example, tons versus kilogram), we also interact the NIC code with a fixed effect for the unit of measurement. The key result from Table (10) is that using a broader product definition causes a slightly larger estimated effect. That is, on average, larger firms are more likely to produce similar—but slightly different—more expensive products.

Table 10: Larger plants produce higher price goods: Robustness to broader product categories

	(1) log(output price)	(2) log(output price)	(3) log(output price)
log(labor)	0.10***	0.16***	0.13***
	(0.010)	(0.015)	(0.031)
Adjusted R^2	0.883	0.698	0.634
Observations	46704	46704	46704
Sample	ASI	ASI	ASI
State x Rural x Product FE	Yes	No	No
State x Rural x 4-digit NIC x Unit FE	N/A	Yes	No
State x Rural x 2-digit NIC x Unit FE	N/A	N/A	Yes
SE clusters:	Product	Product	Product

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table examines whether larger plants produce higher price goods controlling for different degrees of product categories. Column 1 repeats the main regression from the paper. Column 2 and column 3 use increasingly broader categories than column 1. Specifically, column 2 (3) interacts the State x Rural fixed effect with a fixed effect for the 4-digit (2-digit) NIC code. Moreover, since the units for different products within the same 4-digit NIC code may differ (for example, tonnes versus kilogram), we also interact the NIC code with a fixed effect for the unit of measurement.

3. Model

This section develops a general equilibrium model which matches the facts described in Section 2. In particular, we model consumers choice between different quality levels with richer households more likely to buy high quality goods. On the production side, we assume that production of better quality requires larger fixed costs which along with free entry implies that high quality producers are larger on average.

3.1. Households

There are a mass L of households in the economy indexed by the subscript j. Share h of the households are skilled and earn wage w_S (determined endogenously in equilibrium) while share $1-h$ are unskilled and earn wage w_U. Unskilled wage w_U is assumed to be the numeraire and is normalized to 1.[18]

There are N quality levels. $Q = \{q_1, q_2, ..., q_N\}$ denotes the the set of qualities available in the economy. The quality indexes, q_n, are arranged in ascending order of quality. Therefore q_1 is the lowest quality level and q_N is the highest quality level.

The utility derived by household j from consuming quality level q_n is given by

$$u_{j,q_n}\left(c_{j,q_n}, \varepsilon_{j,q_n}\right) = a_{q_n} + q_n \log\left(c_{j,q_n}\right) + \varepsilon_{j,q_n} \quad \forall\, q_n \in Q, \tag{2}$$

where a_{q_n} is a constant in the utility function which can vary by quality level, c_{j,q_n} is the quantity consumed of quality level q_n by household j, and ε_{j,q_n} is a random utility component which represents the idiosyncratic valuation of quality level q_n by household j. The fact that higher quality levels have higher indexes, q_n, ensures that for a given level of quantity consumed, households get more utility from consuming higher quality goods.

The random utility component ε_{j,q_n} is assumed to be independently and identically distributed with a Gumbel Type 1 Extreme Value distribution with density

$$f\left(\varepsilon_{j,q_n}\right) = e^{-\varepsilon_{j,qn}} e^{e^{-\varepsilon_{j,qn}}}.$$

As shown by McFadden [1974] (see also Chapter 3 of Train [2009]), assuming a Gumbel distribution for the random utility component implies simple closed form expressions for demands.

We assume that a household can choose to consume only one quality level and hence will spend their entire income on that quality level. Therefore, if the household j chooses to consume quality level q_n, the indirect utility function can be written as:

$$v_{j,q_n}\left(w_j, P_{q_n}, \varepsilon_{j,q_n}\right) = a_{q_n} + q_n \log\left(\frac{w_j}{P_{q_n}}\right) + \varepsilon_{j,q_n}\, \forall\, q_n \in Q, \tag{3}$$

where we have substituted the household's wage (w_j) divided by the price of quality level q_n (P_{q_n}) for the household's consumption of good j in equation (2).

[18] Having two skill levels with different wages is crucial for our exercise as it generates cross-sectional differences in income levels in the model. This cross-sectional variation in income levels allows us to calibrate the extent of non-homotheticity in the model to match the price-income slope documented in Section 2.1.

Each household j receives draws of the random utility component ε_{j,q_n} for each quality level q_n and given these draws, chooses to consume the quality level which gives it the highest utility level. Therefore, household j chooses to consume quality level q_n if and only if

$$v_{j,q_n}\left(w_j, P_{q_n}, \varepsilon_{j,q_n}\right) > v_{j,q_m}\left(w_j, P_{q_m}, \varepsilon_{j,q_m}\right) \quad \forall n \neq m.$$

Let $\rho\left(q_n|w\right)$ be the share of households with wage w who choose to consume quality level q_n. Given the assumption that ε_{j,q_n} is independently and identically distributed with a Gumbel distribution, this share takes the simple logit form

$$\begin{aligned} \rho\left(q_n|w\right) &= \frac{e^{a_{q_n}+q_n \log\left(\frac{w}{P_{q_n}}\right)}}{\sum_{i=1}^{N} e^{a_{q_i}+q_i \log\left(\frac{w}{P_{q_i}}\right)}} \quad \forall\, q_n \in Q \\ &= \frac{e^{a_{q_n}}\left(\frac{w}{P_{q_n}}\right)^{q_n}}{\sum_{i=1}^{N} e^{a_{q_i}}\left(\frac{w}{P_{q_i}}\right)^{q_i}} \quad \forall\, q_n \in Q. \end{aligned} \tag{4}$$

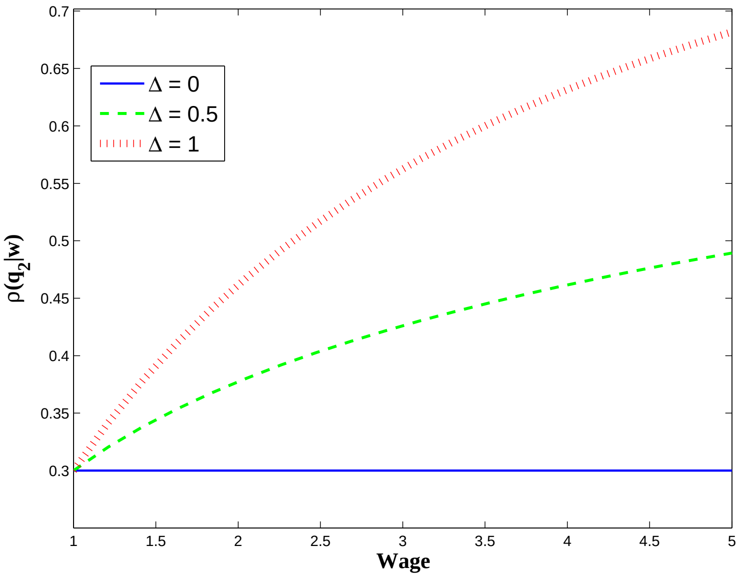
Analyzing how $\rho\left(q_n|w\right)$ changes as wage changes can help understand how this preference structure leads to non-homotheticity with respect to quality choice. Define $\gamma_{\rho(q_n),w}$ to be the elasticity of $\rho\left(q_n|w\right)$ with respect to wages w. Taking logs and differentiating equation (4) with respect to $\log(w)$ yields

$$\gamma_{\rho(q_n),w} = \frac{\partial \log\left[\rho\left(q_n|w\right)\right]}{\partial \log(w)} = q_n - \sum_{i=1}^{N} q_i \rho\left(q_i|w\right).$$

The elasticity of $\rho\left(q_n|w\right)$ with respect to wages w is simply the quality index q_n minus a weighted average of all the quality indexes, where the weights are the share of households with wage w who buy each quality level. A positive elasticity $\left(q_n > \sum_{i=1}^{N} q_i \rho\left(q_i|w\right)\right)$ implies that as wages increase, a larger share of the households buy the quality q_n. As lower quality goods have a lower quality index $\left(q_n > q_m\ \forall n > m\right)$, the lowest quality level will always have a negative elasticity, that is, the share of household who buy the lowest quality level will always go down as wages increase.

In our model, the non-homotheticity with respect to quality operates on the extensive margin. As a household becomes richer, it is more likely to choose the higher quality goods. There is a

Figure 9: Quality Engel Curve



Notes: The figure plots the share of households who purchase the high quality product for different wage levels. There are only 2 quality level $(N = 2)$ which have prices $P_{q_1} = 1$. Quality index for the low quality is set to one, that is, $q_1 = 1$. The three lines correspond to three different values of Δ where $q_2 = 1 + \Delta$. a_{q_2}, the constant for the high quality is chosen such that 30 percent of households with wage equal to one choose the high quality.

positively sloped "quality Engel curve" where households with higher levels of wages will, on average, spend a larger share of their expenditure on higher quality goods. This arises because the utility function in equation (2) features complementarity between quantity consumed and quality. As wages increase, the household can consume more quantity of whichever quality level that it chooses. Complementarity between quantity and quality implies that the marginal increase in utility from a given increase in wage is larger for higher quality goods which leads to more households choosing higher quality levels as wages increase (given the draw of ε_{j,q_n}).

The steepness of the quality Engel curve is determined by the differences in the quality indexes across quality levels. One way of parameterizing the quality indexes is to set the index for the lowest quality level to be one and assume that each higher quality level has an index which is a constant Δ larger than the previous quality index.[19] In this case, the size of the constant Δ determines the extent of non-homotheticity with a larger Δ implying that demand shifts to higher quality faster as wages increase.

[19] For example, $q_1 = 1$ and $q_n = q_{n-1} + \Delta$

Consider the following simple example which illustrates this relation between the size of Δ and the extent of the non-homotheticity. Assume that there are only two quality level $(N = 2)$ which have prices $P_{q_1} = 1$ and $P_{q_2} = 1.5$ and quality indexes $q_1 = 1$ and $q_2 = 1 + \Delta$.[20] Figure 9 plots the share of households who choose the high quality level q_2 as a function of wages for different value of Δ.[21] For each value of Δ, the constant in the utility function a_{q_2} is chosen such that 30 percent of the households with wage equal to one choose the high quality q_2.[22]

For the case with $\Delta = 0$ (blue line in Figure (9)), there is no change in the share of households who buy the high quality as wage increases. This is expected as $\Delta = 0$ ensures there is no quality distinction between the goods. For positive values of Δ, there is an increase in the share of households who buy the high quality good as wages increase, and this increase is larger for higher values of Δ (compare the green line with the red line).

Given prices and the wages of skilled and unskilled workers, the total demand for quality level q_n is given by

$$C_{q_n} = \underbrace{Nh\rho\left(q_n|w_S\right)\frac{w_S}{P_{q_n}}}_{\text{demand from skilled households}} + \underbrace{N\left(1-h\right)\rho\left(q_n|w_U\right)\frac{w_U}{P_{q_n}}}_{\text{demand from unskilled households}} \quad \forall\, q_n \in Q. \tag{5}$$

The first term is the demand for quality q_n from skilled households which is the product of the number of skilled households (Nh), the share of skilled households who choose quality q_n $(\rho\left(q_n|w_S\right))$, and the quantity consumed by each skilled household who consumes quality q_n $\left(\frac{w_S}{P_{q_n}}\right)$. Similarly, the second term is the demand for quality q_n from unskilled households.

In summary, the consumers choose between different quality levels and complementarity between quality and quantity implies that richer households are more likely to consume higher quality. This non-homotheticity with respect to quality will help match the patterns seen in Table 1 (that richer households buy higher price goods).

[20] In the full calibration done in Section 4, there is a richer quality space with $N = 12$. Here, to illustrate the non-homotheticity, the simplifying assumption of $N = 2$ is made.

[21] The results in Figure 9 can be viewed as the choice made by an individual *if* they faced a continuous wage profile. However, only two wages will exist in equilibrium (the unskilled and the skilled wages).

[22] Only $N-1$ constants in the utility function are identified as what matters for consumer choice is the difference in utility across quality levels. Therefore, for the case with $N = 2$, only one constant needs to be calibrated.

3.2. Final Goods Producers

There are N competitive final goods producers, one for each quality level. In addition to the vertical differentiation across quality levels, there is horizontal differentiation in products within a quality level. The final goods producer of quality q_n combines intermediate varieties (horizontal differentiation) of quality q_n to produce the composite final good of that quality. Each final goods producer has a constant elasticity of substitution (CES) production function given by

$$Y^s_{q_n} = \frac{1}{M_{q_n}^{\frac{1}{\sigma-1}}} \left(\sum_{i=1}^{M_{q_n}} x_{i,q_n}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \quad \forall q \in Q$$

where i indexes varieties, M_{q_n} is the number of varieties (or plants) of quality q_n present in the economy which will be determined by free entry, x_{i,q_n} is the quantity of variety i of quality q_n used by the final quality producer of quality q_n,[23] and σ is the elasticity of substitution between different varieties of the same quality.

The multiplicative factor $\frac{1}{M_q^{\frac{1}{\sigma-1}}}$ in the production function scales out the love of variety from the CES production function. This ensures that the price difference between different quality levels does not reflect differences in number of varieties available. We maintain this assumption of no love of variety in the baseline specification for two reasons. Firstly, assuming no love of variety is the conservative choice as changes in the size distribution in the counterfactual exercises are smaller in this case as opposed to the case with love of variety. Secondly, allowing for love of variety makes the changes in size distribution in the counterfactual sensitive to the average level of the quality indexes q_n which is a difficult parameter to calibrate as it represents the own price elasticity of each quality level with respect to the unobserved CES price index of that quality.[24] Therefore, while the baseline results presented in Section 5.1 maintains the assumption of no love of variety, Section 5.3 provides results when allowing for love of variety and further discuses the

[23] Note that the pair (i, q_n) together identifies a variety uniquely in the economy. i represents the horizontal differentiation dimension while q_n represents the vertical differentiation dimension. For example, $(i = 1, q_1)$ represents the first variety of lowest quality q_1 while $(i = 1, q_N)$ represents the first variety of the highest quality.

[24] As mentioned in Section 3.1, we parametrize the quality indexes using the recursion $q_n = q_{n-1} + \Delta$, where the size of Δ determines the steepness of the quality Engel curves. With no love of variety, the choice of the level of q_1 (which given a Δ determines the average level of the quality indexes) does not impact the changes in size distribution in the counterfactual. However, when allowing for love of variety, the results become sensitive to the choice of q_1.

sensitivity of the results to the average level of the quality indexes q_n.

The final quality producers take the prices of intermediate varieties, p_{i,q_n}, as given and solve their cost minimization problem

$$\min_{x_{i,q_n}} \sum p_{i,q_n} x_{i,q_n}$$
$$s.t.\ Y^s_{q_n} = \frac{1}{M_{q_n}^{\frac{1}{\sigma-1}}} \left(\sum_{i=1}^{M_{q_n}} x_{i,q_n}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \quad \forall q_n \in Q.$$

This yields their demand curves

$$x_{i,q_n} = p_{i,q_n}^{-\sigma} M_{q_n}^{\frac{1}{\sigma-1}} Y^s_{q_n} \left(\sum_{i=1}^{M_{q_n}} p_{i,q_n}^{1-\sigma} \right)^{\frac{\sigma}{1-\sigma}} \qquad \forall q_n \in Q, \tag{6}$$

which are taken as given by downstream intermediate producers. The final quality producers make zero profits. The price that they charge consumers is given by

$$P_{q_n} \quad = \quad \frac{\sum_{i=1}^{M_{q_n}} p_{i,q_n} x_{i,q_n}}{Y^s_{q_n}}, \quad \forall q_n \in Q.$$

Given the assumption of no love of variety, P_{q_n} will be independent of the number of varieties M_{q_n} available in the economy.

3.3. Intermediate Goods Producers

Each variety of each quality is produced by a monopolistically competitive intermediate producer. The intermediate producers combine skilled and unskilled labor and their production function is given by

$$x(A_i, q_n) = A_{i,q_n} \left(\theta_{q_n} \left(l^U_{i,q_n} \right)^{\frac{\sigma_{su}-1}{\sigma_{su}}} + \left(1-\theta_{q_n}\right) \left(l^S_{i,q_n} \right)^{\frac{\sigma_{su}-1}{\sigma_{su}}} \right)^{\frac{\sigma_{su}}{\sigma_{su}-1}}, \tag{7}$$

where l^U_{i,q_n} is the quantity of unskilled labor hired by variety i producer of quality q_n, l^S_{i,q_n} is the quantity of skilled labor hired by variety i producer of quality q_n, σ_{su} is the elasticity of substitution between the two types of labor, A_{i,q_n} is the idiosyncratic productivity level of variety i producer of quality q_n, and θ_{q_n} is the share parameter of unskilled labor for quality q_n producers.

Solving the cost minimization problem of the intermediate goods producer subject to the production function given in equation (7) yields the marginal cost of production for variety i of quality q_n which is given by

$$\kappa\left(A_i, q_n\right) = \frac{1}{A_{i,q_n}\left(\theta_{q_n}^{\sigma_{su}}\left(\frac{1}{w_U}\right)^{\sigma_{su}-1} + \left(1-\theta_{q_n}\right)^{\sigma_{su}}\left(\frac{1}{w_S}\right)^{\sigma_{su}-1}\right)^{\frac{1}{\sigma_{su}-1}}}.$$

The marginal costs is a function of skilled and unskilled wage, and is inversely proportional to the productivity level A_{i,q_n}.

Intermediate quality producers will take the demand curve of final quality producers (equation 6) as given and will maximize profits. As the demand curve of final quality producers is of the constant elasticity form, the optimal price charged by intermediate producers will be a constant markup over marginal cost and is given by

$$p\left(A_i, q_n\right) = \frac{\sigma}{\sigma-1}\kappa\left(A_i, q_n\right). \tag{8}$$

To start an intermediate goods plant of quality q_n requires f_{q_n} units of labor. Share α_{q_n} of the entry labor needs to be skilled and this share is different for different quality levels. On paying the fixed cost f_{q_n}, entrant receive a productivity draw from a log normal distribution given by

$$\log\left(A_{i,q_n}\right) \sim g_{q_n} \sim N\left(\mu_{q_n}, \nu^2\right).$$

Note that the mean of the log of the productivity draw can differ across quality levels but the variance is the same.

Free entry requires that the fixed cost payed must equal the ex-ante expected profit i.e.

$$\alpha_{q_n} f_{q_n} w_S + \left(1-\alpha_{q_n}\right) f_{q_n} w_U \quad = \quad \int \pi\left(A_i, q_n\right) g_{q_n}\left(A_i\right) dA_i \quad \forall q_n \in Q \tag{9}$$

where $\pi(A_i, q_n)$ is the flow profit earned by an intermediate quality producer of quality q_n with productivity draw A_i and is given by

$$\pi(A_i, q_n) = [p(A_i, q_n) - \kappa(A_i, q_n)]\, x(A_i, q_n).$$

The number of varieties M_{q_n} will adjust to ensure that the free entry condition holds for all quality levels.

If fixed costs for higher quality levels is larger than for lower quality levels, then for the free entry condition to hold, the scale of production $x(A_i, q_n)$ will have to be larger for higher quality producers. Furthermore, if $\theta_{q_n} > \theta_{q_m} \;\; \forall n > m$ then higher quality producers will use skilled labor more intensively and will have a higher cost of production. Finally, differences in μ_{q_n} will also translate into differences in prices between different quality levels as marginal costs and prices are proportional to productivity.

One simplifying assumption in our model is that output prices are fully determined by production costs and do not serve, in equilibrium, as signals of quality. This outcome arises because we assume perfectly competitive final goods markets with zero economic profits, and monopolistically competitive intermediate producers with constant markups. As a result, prices and average firm size reflect the cost of intermediate inputs and fixed costs. Therefore, higher quality firms charge higher prices and are larger on average only due to the technology requirement of higher quality production requiring higher fixed and variable costs. While this abstraction allows us to isolate the supply-side forces driving the quality–firm size relationship, it omits a potential demand-side mechanism. We acknowledge that in many real-world settings, firms producing higher quality goods may also enjoy pricing power. Nevertheless, the model delivers a positive relationship between price and quality via firm input choices, which aligns with the empirical patterns we observe: higher-priced firms invest more, use more capital, and employ more skilled labor. We interpret this as indirect evidence of quality, while recognizing the model's limitations in capturing price as a signal of consumer valuation.

3.4. Equilibrium

The equilibrium in this economy is a set of prices $\left(w_S, \left\{\left\{p_{i,q_n}\right\}_{i\in M_{q_n}}, P_{q_n}\right\}_{q_n\in Q}\right)$, allocations $\left\{\left\{c_{j,q_n}\right\}_{j\in L}, C_{q_n}, \left\{x_{i,q_n}\right\}_{i\in M_{q_n}}, Y_{q_n}\right\}_{q_n\in Q}$, and mass of entrants M_{q_n} such that

- Given prices P_{q_n}, wages, and draws of the random utility component $\left(\varepsilon_{j,q_n}\right)$, consumers choose their optimal quality level (equations 4 and 5 hold)
- Given prices, final quality producers demand optimal amounts of intermediate goods (demand follows equation 6)
- Intermediate good producers maximize profits (charge the constant markup price given by equation 8)
- Free entry conditions hold for all quality levels (equation 9)
- Markets clear

$$Y_{q_n} = C_{q_n} \quad \forall q_n \in Q$$

$$L(1-h) = \sum_{q_n} M_{q_n} \int l^U(A_i, q_n)\, g_{q_n}(A_i)\, dA_i + \sum_{q_n} M_{q_n}\left(1-\alpha_{q_n}\right) f_{q_n} \tag{10}$$

$$Lh = \sum_{q_n} M_{q_n} \int l^S(A_i, q_n)\, g_{q_n}(A_i)\, dA_i + \sum_{q_n} M_{q_n} \alpha_{q_n} f_{q_n} \tag{11}$$

Equations (10) and (11) are the labor market clearing conditions. Equation (10) requires that the demand for unskilled labor for production by the intermediate producers (summing over all quality levels) and entry requirements must equal the supply of unskilled labor. Similarly, Equation (11) requires that the demand for skilled labor from intermediate producers and entry requirements must equal the supply of skilled workers.

4. Calibration

This section calibrates the model to match the cross-sectional facts documented in Section 2 and some additional moments taken from the Indian data. The key parameters in the counterfactual exercises that determine the *change* in the size distribution are the degree of non-homotheticity (Δ) for consumers and the price-size relationship for producers. These parameters are calibrated independently of the aggregate relationship between the share of employment in small plants and income levels seen across Indian states (which is what we want to explain in the counterfactual). In particular, we use the micro-facts documented in Section 2 (richer households buy higher priced goods and larger plants produce higher priced goods) to discipline these parameters of the model.

4.1. Production Parameters

For the calibration, we define an individual with less than ten years of education as unskilled. h, the share of the labor force which is skilled, is set to 0.24, which is the share of manufacturing workers with at least ten years of education in India in 2004-05. σ_{su}, the elasticity of substitution between skilled and unskilled workers in the intermediate goods production function (equation 7), is assumed to be 1.75 which is in the range of estimates for developing countries in Behar [2009].

The elasticity of substitution between varieties for the final goods producer, σ, is set to 5, which implies a markup over cost of 25 percent for the intermediate producers and is in the range of estimates in Broda and Weinstein [2006].

This leaves five sets of parameters to be calibrated on the production side: (1) f_{q_n}, the fixed cost for each quality level; (2) θ_{q_n}, the share of unskilled workers in the production function for each quality level; (3) μ_{q_n}, the mean of the log of the productivity draw for each quality level; (4) α_q, the share of skilled labor needed for entry for each quality level; and (5) ν^2, the variance of the productivity draw which is common across all quality levels. These parameters (along with the utility parameters) are jointly calibrated as there is no one-to-one mapping between the parameters and the target moments. However, for expositional purposes, we explain the calibration of each parameter in terms of the moments which are most informative about the parameter.

Table 11: Unskilled to Skilled Ratio for Different Size Categories

	U/S Ratio	Ratio Relative to Smallest
$L <= 5$	5.05	1.00
$5 < L <= 20$	2.92	0.58
$L > 20$	1.25	0.25

Notes: The data is from the Employment-Unemployment Survey of 2004-05. The rows of the table represent the size category of the establishment in which an individual works. The first column gives the ratio of skilled to unskilled workers in each size category where the definition of skilled is assumed to be an individual with at least ten years of education. The second column gives the ratio of skilled to unskilled relative to the smallest size category.

The number of quality levels N is set to 12.[25]

The fixed costs, f_{q_n}, determines the average scale of operation of the intermediate producers of each quality level. A larger fixed cost will mean that the average size (in terms of output and employment) of intermediate producers will need to be larger in order for the the free entry condition to hold. As shown in Section 2.2.1, larger plants tend to produce higher price products, which is indicative of higher quality goods being produced in larger plants. Therefore, the fixed costs are chosen such that the average employment (skilled plus unskilled workers) in intermediate producers of the lowest quality levels is 1.25 workers and each higher quality level has double the average size of the previous quality level, that is, the average employment of the intermediate producers of the different quality levels are $size_{q_n} = \{1.25,\ 2.5,\ 5, ..., 2560\}$.[26]

The level of $\theta'_{q_n}s$ determine the demand for unskilled labor relative to skilled labor and are informative about the wage premium, w_S, in the economy. The ratio of skilled to unskilled workers in any quality level relative to the lowest quality is also a function of the $\theta'_{q_n}s$ and is given by

$$ratio_{q_n}^{U,S} = \left(\frac{L^U_{qn}}{L^S_{qn}}\right) / \left(\frac{L^U_{q1}}{L^S_{q1}}\right) = \left(\frac{\theta_{qn}}{1-\theta_{qn}}\right)^{\sigma_{us}} / \left(\frac{\theta_{q1}}{1-\theta_{q1}}\right)^{\sigma_{us}} \quad \forall q_n \in Q. \tag{12}$$

[25] The results discussed in Section 5 are not very sensitive to the choice of N. For example, if we instead choose N to be 6, and choose all the other parameters in the same way as described below, then the model explains 45 percent instead of 43 percent of the differences in share of employment in small plants in rich versus poor states (the baseline results discussed in Section 5.1).

[26] Different intermediate producers of the same quality will have different levels of employment due to heterogeneity in the productivity draw. Within the same quality level, intermediate producers with higher productivity draws will be larger compared to those with lower productivity draws. The fixed costs are chosen such that the average employment level of the producers within a quality level matches the target $size_{q_n} = \{1.25,\ 2.5,\ 5, ..., 2560\}$.

Therefore, the twelve $\theta'_{q_n}s$ are chosen to match a target for the wage premium and eleven targets for unskilled to skilled ratio in different quality levels relative to the lowest quality level.

The targets for these moments are obtained from the Employment-Unemployment Survey conducted by the NSS in 2004-05 (see Section 2.2.2 and Appendix A.4 for details about the dataset).

The target for the wage premium is set at 1.6, and is obtained from running Mincerian regressions on data from the Employment-Unemployment Survey.[27] Table 11 gives the ratio of unskilled to skilled workers for three different size categories, along with the ratio relative to the smallest size category, as computed from the Employment-Unemployment Survey. Smaller plants have a much higher ratio of unskilled to skilled workers indicating that low quality producers have higher $\theta'_{q_n}s$. Unfortunately, the size categories reported in the Employment-Unemployment survey are very coarse, and therefore cannot be used to compute eleven ratios for equation (12) for eleven different quality (size) levels. We use the first two data points reported in Table 11 for the unskilled to skilled ratio (column 1) and extrapolated the relationship to larger sizes (with a minimum of 0.5) to compute eleven ratios, one for each quality (size) level.

μ_{q_n}, the mean of the log of the productivity draw for each quality level, is informative about the average price of each quality level as $p(A_i, q_n) \propto \frac{1}{A_i}$. If the mean of the productivity draw for a particular quality is high, then the average price of that quality level will be lower. Therefore, the μ_{q_n} for each quality level is chosen to match the price-size relationship seen in Table 3.[28]

The share of skilled labor needed for entry for each quality level, α_{q_n}, is chosen to match the share of skilled labor used in the production of that quality. Therefore, high quality producers use a more skill intensive production process (lower θ_{q_n}) and also have more skill intensive entry requirement.[29]

[27] Werun a regression of log wages on a dummy of whether the individual is skilled (at least ten years of education) for all manufacturing workers, controlling for potential experience, sex, state, industry, occupation, and whether the individual is residing in a rural or urban area. Individuals with ten or more years of education on average make 56.8 percent more than workers with less than ten years of education which is rounded up to a wage premium of 1.6. Appendix G reports more details and the regression results.

[28] In particular, the $\mu'_{q_n}s$ are chosen to match a price-size slope of 0.1. Note that plants of each higher quality level are calibrated to be two times the size of the previous quality level. Therefore, the $\mu'_{q_n}s$ are chosen such that each higher quality level charges a log price which is $0.1 * \log(2)$ higher than the previous quality levels log price.

[29] The ratio of skilled to unskilled labor used by plants of quality q_n is given by $\frac{l^U_{i,q_n}}{l^S_{i,q_n}} = \left(\frac{w_S}{w_U}\frac{\theta_q}{1-\theta_q}\right)^{\sigma_{us}}$ and is independent of the productivity draw of the plant. Therefore, the share of skilled workers used in production of quality q_n is

Table 12: Calibration

Param.	**Description**	**Targets**
f_{q_n}	Fixed costs	$size_{q_n} = \{1.25,\ 2.5,\ 5,\ ...,\ 2560\}$
μ_{q_n}	Mean of productivity draws	Price-size slope of 0.1
θ_{q_n}	Sh of U in production	$w_S = 1.6$; and $\frac{L^U_{qn}}{L^S_{qn}}$ across qualities
α_{q_n}	Sh of skilled in entry	$\frac{L^S_{qn}}{L^S_{qn}+L^U_{qn}}$ in production
ν^2	Variance of productivity draw	Std dev of employment = 0.64
$q_n\ (\Delta)$	Utility from quality	Price-income slope of 0.1
a_{q_n}	Constant in utility function	Size distribution

Finally, ν^2 , the variance of the log of the productivity draw (common across qualities), is chosen to match the standard deviation of the log of employment in the combined ASI and SUM dataset which was 0.64.

4.2. Utility Parameters

The utility function in the model takes the form

$$u_{j,q_n}\left(c_{j,q_n}, \varepsilon_{j,q_n}\right) = a_{q_n} + q_n \log\left(c_{j,q_n}\right) + \varepsilon_{j,q_n} \quad \forall\, q_n \in Q. \tag{13}$$

Two sets of parameters need to be calibrated: (1) q_n, the quality indexes; and (2) a_{q_n}, the quality specific constant in the utility function.

As mentioned in Section 3.1, the quality indexes are parametrized as follows: $q_1 = 1$ and $q_n = q_{n-1} + \Delta$.[30] The value of Δ determines the steepness of the quality Engel curve, that is, how quickly does demand move to higher quality as income levels increase. In the model, skilled workers earn wage w_S (which is calibrated to be 1.6) and unskilled workers earn wage w_U (which is normalized

simply $\frac{1}{1+\left(\frac{w_S}{w_U}\frac{\theta_q}{1-\theta_q}\right)^{\sigma_{us}}}$.

[30] Setting q_1 to be one is not a normalization in the model. However, for the baseline specification with no love of variety, the results are not sensitive to the choice of q_1. This issue is discusses further in Section 5.3.

to one as the numeraire). Δ is chosen to match the price-income relationship documented in Table 1 of Section 2.1. In particular, Δ is chosen such that the price-income elasticity in the model is 0.1, that is, the average log price paid by skilled households is $0.1 * \log\left(\frac{w_S}{w_U}\right)$ more than for unskilled households. As higher quality producers in the model have higher prices, this in effect determines the extent to which demand shifts towards high quality as we move from unskilled wages to skilled wages.

The quality specific constant in the utility function, a_{q_n}, determines the absolute levels of demand for different quality levels i.e. it determines $\rho\left(q_n|w\right)$ given in equation (4). A higher a_{q_n} for a specific quality means that a larger share of households are likely to buy that quality (irrespective of income level). Therefore, we choose a_{q_n} such that the size distribution in the model matches the size distribution for India as a whole in 2005-06.

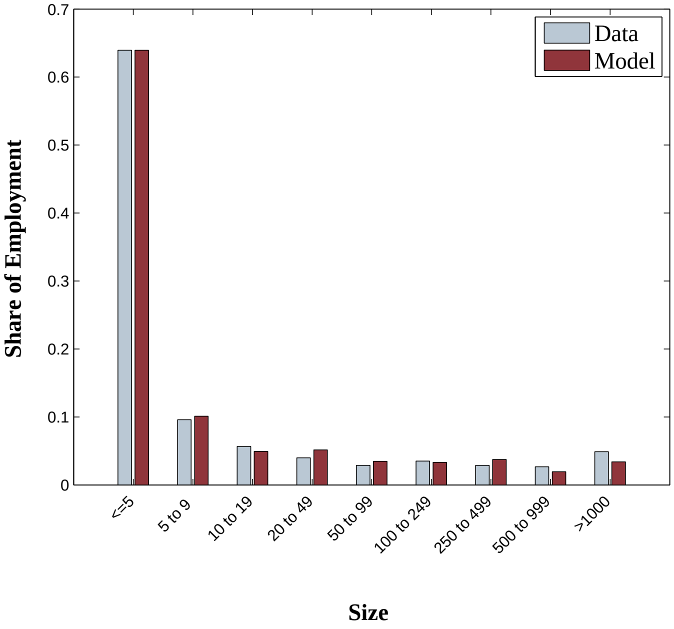
In summary, a_{q_n} pins down the absolute level of demand for the different qualities and are calibrated to match the size distribution in the model to the Indian data. Δ determines the differences in demand for high versus low quality levels between skilled and unskilled workers and is calibrated to match the price-income elasticity seen in the data.

Table 12 summarizes the calibration. Figure 10 plots the share of workers in plants of different size categories for the calibrated model and the data (combining the ASI and the SUM for 2005-06). As the model parameters were chosen to match the size distribution, it is not surprising to see that the size distribution in the model matches the data very closely. However, the model was not calibrated to match the change in size distribution as income levels change. The extent to which the size distribution changes in the model as income levels change depends crucially on the degree of non-homotheticity (Δ) on the consumer side and the price-size relation on the producer side and these parameters were calibrated using micro-data from consumer and producer surveys.

5. Results

This section first conducts counterfactual exercises that simulate differences in per-capita income levels and examine how this effects the size distribution. Second, we explore the sensitivity of the results to some important parameters.

Figure 10: Size Distribution - Data vs Model



Notes: The figure plots the share of employment in different size categories in the data and in the calibrated baseline of the model. The data is for the manufacturing sector in India for 2005-06. It combines the ASI and the SUM (same as Figure 1).

5.1. Cross-section of Indian States

How much of the cross-state differences in the size distribution seen in the data can be explained by the model if per-capita income in the model varies by the same amount as it varies across Indian states? To answer this question, we conduct counterfactual exercises in which we vary three sets of parameters in the model while keeping all the other parameters unchanged:

1. The share of the households in the model who are skilled, h, is varied in the counterfactual exercises to match the share of workers with ten or more years of education across rich and poor states. About 13 percent of the manufacturing workers in the poorest states are skilled as compared to 43 percent in the richest states.

2. The share parameter of unskilled labor for intermediate producers, θ_{q_n}, is changed across the counterfactuals to keep the wage premia unchanged.[31] This can be viewed as skill biased technical change with richer states having a higher supply of skilled labor and also using skilled labor more intensively in the production of all quality levels.[32]

3. The mean of the productivity draw of intermediate producers, μ_{q_n}, is changed to match the differences in per-capita income across states and to maintain the price-size slope of 0.1 across the counterfactuals.[33] Per-capita income of the poorest Indian state (Bihar) is

[31] If we do not change θ_{q_n}, then the wage premia falls in the counterfactual for the richer states due to the higher supply of skilled workers. However, in the data, wage premia does not vary systematically across states. In particular, if we run a Mincerian regression of log of wages on a dummy which takes value 1 if the person is skilled and also include the interaction of the dummy with per-capita state NDP (controlling for industry, occupation, sex, experience etc), then the coefficient on the interaction is not significantly different from zero.

[32] In effect, θ_{q_1} for each counterfactual is chosen to maintain the wage premia ($w_S = 1.6$). All the other $\theta'_{q_n}s$ are picked as described in equation (12) to match the ratio of skilled to unskilled in different quality levels relative to the worst quality level. Furthermore, in the counterfactual, the share of entry labor which needs to be skilled workers (α_{q_n}) is also changed to match the share of skill in production for each quality level. That is, the richer states do not just use more skill intensive production techniques but also use more skill in the entry process.

[33] As mentioned in Section 4.1, μ_{q_n} for each quality level was chosen to match the price-size elasticity of 0.1. In the counterfactual exercises, as the $\theta'_{q_n}s$ are changed, this can lead to changes in prices of the high quality relative to low quality even though there is no change in wage premia. These changes in relative prices can cause a shift in demand and thus changes in the size distribution for reasons other than changes in real income which is what we want to focus on. Therefore, in the counterfactual, in addition to scaling all the $\mu'_{q_n}s$ by a constant (to match the differences in per-capita income seen across Indian states), we also change the relative $\mu'_{q_s}s$ of different qualities to maintain the same relative prices of different quality levels. This eliminates any substitution effects due to relative price changes and only focuses on changes in demand (caused by the non-homotheticity in the preferences) due to changes in per-capita income levels.

0.39 times India's per-capita income while that of the richest state (Maharashtra) is 1.57 times India's per-capita income. To generate similar differences in per-capita income in the model, the poorer states in the counterfactual exercise have lower average productivity levels compared to the richer states.[34]

To summarize, three sets of parameters are changed in the counterfactual exercises: the share of skilled in the population, the skill intensity of the production process, and the means of the productivity draws of intermediates. These parameters are changed to match the differences in skill composition and per-capita income levels across Indian states while keeping the wage premia and the relative prices of different quality levels unchanged.[35]

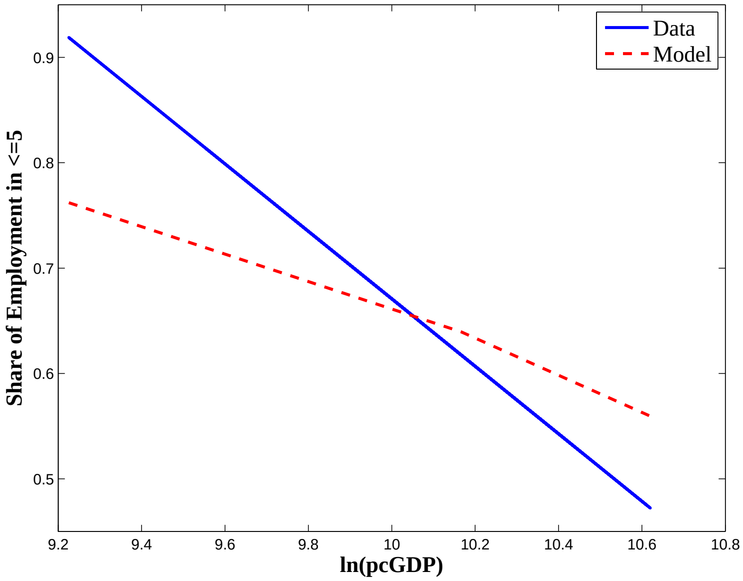
An increase in the productivity of intermediate producers and in the supply of skill translates into an increase in real income levels in the model. The increase in real income level leads to demand shifting towards higher quality goods due to the non-homotheticity in the preferences. This change in demand leads to a shift in the production side. The number of plants producing low quality goods declines while those producing high quality increases. This in turn implies that there is a shift in the size distribution with the share of employment in small plants falling.

The red dashed line in Figure 11 plots the share of employment in plants of size five or less that is predicted by the model when conducting the counterfactual exercises. In the calibrated baseline, the share of employment in small plants in the model is 63.9 percent. When productivity and supply of skill is lowered such that per-capita income levels decrease by a factor of 0.39 (0.94 log points lower), the share of employment in plants of size five or less increases to 75.6 percent. On the other hand, when productivity and supply of skill is increased such that per-capita income levels increase by a factor of 1.57 (0.43 log points higher) compared to the calibrated baseline, the

[34] In order to define per-capita GDP in the model, we need to define a set of base prices. We use the prices of intermediates in the calibrated baseline as the base prices and value output in the counterfactuals using these prices.

[35] Much of the variation in income levels across states in the model is captured by differences in productivity, as higher productivity directly raises the amount of output that can be produced by labor. On the other hand, human capital plays a relatively minor role in explaining income differences across counterfactuals because skilled workers in the model are not intrinsically more efficient than unskilled workers; that is, they do not possess more effective units of labor. Rather, skilled and unskilled workers are just two distinct inputs in the production function and the relatively low supply of skilled workers compared to demand results in the skilled workers getting a wage premia. As such, varying the share of skilled workers does not necessarily increase aggregate per capita income levels in the model. This result is consistent with development accounting exercises which also find that residual TFP explains the majority of the differences in per-capita income across Indian states (see Chanda [2011]).

Figure 11: Counterfactual Across Indian States - Data vs Model



Notes: The figure plots the share of employment in plants of size five or less across Indian states in the data and for the counterfactual exercise in the model. The blue line is the linear regression line of share of employment in plants of size five or less in different Indian states on log of per-capita GDP of the state. The red line is the model predicted share of employment in plants of size five or less when conducting the counterfactual exercise.

share of employment in small plants falls to 56.3 percent.

The solid blue line in Figure 11 plots the projection from a linear regression of the share of employment in plants of size five or less on log of per-capita State NDP across Indian states. The share of employment in small plants is computed by combining the ASI and the SUM (the same data as in Figure 2). In the data, the poorest Indian states have about 91.9 percent of employment in small pants while the richest have 47.2 percent employment in small plants.

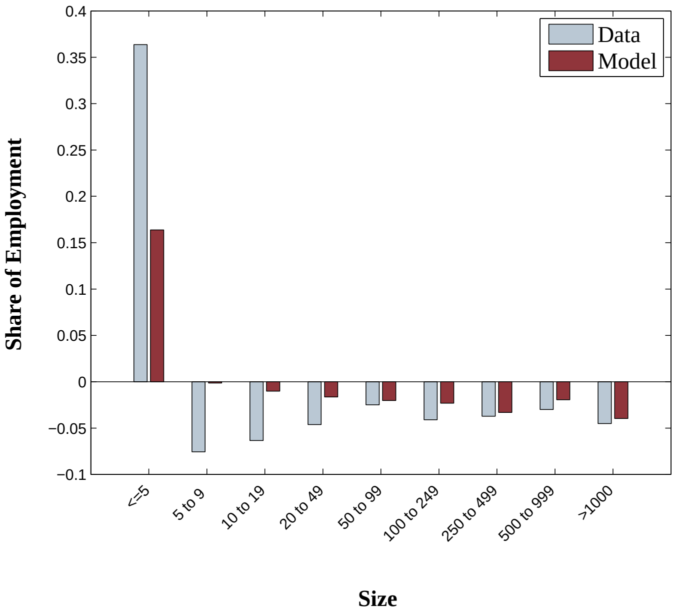
While the share of employment in small plants varies by 44.7 percentage points across Indian states in the data, the model predicts an 19.3 percentage points difference. Therefore, the model explains about 43 percent of the difference in share of employment in small plants seen across Indian states.

Figure 12 compares how the entire size distribution (as opposed to just the share of employment in plants of size five or less) changes in the model as compared to the data as we change income levels. In the data, we pool together the three poorest states and the three richest states and compute the share of employment in different size categories for these groups of states.[36]

The light blue bars in Figure 12 show the difference (in percentage points) in the share of employment in the three poorest states compared to the three richest states for each size category. The poorest states have about 36 percentage points more employment in plants of size five or less as compared to the richest states. The richer states have a larger share of their employment in all the larger size categories as compared to the poor states, which is why the the blue bars lie below zero for all these size categories. The red bars represent the same difference in share of employment for different size categories that the model predicts when productivity and skill levels in the model are varied to match the incomes differences across these groups of states. The model predicts that the share of employment in plants of size five or less is about 15 percentage points higher in the poorer states as compared to richer states, which again accounts for about 42 percent of the difference seen in the data. Again, like the data, the red bars lie below zero for all the other size categories, indicating that the model predicts a larger share of employment in richer states for these size categories.

[36] We pool the three richest and poorest states in order to avoid having the results being driven by an outlier state. The results are similar if we just compare the richest state to the poorest state.

Figure 12: Counterfactual: Changes in Distribution for 3 Richest vs 3 Poorest States



Notes: The figure plots the share of employment in the three poorest states minus the share in the three richest states for different size categories in the data and in the model (when productivity and skill levels are varied to match the differences in per-capita income across these groups of states). The data is from the ASI and SUM for 2005-06.

5.2. India Over Time

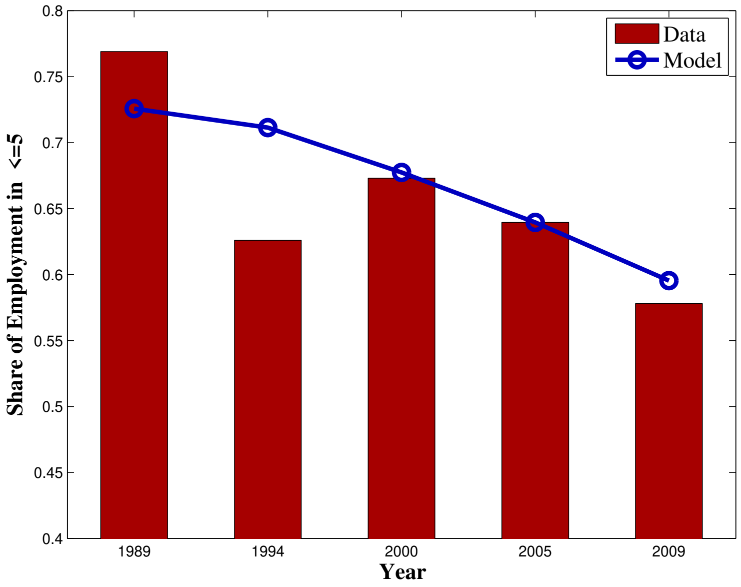
This subsection examines how well the model can explain the changes in the size distribution of manufacturing plants in India over time. Five waves of the Survey of Unorganized Manufacturing (SUM) have been conducted in Indian between 1989-90 and 2010-11. These can be combined with the corresponding years of the Annual Survey of Industries (ASI) to get five data points for how the size distribution has evolved over time in India.

The bars in Figure 13 show the share of employment in plants of size five or smaller for 1989, 1994, 2000, 2005, and 2009.[37] As can be seen, the share of employment in small plants has decreased from 77 percent of total employment in 1989 to 58 percent in 2009.

Per-capita income in 1989 was 0.54 times the 2005 level of per-capita income while the share of manufacturing workers with ten or more years of schooling was just 14 percent. In 2009 per-capita income levels were 1.30 times the 2005 level while the share of manufacturing workers with ten

[37] More details of the surveys are given in Appendix A.1 and A.2.

Figure 13: Counterfactual India Over Time - Data vs Model



Notes: The red bars in the figure plot the share of employment in plants of size five or less for five years for India. The data for each year pools the SUM and and the the ASI for that year. The blue line plots the model predicted share of employment for each year when productivity and skill levels are varied to match the differences in per-capita income in India over time.

or more years of schooling had increased to 31 percent. The blue line in Figure 13 plots the share of employment in plants of size five or less as predicted by the model when productivity and skill supply in the model is varied to the extent required to match the differences in per-capita income levels and share of skilled in the data. The model was calibrated to match the share of employment in small plants in 2005, therefore, the fit in 2005 is very good by construction. The model predicts that 72 percent of employment would be in plants of size five or less in 1989, which is a little less than the 77 percent seen in the data. Similarly, the model under-predicts the change in the size distribution going from 2005 to 2009 by a small amount. Overall, the model predicts 65 percent of the change in share of employment in small plants seen in the data between 1989 to 2009.[38]

[38] A related question can also be asked: Did states which grow more over time see a larger drop in share of employment in small plants? If we look at the change in the size distribution between 1989 and 2009, then this does seem to be the case.

Table 13: Love of Variety: Percent of Cross-State Difference Explained

	$q_1 = 1$	$q_1 = 0.1$
$\eta = \frac{1}{\sigma-1}$	43.1%	43.1%
$\eta = 0$	71.2%	53.1%

Notes: The table shows the percent of cross-state variation in share of employment in plants of size five or less that is explained by the model counterfactual for different parameter values of η and q_1.$\eta = \frac{1}{\sigma-1}$ is the baseline specification of no love of variety while $\eta = 0$ is the case of full love of variety.

5.3. Parameter Sensitivity: Love of Variety

As mentioned in Section 3.2, the baseline specification of the model assumed that the final goods producers production function had no love of variety. A generalization of the the production function of the final goods producer of quality q_n is given by

$$Y^s_{q_n} = \frac{1}{M^{\eta}_{q_n}} \left(\sum_{i=1}^{M_{q_n}} x_{i,q_n}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \qquad \forall q \in Q.$$

In the baseline specification, η was set equal to $\frac{1}{\sigma-1}$, which corresponded to the case of no love of variety. In this section, we provide results for the case when $\eta = 0$ (the case with full love of variety) and compare this to the baseline. As mentioned in Section 3.2, the no love of variety assumption is the conservative case, with changes in the size distribution in the counterfactual being larger when we allow for love of variety. Furthermore, when allowing for love of variety, the results become more sensitive to the choice of q_1, the quality index of the lowest quality level (note that given q_1, all subsequent quality indexes are given by the recursion $q_n = q_{n-1} + \Delta$).

Table 13 shows how much of the cross-state differences in share of employment in small plants is explained by the model for different values of η and q_1. The first row and first column corresponds to the baseline specification, with $\eta = \frac{1}{\sigma-1}$ (no love of variety) and $q_1 = 1$. As mentioned in Section 5.1, when varying productivity and supply of skill to match the differences in per-capita incomes across states, the model explains 43.1 percent of the difference in the share of employment in small plants as compared to the data.

Now consider the model with love of variety ($\eta = 0$). When allowing for love of variety, all other parameters are recalibrate to match the same moments as in the baseline. We then run the same counterfactual exercises as in Section 5.1. As reported in Table 13, in the case with love of variety, the model can explain 71.2 percent of the differences in size distribution between the rich and poor states.

Why is it that in the case with love of variety, the model generates bigger changes in the size distribution in the counterfactual? The reason is that in the case with love of variety, relative prices of different quality levels change in the counterfactual, due to changes in the relative varieties of the different qualities. In particular, the CES price index (the price charged by the final producer to the consumer) for quality q_n is given by

$$P_{q_n} = M_{q_n}^{\eta - \frac{1}{\sigma-1}} \left(\int \left(p\left(A_i, q_n\right)\right)^{1-\sigma} g_{q_n}\left(A_i\right) dA_{q_i} \right)^{\frac{1}{1-\sigma}} \quad \forall q \in Q.$$

In the baseline specification, because $\eta = \frac{1}{\sigma-1}$, the price index for q_n was independent of the of the number of varieties M_{q_n}. However, when $\eta = 0$, the CES price index of a quality level, P_{q_n}, is inversely related to the number of varieties of that quality (M_{q_n}) available in the economy. In the counterfactual, as income levels increase, demand shifts towards higher quality, and this induces more entrants of the higher quality levels. The increase in number of varieties of high quality intermediate producers causes the relative price of high quality goods to fall in the counterfactual when $\eta = 0$. This causes a further shift in demand towards high quality which in turn causes more entry into higher quality goods. The additional increase in demand for high quality which acts through relative price changes due to change in number of varieties does not occur in the baseline specification when $\eta = \frac{1}{\sigma-1}$. Hence, the change in size distribution in the counterfactual in the baseline specification is less than in the case with love of variety. In effect, the baseline specification focuses attention on the changes in demand caused by changes in income levels alone. It abstracts away from any changes in relative prices caused by changes in number of varieties in the counterfactual.

Furthermore, when allowing for love of variety, the change in the size distribution in the counterfactual, becomes more sensitive to the choice of q_1, the quality index for the lowest quality level. When q_1 is set to 0.1 and $\eta = 0$, the model counterfactual explains only 53.1 percent of the difference in size distribution as opposed to 71.2 percent when $q_1 = 1$. As shown in Section 3.1, the share of households with wage w who choose quality level q_n is given by

$$\rho\left(q_n|w\right) = \frac{e^{a_{q_n}}\left(\frac{w}{P_{q_n}}\right)^{q_n}}{\sum_{i=1}^{N} e^{a_{q_i}}\left(\frac{w}{P_{q_i}}\right)^{q_i}} \quad \forall\, q_n \in Q.$$

As P_{q_n} is raised to the power q_n in the numerator, the absolute levels of q_n approximately determine the own price elasticity of demand for a quality level. Lower absolute levels of the quality indexes imply that demand is less sensitive to changes in relative prices (of the CES price indexes). Therefore, a lower value for q_1 (which translates into lower values for all the quality indexes) makes the model less sensitive to the changes in relative prices induced by changes in varieties.

6. Conclusion

The size distribution in developing countries usually has a thick left tail compared to developed countries. The same holds across Indian states, with richer states usually having a much smaller share of their manufacturing employment in small plants. This paper explores the hypothesis that this income-size relationship arises from the fact that low income countries and states have high demand for low quality products which can be produced efficiently in small plants. We find compelling consumer and producer evidence that is consistent with this hypothesis. We show that richer households buy higher price goods and larger plants produce more expensive products (and use more expensive inputs). Finally, we develop a model that features non-homothetic preferences with respect to quality and is calibrated to match the cross-sectional empirical findings. Our calibrated model indicates that up to 41 percent of the cross-state variation seen in the left tail of manufacturing plants in India can be explained by the model.

A key simplification in our model is the assumption that product prices are determined entirely by input costs—directly following from our model of free entry and zero profits for final producers and monopolistic competition with constant markups for intermediate producers-—abstracting from the possibility that consumers may pay explicitly for quality. This assumption allows us to focus sharply on how technology differences impact production of quality and firm size. However, it also limits the model's ability to capture settings where firms can charge quality-dependent markups or where prices reflect consumer willingness to pay. We view this as an important avenue for future research. Extending the model to allow interaction between the demand-side for quality and firm pricing decisions would allow for a richer mapping between quality, price, and firm scale—potentially shedding further light on the variation observed across industries and countries.

To sum, this paper suggests that a large part of the differences in size distribution that we see across countries and states is a natural consequence of the low levels of income in developing countries and is not caused by policies which discriminate against large productive plants in favor of small unproductive plants. The presence of small plants in developing countries should not be viewed as originating necessarily from policy failures.

References

M. Aguiar and E. Hurst. Life-cycle prices and production. *American Economic Review*, 97(5):1533–1559, December 2007. URL http://ideas.repec.org/a/aea/aecrev/v97y2007i5p1533-1559.html.

U. Akcigit, H. Alp, and M. Peters. Lack of selection and limits to delegation: firm dynamics in developing countries. *American Economic Review*, 111(1):231–275, 2021.

L. Alfaro, A. Charlton, and F. Kanczuk. Plant-size distribution and cross-country income differences. In *NBER International Seminar on Macroeconomics 2008*, NBER Chapters, pages 243–272. National Bureau of Economic Research, Inc, April 2009. URL http://ideas.repec.org/h/nbr/nberch/8244.html.

D. Atkin and D. Donaldson. Who's getting globalized? The size and nature of intranational trade costs. Technical report, Yale University, 2012.

D. Atkin, A. K. Khandelwal, and A. Osman. Exporting and firm performance: Evidence from a randomized experiment. *The quarterly journal of economics*, 132(2):551–615, 2017.

O. P. Attanasio and C. Frayne. Do the poor pay more? Technical report, January 2006.

A. Banerji and S. Jain. Quality dualism. *Journal of Development Economics*, 84(1):234–250, September 2007. URL http://ideas.repec.org/a/eee/deveco/v84y2007i1p234-250.html.

L. Barseghyan and R. DiCecio. Entry costs, industry structure, and cross-country income and TFP differences. *Journal of Economic Theory*, 146(5):1828–1851, 2011.

V. Bassi, J. H. Lee, A. Peter, T. Porzio, R. Sen, and E. Tugume. Self-employment within the firm. Technical report, National Bureau of Economic Research, 2023.

A. Behar. Directed technical change, the elasticity of substitution and wage inequality in developing countries. Economics Series Working Papers 467, University of Oxford, Department of Economics, Dec 2009. URL http://ideas.repec.org/p/oxf/wpaper/467.html.

P. Bento and D. Restuccia. Misallocation, establishment size, and productivity. *American Economic Journal: Macroeconomics*, 9(3):267–303, 2017. doi: 10.1257/mac.20140137.

P. Bento and D. Restuccia. Financial frictions at entry, average firm size, and productivity. *B.E. Journal of Macroeconomics*, forthcoming.

M. Bils and P. J. Klenow. Quantifying quality growth. *American Economic Review*, 91(4):1006–1030, September 2001. URL http://ideas.repec.org/a/aea/aecrev/v91y2001i4p1006-1030.html.

N. Bloom and J. Van Reenen. Measuring and explaining management practices across firms and countries. *The quarterly journal of Economics*, 122(4):1351–1408, 2007.

C. Broda and D. E. Weinstein. Globalization and the gains from variety. *The Quarterly Journal of Economics*, 121(2):541–585, May 2006. URL http://ideas.repec.org/a/tpr/qjecon/v121y2006i2p541-585.html.

A. Chanda. Accounting for bihar's productivity relative to india's: What can we learn from recent developments in growth theory. Technical Report 11/0759, International Growth Centre, August 2011.

Y. C. Choi, D. Hummels, and C. Xiang. Explaining import quality: The role of the income distribution. *Journal of International Economics*, 77(2):265–275, April 2009. URL http://ideas.repec.org/a/eee/inecon/v77y2009i2p265-275.html.

X. Cirera, D. Comin, and M. Cruz. *Bridging the technological divide: Technology adoption by firms in developing countries*. World Bank Publications, 2022.

M. Dalgin, D. Mitra, and V. Trindade. Inequality, nonhomothetic preferences, and trade: A gravity approach. *Southern Economic Journal*, 74(3):747–774, January 2008. URL http://ideas.repec.org/a/sej/ancoec/v743y2008p747-774.html.

H. De Soto. *The other path*. Harper & Row New York, 1989.

A. Deaton and O. Dupriez. Spatial price differences within large countries. Working Papers 1321, Princeton University, Woodrow Wilson School of Public and International Affairs, Research Program in Development Studies., July 2011. URL http://ideas.repec.org/p/pri/rpdevs/1321.html.

Y. Dikhanov. Income effect and urban-rural price differentials from the household survey perspective. Technical report, ICP Global Office, 2010.

S. Djankov, R. L. Porta, F. Lopez-De-Silanes, and A. Shleifer. The regulation of entry. *The Quarterly Journal of Economics*, 117(1):1–37, February 2002. URL http://ideas.repec.org/a/tpr/qjecon/v117y2002i1p1-37.html.

P. Fajgelbaum, G. M. Grossman, and E. Helpman. Income distribution, product quality, and international trade. *Journal of Political Economy*, 119(4):721 – 765, 2011. URL http://ideas.repec.org/a/ucp/jpolec/doi10.1086-662628.html.

H. Flam and E. Helpman. Vertical product differentiation and north-south trade. *American Economic Review*, 77(5):810–22, December 1987. URL http://ideas.repec.org/a/aea/aecrev/v77y1987i5p810-22.html.

M. García-Santana and J. Pijoan-Mas. Small scale reservation laws and the misallocation of talent. Working papers, CEMFI, Dec 2010. URL http://ideas.repec.org/p/cmf/wpaper/wp2010_1010.html.

L. Garicano, C. Lelarge, and J. Van Reenen. Firm size distortions and the productivity distribution: Evidence from France. *American Economic Review*, 106(11):3439–3479, 2016.

E. Ghani, A. G. Goswami, and W. R. Kerr. Is India's manufacturing sector moving away from cities? Policy Research Working Paper Series 6271, The World Bank, Nov. 2012. URL http://ideas.repec.org/p/wbk/wbrwps/6271.html.

D. Gollin. Do taxes on large firms impede growth? Evidence from Ghana. Bulletins 7488, University of Minnesota, Economic Development Center, 1995. URL http://ideas.repec.org/p/ags/umedbu/7488.html.

N. Guner, G. Ventura, and X. Yi. Macroeconomic implications of size-dependent policies. *Review of Economic Dynamics*, 11(4):721–744, October 2008. URL http://ideas.repec.org/a/red/issued/07-73.html.

J. C. Hallak. Product quality and the direction of trade. *Journal of International Economics*, 68(1):238–265, January 2006. URL http://ideas.repec.org/a/eee/inecon/v68y2006i1p238-265.html.

J. C. Hallak and J. Sivadasan. Product and process productivity: Implications for quality choice and conditional exporter premia. Technical Report 1, 2013.

R. Hasan and K. R. Jandoc. The distribution of firm size in India: What can survey data tell us? *Asian Development Bank Economics Working Paper Series*, (213), 2010.

R. Hillberry and D. Hummels. Trade responses to geographic frictions: A decomposition using micro-data. *European Economic Review*, 52(3):527–550, April 2008. URL http://ideas.repec.org/a/eee/eecrev/v52y2008i3p527-550.html.

C.-T. Hsieh and P. J. Klenow. The life cycle of plants in india and mexico. *The Quarterly Journal of Economics*, 129(3):1035–1084, 2014.

C.-T. Hsieh and B. A. Olken. The missing 'missing middle'. *Journal of Economic Perspectives*, 28(3):89–108, 2014.

D. Hummels and P. J. Klenow. The variety and quality of a nation's exports. *American Economic Review*, 95(3):704–723, June 2005. URL http://ideas.repec.org/a/aea/aecrev/v95y2005i3p704-723.html.

L. Iacovone and B. Javorcik. Getting ready: Preparation for exporting. CEPR Discussion Papers 8926, C.E.P.R. Discussion Papers, Apr. 2012. URL http://ideas.repec.org/p/cpr/ceprdp/8926.html.

M. Kugler and E. Verhoogen. Prices, plant size, and product quality. *Review of Economic Studies*, 79(1):307–339, 2012. URL http://ideas.repec.org/a/oup/restud/v79y2012i1p307-339.html.

R. La Porta and A. Shleifer. The unofficial economy and economic development. NBER Working Papers 14520, National Bureau of Economic Research, Inc, Dec. 2008. URL http://ideas.repec.org/p/nbr/nberwo/14520.html.

D. Lagakos. Explaining cross-country productivity differences in retail trade. *journal of political economy*, 124(2):579–620, 2016.

I. Little, D. Mazumdar, and J. M. Page Jr. *Small Manufacturing Enterprises: A Comparative Analysis of India and Other Economies*. NY: Oxford U. Press, 1987.

N. V. Loayza. The economics of the informal sector: A simple model and some empirical evidence from latin america. *Carnegie-Rochester Conference Series on Public Policy*, 45(0):129 – 162, 1996. ISSN 0167-2231. doi: http://dx.doi.org/10.1016/S0167-2231(96)00021-8. URL http://www.sciencedirect.com/science/article/pii/S0167223196000218.

N. V. Loayza, A. M. Oviedo, and L. Serven. The impact of regulation on growth and informality - cross-country evidence. Policy Research Working Paper Series 3623, The World Bank, May 2005. URL http://ideas.repec.org/p/wbk/wbrwps/3623.html.

N. V. Loayza, L. Serven, and N. Sugawara. Informality in latin america and the caribbean. Policy Research Working Paper Series 4888, The World Bank, Mar. 2009. URL http://ideas.repec.org/p/wbk/wbrwps/4888.html.

B. R. Mandel. Heterogeneous firms and import quality: evidence from transaction-level prices. International Finance Discussion Papers 991, Board of Governors of the Federal Reserve System (U.S.), 2010. URL http://ideas.repec.org/p/fip/fedgif/991.html.

K. Manova and Z. Zhang. Export prices across firms and destinations. *The Quarterly Journal of Economics*, 127(1):379–436, 2012. URL http://ideas.repec.org/a/oup/qjecon/v127y2012i1p379-436.html.

D. F. McFadden. *Conditional Logit Analysis of Qualitative Choice Behavior*, pages 105–142. Academic Press: New York, 1974.

D. Mitra and V. Trindade. Inequality and trade. *Canadian Journal of Economics*, 38(4):1253–1271, November 2005. URL http://ideas.repec.org/a/cje/issued/v38y2005i4p1253-1271.html.

S. Nataraj. The impact of trade liberalization on productivity: Evidence from India's formal and informal manufacturing sectors. *Journal of International Economics*, 85(2):292–301, 2011. URL http://ideas.repec.org/a/eee/inecon/v85y2011i2p292-301.html.

M. Poschke. The firm size distribution across countries and skill-biased change in entrepreneurial technology. *American Economic Journal: Macroeconomics*, 10(3):1–41, 2018.

D. Restuccia and R. Rogerson. Misallocation and productivity. *Review of Economic Dynamics*, 16 (1):1–10, January 2013. URL http://ideas.repec.org/a/red/issued/13-0.html.

P. K. Schott. Across-product versus within-product specialization in international trade. *The Quarterly Journal of Economics*, 119(2):646–677, May 2004. URL http://ideas.repec.org/a/tpr/qjecon/v119y2004i2p646-677.html.

D. Scur, S. Ohlmacher, J. V. Reenen, M. Bennedsen, N. Bloom, A. Choudhary, L. Foster, J. Groenewegen, A. Grover, S. Hardeman, L. Iacovone, R. Kambayashi, M.-C. Laible, R. Lemos, H. Li, A. Linarello, M. Maliranta, D. Medvedev, C. Meng, J. M. Touya, N. Mandirola, R. Ohlsbom, A. Ohyama, M. Patnaik, M. Pereira-Lopez, R. Sadun, T. Senga, F. Qian, and F. Zimmermann. The international empirics of management. *Proceedings of the National Academy of Sciences*, 121(45):e2412205121, 2024.

A. Shaked and J. Sutton. Product differentiation and industrial structure. *The Journal of Industrial Economics*, pages 131–146, 1987.

J. Sutton. *Sunk costs and market structure: Price competition, advertising, and the evolution of concentration*. MIT press, 1991.

K. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

J. R. Tybout. Manufacturing firms in developing countries: How well do they do, and why? *Journal of Economic Literature*, 38(1):11–44, March 2000. URL http://ideas.repec.org/a/aea/jeclit/v38y2000i1p11-44.html.

G. Ulyssea. Firms, informality, and development: Theory and evidence from brazil. *American Economic Review*, 108(8):2015–2047, 2018.

E. Verhoogen. Firm-level upgrading in developing countries. *Journal of Economic Literature*, 61 (4):1410–1464, 2023.

E. A. Verhoogen. Trade, quality upgrading, and wage inequality in the mexican manufacturing sector. *The Quarterly Journal of Economics*, 123(2):489–530, 05 2008. URL http://ideas.repec.org/a/tpr/qjecon/v123y2008i2p489-530.html.

Appendix

A. Data

This paper uses data from the following surveys from India:

1. Annual Survey of Industries of 2005-06, 1989-90, 1994-95, 2000-01, and 2009-10
2. Survey of Unorganized Manufacturing of 2005-06, 1989-90, 1994-95, 2000-01, and 2010-11
3. Consumer Expenditure Survey of India of 2003 and 2004-05
4. Employment-Unemployment Survey of India of 2004-05

This section provides some more details regarding these surveys. It also provides a brief description of the County Business Database of the US.

A.1. Annual Survey of Industries

The Annual Survey of Industries (ASI) is conducted by the Central Statistics Office of the Government of India every year. It covers all factories registered under Sections 2m(i) and 2m(ii) of the Factories Act, 1948, that is, those factories employing ten or more workers using power, and those employing twenty or more workers without using power.

The paper primarily uses data from the 2005-06 ASI (as the SUM was also conducted in 2005-06) which reports data for the financial year ending March 2006. The geographical coverage of the 2005-06 ASI was all of India except the states of Arunachal Pradesh, Mizoram, and Sikkim and the Union Territory of Lakshadweep.

ASI 2005-06 uses the National Industrial Classification (NIC) 2004 (which is closely based on International Standard of Industrial Classification (ISIC) Rev 3.1) to classify economic activity. For all the analysis done in the paper, we restrict the sample to plants which report a 2-digit NIC between 15 to 36 as this constitutes the manufacturing sector and matches the coverage of the SUM. Furthermore, for some of the figures, attention is restricted to 15 large Indian states.[39]

The main variables used in the paper are total employment level of the plant, and details regarding the products produced and inputs used (quantities and rupee values) by each plant.

Plants report the average number of employees working in the plant for seven different categories, namely: male workers employed directly, female workers employed directly, child workers employed directly, workers employed through contractors, supervisory and managerial staff, other employees, and unpaid family workers. The size of the plant (total employment) is defined as the sum across all these categories.

All plants report the output they produce using a standardized classification of products called the ASICC product classification. The ASICC has about 5,500 product categories. Plants can report up to ten main products produced in terms of this ASICC classification. Each product category has an associated standardized unit (kilograms, tonnes, numbers, etc) in terms of which the quantity produced is to be reported. Plants also report the total value of production before taxes and distribution costs for each product which can be combined with the information on quantity produced to infer per-unit prices. As all plants are supposed to report quantities in standardized units, the prices inferred should be comparable for all plants producing the same product. However, there seems to be some misreporting in units and this issue is discussed further in Section F. The same commodity classification and units are used to report the quantity and value of materials inputs used.

In addition to the 2005-06 ASI, Section 5.2 also uses data on level of employment of each plant from four other years of the ASI, namely 1989-90, 1994-95, 2000-01, and 2009-10. As with the 2005-06 survey, the broadest definition of employment was used for all years which included part-time workers and unpaid workers. Arunachal Pradesh, Mizoram, Sikkim, and Lakshadweep were excluded from the sample for all years as these states were not covered in the ASI for many of the waves. Different years of the survey used different industrial classifications (NIC 1987, NIC

[39] The main states included are: Andhra Pradesh, Bihar, Gujarat, Haryana, Himachal Pradesh, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Orissa, Punjab, Rajasthan, Tamil Nadu, Uttar Pradesh, and West Bengal. Three Indian states were split into two in 2000. In order to maintain comparability with some of the time-series results in Section 5.2 and D, the pre-split definition of states is used throughout the paper.

1998, NIC 2004, and NIC 2008). We created a concordance across these different classifications and only industries which corresponded to 2-digit NIC 2004 between 15 and 36 were included in the sample.

Table 17 reports the number of observations, estimated number of establishments (using sampling weights provided by the ASI), and the estimated total number of workers employed based on the ASI for all five years that are used in the paper.

More details about the ASI can be found on the website of the Ministry of Statistics and Programme Implementation, Government of India (http://mospi.nic.in/).

A.2. Survey of Unorganized Manufacturing

The Survey of Unorganized Manufacturing (SUM) is conducted by the National Sample Survey Office (NSS) of India. The coverage of the survey includes all manufacturing enterprises not registered under Sections 2m(i) and 2m(ii) of the Factories Act, 1948. The SUM is usually conducted every five years. The last five waves were done in 1989-90, 1994-95, 2000-01, 2005-06, and 2010-11.

The paper primarily uses data from the 2005-06 SUM (62^{nd} Round of the NSS). The survey period was from July 2005 to June 2006.[40]

The geographical coverage of the survey was comprehensive and included all States and Union-Territories of India, with only Leh and Kargil districts of Jammu and Kashmir and a few remote villages in Nagaland and Andaman and Nicobar Islands being excluded. The states of Arunachal Pradesh, Mizoram, and Sikkim and Union Territory of Lakshadweep were dropped to maintain comparability with the coverage of the 2005-06 ASI.

Like the ASI, the SUM 2005-06 uses the National Industrial Classification (NIC) 2004 to classify economic activity. For all the analysis done in the paper, we restrict the sample to plants which report a 2-digit NIC between 15 to 36. Furthermore, for some of the figures, attention is restricted to 15 large Indian states.

The main variables used in the paper are the total employment, and details regarding the products

[40] Note that there is a three month difference in coverage period between the ASI and SUM.

produced and inputs used (quantities and rupee values) by each plant.

Plants report the average number of employees working in the plant for the reference period for which the data is collected (for most plants this was one month). The plants reported the average number of hired workers, working owners, and other workers that they employed on a part-time and full-time basis. Like the ASI, the broadest definition of employment is used with the size of the plant (total employment) being defined as the sum across all these categories.

All plants report the output they produce and material inputs consumed using the same standardized classification of products as is used by the ASI plants. Plants can report up to five main products produced in terms of this product classification. However, unlike the ASI, SUM plants can choose the units in which they are reporting quantities and prices. For example, all ASI plants which produce matchsticks must report quantities in kilograms. However, different SUM plants report quantities and prices of matchsticks in different units including kilograms, tonnes, and numbers (number of matchsticks). We concord units across the two surveys when combining the two surveys. If the same product is being reported in different units which are simple scalar multiples of each other (kilograms and tonnes for example), then we convert the units so that all quantities and prices are being measured in the same unit i.e., divide quantities and prices of all SUM units which report quantities of matchsticks in tonnes by 1000, to get per kilogram prices which are comparable to ASI prices. However, if a SUM plant is reporting the output of matchsticks in numbers, then it is not possible to make this comparable to the the ASI plants which are reporting in kilograms. In such cases, the SUM products are treated as a separate product category.

In addition to the 2005-06 SUM, Section 5.2 also uses data on level of employment of each plant from four other years of the SUM, namely 1989-90, 1994-95, 2000-01, 2005-06, and 2010-11. As with the 2005-06 survey, the broadest definition of employment was used for all years which included part-time workers and unpaid workers. The same sampling on states and industries was done as in the ASI.

Table 17 reports the sample size, number of establishments (using sampling weights provided by the SUM), and the total number of workers employed based on the SUM for all five years that are used in the paper.

More details about the SUM can be found on the website of the Ministry of Statistics and Programme Implementation, Government of India (http://mospi.nic.in/).

A.3. Consumer Expenditure Surveys

The National Sample Survey Office of India (NSS) conducts an annual Consumer Expenditure Surveys (Schedule 1.0) in India. From 1972-73, the NSS started a quinquennial series in which every five years, it conducts a survey with a sample size which is about four times larger than the annual survey.

The paper uses data mainly from the 2004-05 (61^{st} Round of the NSS) Consumer Expenditure Survey which was part of the quinquennial series and interviewed about 125,000 households. The geographical coverage of the survey was comprehensive and included all States and Union-Territories of India, with only Leh and Kargil districts of Jammu and Kashmir and a few remote villages in Nagaland and Andaman and Nicobar Islands being excluded. The survey period was from July 2004 to June 2005.

The Consumer Expenditure Surveys of 2004-05 asks households to report the value of consumption for 339 different goods. Households report quantities and rupee values separately for 209 goods, which can be used to compute prices for these goods. 156 of these 209 goods are food items, 10 fall under the "fuel and light" category, another 24 are clothing and footwear, while the remaining are durables.

For food items, households report consumption out of home production (quantities and imputed rupee values) and total consumption (which includes home production and market purchases). The price computed divides total value of consumption by total quantity consumed, thus averaging across home and market consumption.

The reference period for consumption of all food items is 30 days, i.e., households report quantity consumed and rupee values for food consumption for the last 30 days. For clothing and footwear categories, households report consumption for a reference period of 30 days as well as 365 days. The 365 day reference period for these categories is used as many households report zero purchases for these items for the 30 day reference period but positive amounts for the 365 day reference period.

Table 14 uses data from the 2003 (59^{th} Round) Consumer Expenditure Survey which was not part of the quinquennial series and interviewed about 41,000 households. The geographical coverage of the 2003 survey was similar to the 2004-05 survey. The survey period was from January 2003 to December 2003. The consumption items recorded across the two surveys were also very similar with only a few minor differences.

Table 18 reports some summary statistics for the Consumer Expenditure Survey of 2004-05. It reports the number of items and share of expenditure for five broad expenditure heading and also the share of expenditure within the heading for which prices could be computed. The summary statistics for the 2003 survey are very similar and are not reported.

More details about the dataset can be found on the website of the Ministry of Statistics and Programme Implementation, Government of India (http://mospi.nic.in).

A.4. Employment-Unemployment Survey

The National Sample Survey Office of India (NSS) conducts an Employment-Unemployment Survey (Schedule 10.0) as part of its quinquennial series. This paper uses the Employment-Unemployment Survey of 2004-05 (61^{st} Round of the NSS). The geographical coverage of the survey was comprehensive and included all States and Union-Territories of India, with only Leh and Kargil districts of Jammu and Kashmir and a few remote villages in Nagaland and Andaman and Nicobar Islands being excluded. The survey period was from July 2004 to June 2005. In this survey it interviews about 125,000 households (about 600,000 individuals).

The survey asks all the individuals in the household to report demographic characteristics like age, education etc. It also asks individuals to report the main industry in which they work, the size of establishment in which they work, and the wage they earned in the last week. To maintain comparability with the production surveys, only individuals who report a 2-digit NIC 2004 between 15 and 36 are used.

The main variables used from this survey are the education level of individuals and the size category of the establishment in which they work.

The survey asks individuals to report the level of general education that they have achieved.

The possible responses are: illiterate, literate but not through formal schooling, primary, middle, secondary, higher secondary, diploma/certificate course, graduate, and post graduate or above. For the purpose of the model, a person was defined as skilled if he or she had finished at least secondary education (Grade ten).

Individuals were also asked to report the size category of the establishment in which they worked. They could report one of the following options: establishment of size less than 6, between 6 and 9, between 10 and 19, 20 or greater, and unknown size

The calibration of the wage premium in Section 4.1 also makes use of wage data from this survey. More details regarding the construction of this variable along with the Mincerian regression results are provided in Section G.

More details about the Employment-Unemployment Survey can be found on the website of the Ministry of Statistics and Programme Implementation, Government of India (http://mospi.nic.in/).

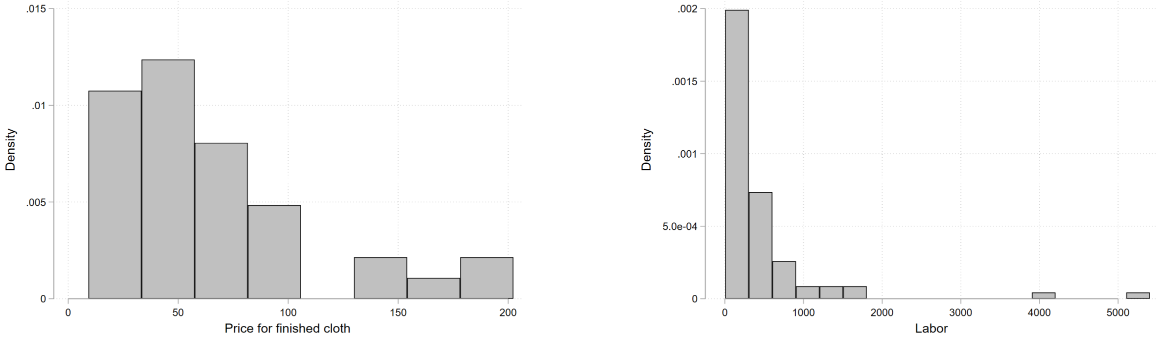
A.5. County Business Patterns Database (US)

The County Business Patterns Database maintained by the US Census Bureau provides level of employment for each 6-digit NAICS for each US county. The employment level is as on the week of March 12^{th} of that year.

The paper uses the 2006 release of the data. For many industry-county cells the exact level of employment is not reported. Instead, the dataset reports an employment size class for that cell. In these cases, the employment in the cell is assigned the midpoint of the size class reported. For example, if a NAICS-County cell reports employment in the size class *'B'* which represents 20-99 employees, then the cell is assigned an employment level of 60.

The data can be downloaded from http://www.census.gov/econ/cbp/.

Figure 14: Distribution of prices and employees for firms that produce “Finished Cotton Cloth”



This figure shows the histogram of average price charged (left) and the number of employees (right) by each firm that produces “finished cotton cloth” in the Annual Survey of Industries.

B. Case study: Finished Cotton Cloth

To build intuition for why larger firms, on average, are more likely to sell higher priced products, we present a deep dive on firms that produce “finished cotton cloth.” To ensure comparability and a more detailed analysis, we restrict attention to the firms that are reported in the ASI. First, we detail the observed variation in this industry; second, establish that larger firms charge higher prices, on average, and finally, explain why larger firms have a comparative advantage in producing more expensive goods.

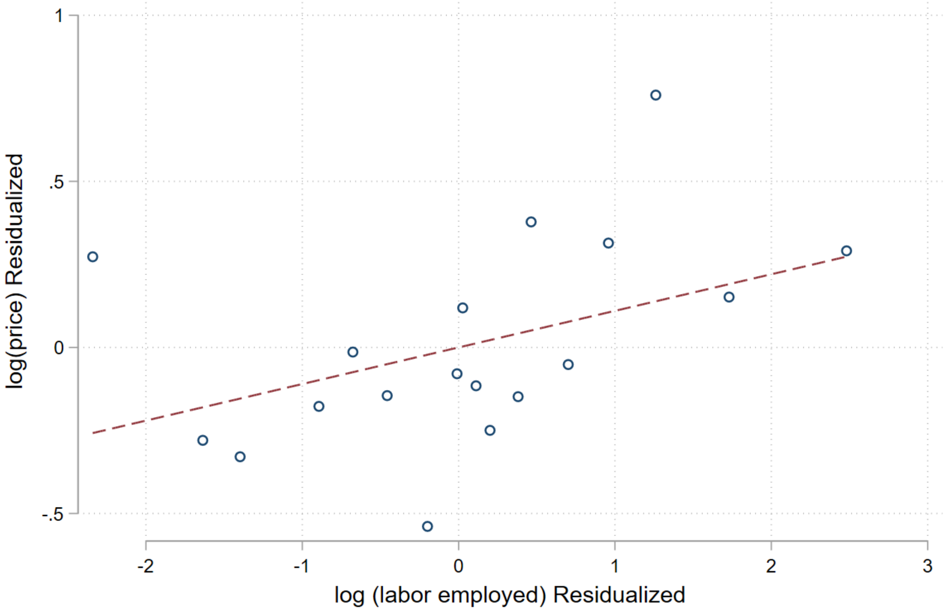
There is substantial variation in the price charged for a meter of finished cotton cloth (left panel of figure 14). The median firm charges an average of 51 Indian rupees, but a firm at the 10th percentile charges only 22 rupees (equivalent to US $0.50 in 2006), and the 90th percentile firm charges 140 rupees (equivalent to US $3.16)—a difference of more than 500 percent. Alongside this variation in the price charged, the variation in firm size is substantially more (right panel of figure 14). The median firm employs 197 people but there is substantial mass of firms that are very small with 20 percent of firms with less than 35 people.[41] At the same time, there are two very large firms with close to 4000 and 5000 employees each.

Consistent with the main argument in our paper, we find substantial correlation between these

[41] Recall that the ASI omits the smallest firms since the ASI only includes manufacturing plants employing twenty or more workers and not using electricity or employing ten or more workers and using electricity.

two measures. Figure (15) shows that, on average, larger cloth firms charger a higher price (we plot the residuals for the natural logarithm of employees in a firm and for the natural logarithm of product price after controlling for the interaction of state, and rural fixed effects). Moreover, consistent with our earlier argument, on average, larger firms also employ significantly more capital (figure 16) suggesting that higher quality products require more capital investment.

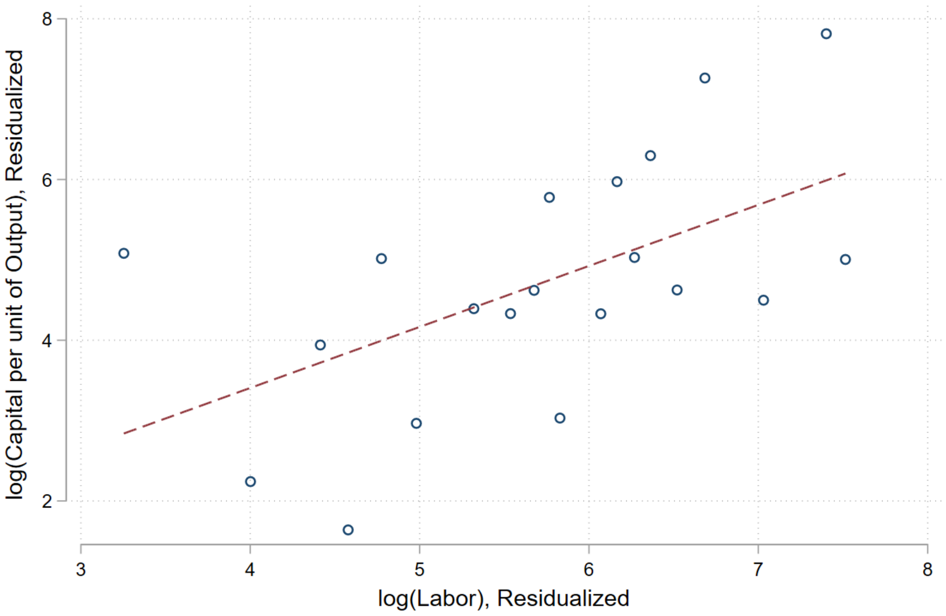
Figure 15: Finished cotton cloth: Relationship between size of firm and prices charged



This figure shows the binned scatterplot and line of best fit for the relationship between the price charged and the number of employees in each firm in the finished cotton cloth industries using the ASI. We plot the residuals for log ("price") and log ("labor employed") after controlling for state and rural fixed effects.

To provide more color beyond what is available in the survey data, we can examine the largest cloth manufacturing clothing companies in India. The largest primarily textile company with financial information on Dun&Bradstreet (a commercial data provider with a large Indian presence) is Vardhman Textiles Limited with more than $1 billion in sales (2024). Given the size of the company, it retail products at many different price points. However, a close examination of a single line of products—men's formal shirts—suggests a higher-end product with a back-of-the-envelope calculation for the average shirt price of $8.40 (which is likely the sale price to other intermedi-

Figure 16: Finished cotton cloth: Relationship between size of firm and capital stock



This figure shows the binned scatterplot and line of best fit for the relationship between the price charged and the number of employees in each firm in the finished cotton cloth industries using the ASI. We plot the residuals for log (“price”) and log (“labor employed”) after controlling for state and rural fixed effects.

aries, the final retail price will be considerably higher).[42] Other evidence that is consistent with a higher-end product:

- Large capital expenditures, the latest available audited accounts lists US $450 million in fixed assets (“property, plant, and machinery”). Therefore, it has substantial capital relative to total sales (over 30 percent). Moreover, the machinery suggests high-quality equipment with many of these machines imported from more economically advanced countries.

- The company boasts of several textile certifications on its website, such as OEKO-Tex Standard 100. This is consistent with the results in Verhoogen [2008] that proxies product quality using international certifications.[43]

[42] Specifically, the 2023-2024 annual report details revenue of INR 1.12 billion and 1.5 million garments produced.

[43] The OEKO-Tex Standard 100 is a label for textiles tested for harmful substances, https://www.oeko-tex.com/en/our-standards/oeko-tex-standard-100. Items bearing the STANDARD 100 label is certified as having passed safety tests for the presence of harmful substances.

- Customers include high-end global brands (such as "United Colors of Benetton") and Indian domestic brands such as "Raymond."

C. Could higher opportunity cost cause richer households to pay more for the same good?

In section (2.1) we showed that richer households buy goods at a higher unit price which is consistent with the hypothesis that they buy higher quality goods. However, as documented by Aguiar and Hurst [2007], households might be paying different prices for similar goods because households with higher opportunity cost of time tend to shop around less for lower prices. If richer households have a higher opportunity cost of time, then the findings in Table 1 might be a result of less time spent shopping by richer households and not because of purchase of higher quality goods.[44]

The 2003 Consumer Expenditure Survey asked each individual in the household the main activity they were engaged in (whether they were employed, studying, attending to domestic duties, retired etc).[45] We use this to construct a proxy variable which takes value 1 if the household has at least one member between the age of 15 and 70 who is only attending to domestic duties or is retired, and 0 otherwise.[46] We interpret households with a non-worker present as households with low opportunity cost of time and include this variable as a control in the regressions. Column 1 of Table 14 repeats the regression from Column 1 of Table 1, but with the 2003 data instead of the 2004-05 data. Column 2 of Table 14 now adds the measure of "non-worker present" as an additional control. Although the coefficient on the "non-worker present" variable is positive, the key point is that the coefficient of per-capita expenditure does not change substantially. Column

[44] For developing countries, there is evidence that poorer households might in fact be paying more for the same product as opposed to rich households which would imply that the estimates for β are a lower bound for the quality-income relation. For example, Attanasio and Frayne [2006] find that poor people in rural Columbia are less likely to avail of bulk discounts and thus end up paying more for the same product as compared to richer households.

[45] Unfortunately, the 2004-05 Consumer Expenditure Survey does not ask this question so this exercise cannot be conducted using the same data used in Table 1. The 2003 survey has only one fourth the number of households as the 2004-05 survey. However, the point estimates for the elasticity of price with respect to per-capita expenditure (β) are quite similar across the two surveys.

[46] Table 19 in the appendix lists the possible responses for the question regarding main activity of the individual. People who reported codes 92, 93, 94, or 97 were classified as non-workers.

Table 14: Household Regressions: Controlling for Opportunity Cost of Time

Dependent Variable: log(price)						
	(1)	(2)	(3)	(4)	(5)	(6)
log(per-capita expenditure)	0.102***	0.102***	0.094***	0.105***	0.104***	0.099***
	(0.0010)	(0.0010)	(0.0015)	(0.0029)	(0.0029)	(0.0036)
non-worker present		0.020***	-0.059***		0.017***	-0.065**
		(0.0011)	(0.0113)		(0.0027)	(0.0318)
(non-worker present)*pce			0.012***			0.011**
			(0.0017)			(0.0045)
Household Size	All	All	All	1 and 2	1 and 2	1 and 2
Observations	1,822,762	1,822,762	1,822,762	219,390	219,390	219,390
Number of products	169	169	169	169	169	169
Clusters	41,013	41,013	41,013	6,161	6,161	6,161

Notes: The data is from the Consumer Expenditure Survey of 2003. Column 1 reports results for the regression of log of price paid by households for different goods on log of per-capita expenditure (replicating Column 1 of Table 1). Column 2 includes a control for opportunity cost of time, namely a variable which takes value 1 if there is at least one non-working adult in the household. Column 3 also includes the interaction of this variable with per-capita expenditure. Columns 4, 5, and 6 repeat the specifications in 1,2, and 3 but restrict the sample to households of size 1 and 2 only. Regressions include the triple interaction of good interacted with state and rural-urban fixed effect. Standard errors are clustered at the household level. ***p<0.01, **p<0.05.

3 also includes the interaction of the "non-worker present" variable with per-capita expenditure and this does not change the results substantially either. Columns 4, 5, and 6 repeat the regressions from columns 1, 2, and 3 respectively, but restrict the sample to include households with one or two members only. This controls for the fact that larger households are more likely to have non-working adults. Again, the coefficient on per-capita expenditure does not change substantially when including the "non-worker present" variable as a control.

D. Inter-State Trade

The model presented in the main paper implicitly assumed that each state in India can be treated as a closed economy and that differences in income levels across states translate into differences in demand and in the size distribution at the state level. How would the possibility of inter-state trade affect the hypothesis presented in the paper?

A potential confounding effect of inter-state trade could come through the location choice of large plants. For example, if the richer states are more suited for operating large plants (due to availability of skilled labor, better labor laws etc), then all the larger plants might choose to locate in these states and ship their goods to the poor states. In this case, the fact that richer states have a smaller share of employment in small plants would not reflect differences in demand across states but rather just the spatial location choice of large plants.

To address this concern, it would be ideal to have a measure of inter-state trade flows (similar to the Commodity Flow Survey in the US) to see how important this channel could be. Unfortunately, data on extent of inter-state trade is not collected in India. Here we provide indirect evidence to suggest that inter-state trade is not the principal driver of cross-state employment differences.

Firstly, transportation costs in developing countries are often very high which makes it harder for plants to transport goods over large distances to poorer states. Atkin and Donaldson [2012] show that intranational transportation costs in two African countries are seven to fifteen times larger than similar estimates for the US. Furthermore, Hillberry and Hummels [2008] show that even in the US, manufacturing production is extremely localized with local shipments volumes being three times larger than shipments to more distant locations. This suggests that local demand is likely to be an predominant determinant of the the size distribution in any region, especially in developing countries.

Furthermore, if inter-state trade is driving the cross-state employment differences, then we would expect more tradable industries to exhibit larger differences in share of employment in small plants across states as compared to less tradable industries. To test this fact, we construct two measures of tradability (within manufacturing) at the 3-digit level of the National Industrial Classification (NIC) of 2004.[47] These are:

1. Herfindahl index of geographical concentration in the US: The County Business Patterns Database of 2005 released by the United States Census Bureau provides information regarding the number of people working in each 6-digit industry of the North American Industry

[47] Economic activity in India is classified according to the National Industrial Classification (NIC) which closely follows the United Nation's International Standard Industrial Classification (ISIC). Details regarding different NIC revisions and the concordance used between them are given in Appendix E.3.

Classification System (NAICS) for each county in the US.[48] As the tradability index is to be applied to the Indian industry classification, we first create a concordance from 6-digit NAICS to 3-digit NIC and then construct a Herfindahl Index (H-index) of geographical concentration of each 3-digit NIC across US counties.[49]

The H-index is defined as

$$H_i = \sum_{c=1}^{C} \left(sh_{i,c}^{L}\right)^2,$$

where *'i'* indexes industry (according to NIC), *'c'* indexed counties, and $sh_{i,c}^{L}$ represents the share of industry *'i'* employment which is in county *'c'*. The H-index for industry *'i'* is simply the sum across counties of the square of the share of the industries employment which is present in county *'c'*. The industries which are highly concentrated in a few counties in the US (have a high value for Herfindahl index) are considered to be tradable industries while industries which have employment spread over lots of counties (have a low value for the Herfindahl index) are considered non-tradable industries. This measure for tradability of an industry based on US levels of concentration is applied to India.

2. Degree of international trade in India: For each 3-digit NIC in the manufacturing sector, we construct a measure of the degree of international trade carried out in the industry as a share of domestic production. In particular, we define this measure of international trade as the exports plus imports in that industry as a share of gross production of that industry carried out by domestic plants in 2005-06. The data for exports and imports for India is taken from the website of the Department of Commerce, Government of India.[50] The imports and exports data is not at the industry level but rather classified according to the Harmonized Commodity Description and Coding System (HS) product classification. This is converted to 3-digit NIC using the products to industry concordance developed by World Integrated Trade Solutions (WITS).[51] The data on gross domestic production for each industry is computed

[48] The data can be found at http://www.census.gov/econ/cbp/. The exact number of people in many industry-county cells is masked. Instead, the dataset reports an employment size class for that cell. In these cases, the employment in the cell is assigned the midpoint of the size class reported.

[49] The concordance from 6-digit NAICS and 3-digit ISIC Rev 3.1 was based on the Census Bureau's concordance file available at http://www.census.gov/eos/www/naics/concordances/concordances.html. ISIC Rev 3.1 to NIC 2004 is a one to one correspondence at the 3-digit level. Appendix E.1 has more details regarding the concordance.

[50] The data is available from http://commerce.nic.in/eidb/default.asp.

[51] WITS is based on a collaboration of the World Bank with UNCTAD, WTO and other international organization

Table 15: Size Income Relation Across States for Tradables vs. Non-tradables

Dependent Variable: share of employment in <=5 in industry 'i', state 's', time 't'				
	(1)	(2)	(3)	(4)
log(per-capita SNDP)*tradability	0.068* (0.0351)	0.052 (0.0394)	-0.010 (0.0469)	0.000 (0.0498)
Index Cutoff	H-index Median	H-index Quartile	Exp-Imp Median	Exp-Imp Quartile
Observations	3,885	1,826	3,899	1,959

Notes: The data is from five rounds of the ASI and SUM. The table reports regression results for the share of employment in plants of size 5 or less in industry 'i' in state 's' at time 't' on log per-capita state NDP interacted with a dummy which takes value 1 if industry 'i' is classified as a tradable industry. Column 1 classifies an industry as tradable if the Herfindahl Index across US counties for the industry was above the median of Herfindahl Indexes, and non-tradable if it was below the median. Column 2 uses top and bottom quartiles of the Herfindahl Index as cutoffs. Column 3 and 4 use the tradability index based on Indian exports and imports and uses the median and the top and bottom quartiles as cutoffs respectively. All regressions include fixed effects for industry interacted with time and state interacted with time. Each observation is weighted by the share of observations in the state-industry cell out of the total observations in the ASI and SUM combined for the given year. Standard errors are clustered at the state level. *p<0.1.

by combining the ASI and the SUM. Industries in which international trade is a large percent of domestic production are considered to be more tradable.

Table 21 in the appendix lists the 3-digit industries which lie above and below the median value of the two indexes of tradability. The two measures of tradability are weakly positively correlated with the rank correlation coefficient between them being 0.25.

We run regressions of the form

$$sd_{i,s,t} = \alpha_{i,t} + \alpha_{s,t} + \gamma \ln(SNDP_{s,t}) * tradabilty_i + \varepsilon_{i,s,t} \tag{14}$$

where $sd_{i,s,t}$ is the share of employment in plants of size five or less in industry *'i'* in state *'s'* at time *'t'*, $SNDP_{s,t}$ is the per-capita NDP of state *'s'* at time *'t'*, and $tradabilty_i$ is a dummy variable which takes value 1 if an industry is classified as tradable. $\alpha_{i,t}$ represents fixed effects for industry interacted with time and it controls for the fact that different industries might have different average

associated with international trade data. More details regarding the concordance can be found in Appendix E.2.

levels for the share of employment in small plants. $\alpha_{s,t}$ represents fixed effects for state interacted with time and controls for the fact that rich states on average have a lower share of employment in small plants.

The coefficient of interest is γ, the coefficient on the interaction of state per-capita income and the tradability dummy. A positive γ implies that the relationship between the share of employment in small plants and log of per-capita income across states is stronger for non-tradables. This is because the share of employment in small plants and per-capita NDP are negatively related and therefore a positive interaction term implies that the slope for tradable industries is less negative compared to non-tradables. Therefore, a positive value of γ is supportive of the view that inter-state trade is not a major driving force behind the size distribution of plants across states.

An industry is classified as tradable if the tradability index for the industry lies above the median (or in the top quartile) of the index across industries. Data for five waves of the SUM is combined with the corresponding year of the ASI (1989, 1994, 2000, 2005, and 2010). Only the fifteen large Indian states mentioned in footnote 39 are included as the smaller states often have no observations for many industries as the 3-digit level.

Table 15 reports results for equation (14) for both the measures of tradability. Each observation is weighted by the share of observations in the state-industry cell out of the total observations in the ASI and SUM combined for the given year.[52] Column 1 uses the Herfindahl index and classifies an industry as tradable if its Herfindahl Index is above the median value of the Herfindahl Index across industries. The coefficient on the interaction of per-capita NDP and the tradability index is positive and marginally significant at the 10 percent level. Column 2 classifies an industry as tradable if it is in the top quartile in terms of the Herfindahl Index and non-tradable if it is in the bottom quartile. The results are very similar to the first column. Columns 3 and 4 use the median and quartile of the tradability measure based on exports and imports in India. The point estimates of the coefficient on the interaction of per-capita NDP and the tradability index is much smaller in absolute value and statistically insignificant.

[52] This weighting scheme accounts for the fact that the size distribution variable (dependent variable) for some industry-state pairs is based on a lot fewer observations than other cells, and are therefore likely to be measured with less precision. Table 22 in the appendix reports results when all observations are weighted equally. Table 23 reports results when industrial categories which are residual categories (industry categories with descriptions which include words like “not elsewhere covered” or “others”) are excluded.

The results in Table 15 suggest that the size-income relationship across states is not stronger for tradable industries as compared to non-tradable industries.

E. Inter-State Trade: Concordances

E.1. NAICS 2002 to NIC 2004 Concordance for Herfindahl Index

Section D uses a Herfindahl Index of employment concentration across US counties as a measure of tradability for industries in India. While the US County Business Patterns Database uses NAICS to classify economic activity, the Herfindahl Index needs to be based on the Indian classification of economic activity (NIC 2004) for it to be applied using Indian data. In order to construct this index, We created a concordance between 6-digit NAICS 2002 and 3-digit ISIC Rev 3.1 (the classification used by the ASI and SUM is the NIC 2004, which is a one to one match to ISIC Rev 3.1 at the 3-digit level).

The concordance between 6-digit NAICS and 3-digit ISIC Rev 3.1 was based on the Census Bureau's concordance file available at http://www.census.gov/eos/www/naics/concordances/concordances.html. Although this file gives a many to many concordance, this was reduced to a one to one concordance by taking the 3-digit ISIC which was the closest fit for each 6-digit NAICS.

Of the 59 3-digit ISIC industries in the manufacturing sector, three industries (182, 231, and 233) were not represented in this concordance i.e. none of the 6-digit NAICS industries were mapped into these 3-digit ISIC industries. These industries employed only 0.16 percent of the total manufacturing workforce in India in 2005. These industries are dropped for all the analysis which uses the Herfindahl Index.

E.2. HS Product Classification to NIC 2004 Concordance for Export-Import Index

Section D also uses a measure of international trade as a proportion of domestic production in India at the 3-digit level for NIC 2004. The export and import data for India was not at the industry level but rather at the product level using the Harmonized Commodity Description and Coding System (HS). The World Integrated Trade Solution (WITS) provides a one to one concordance from HS 2002 to ISIC Rev 3. WITS is based on a collaboration of the World Bank with UNCTAD, WTO and other international organization associated with international trade data.[53]

Two NIC 04 industries (223 and 273) were not represented in this concordance i.e. none of the HS codes mapped into these industries. These industries employed only 0.37 percent of the total manufacturing workforce in India in 2005. These industries are dropped for all the analysis which uses the Export-Import Index. Furthermore, industry 233 (nuclear fuel) had some imports in the trade data but no local production in India. This industry was also dropped.

E.3. Concordances Across Different NIC Revisions

Different years of the ASI and SUM use different revisions of the NIC. The 1989 and 1994 surveys use NIC87, the 2000 surveys uses NIC98, the 2005 surveys use NIC04 and the 2010 surveys use NIC08. We create a concordance from the different NIC revisions to NIC04 at the 3-digit level as the tradability indexes are constructed for NIC04. The concordances were based on official concordance tables which can be found at http://mospi.nic.in under the “Economic and Social Classification Heading”.

The NIC04 industries 341 (Manufacturing of motor vehicles) and 342 (Manufacture of bodies of motor vehicles, trailers, and semi-trailers) cannot be separately identified in NIC87. Hence, these two industries are merged into one industry group for all the tradability regressions.

F. Units Misreporting Problem in the ASI

As mentioned in footnote 10, there seems to be a misreporting of units and quantities in the ASI. We discuss an example here to clarify the problem. ASICC code 11401 stands for “milk”. All plants who report that they produce milk are supposed to report the quantity produced in kiloliters (1000 liters) which should mean that when we divide rupee values by the quantity, then it should

[53] The concordance can be found at http://wits.worldbank.org/wits/product_concordance.html.

yield the price of milk that the plant charges in kiloliters. Figure 17 plots the log of price charged for milk by different plants in the ASI against log of the number of employees in the plant. As can be seen, the log of the price charged by most plants is about ten. However, there is a group of plants who report a price which is seven log points lower or about 1000 times lower $(exp(7) = 1096)$. This is clearly a case of some plants reporting quantities in liters instead of kiloliters which makes the price computed a price per liter.

Such misreporting can potentially bias the results from regressions of price on size if larger plants are more likely to misreport quantities in terms of larger units. To account for this problem, we manually go through about a 1000 product categories to see which product categories suffer from this problem. We split products which suffer from this problem into two separate product categories based on a sensible price cutoff (for the milk example, all plants charging a log price greater than six were placed in a different product category).[54] As a different product fixed effect is allowed for this new product category, the regressions control for the price level differences arising from misreporting of quantity units. However, the clustering when computing standard errors does not treat the new product category as a separate category which is why the number of product fixed effects exceed the number of clusters in these regressions. Table 24 compares the results when the units correction described above is implemented versus when it is not implemented. Column 1 is the same as the first column of Table 3 (it corrects for the units problem). Column 2 repeats the regression but does not split products with the units problem into different categories. As can be seen, the price elasticity with respect to employment is smaller when the units problem is corrected, implying that the misreporting of units is correlated with size.

In addition to the manual correction, we also implement an algorithm which identifies product categories for which units have been potentially misreported. The algorithm consists of the following steps:

1. If the maximum price reported for a product is less than 50 times the minimum price, then the product is classified as one with no units misreporting.

2. We first arrange prices in ascending order within a product category. If there are two con-

[54] While in the milk example presented here, the units problem and the appropriate price cutoff was obvious, for some other products the problem is harder to clearly identify. In these cases we use our judgment to decide on the price cutoff.

secutive prices which are at least different by a factor of 20, and the average price above the jump is between 500 and 2000 times the average price below the jump, then the product is classified as one with a units misreporting problem and is split into two product categories.

3. We run regressions of log of price on log of employment with a dummy which takes value 1 for all observations below a given price, that is, if a product category has 50 plants producing it, we run 50 separate regressions - in the first the dummy only takes the value 1 for the lowest price plants, for the next regression, the dummy takes value 1 for the two lowest prices and so on. We then compare the highest R-square that we get with the dummies (within the product category) with the R-square when we run a regressions of log of price on log of employment with no dummy. If the difference in R-square is more than 0.75, and the difference in mean prices above the dummy (for the highest R-square) is at least 300 times higher than the average price below the dummy, then the product is classified as one with units misreporting and is split into two product categories.

When implementing the algorithm instead of the manual correction, the elasticity of price to size is 0.1037 in the ASI. The result is not very sensitive to using slightly different thresholds in the three stages of the algorithm. For example, changing the threshold for the R-square in step 3 to 0.8 and 0.7 changes the estimated elasticity to 0.1091 and 0.0986 respectively.

We also implement the algorithm for the input prices regressions and the results are similar to the ones with the manual correction

G. Calibrating Production Parameters - θ_{q_n}

This section provides more details on the calibration of θ_{q_n}, the share of unskilled workers in the intermediate producers production function. As mentioned in the paper, θ_{q_n} is chosen to match the wage premium and the ratio of unskilled to skilled workers for different qualities relative to the lowest quality level.

The target for the wage premium is obtained by running Mincerian type regression using the Employment-Unemployment Survey of 2004-5. Each individual is asked to report the main activities he or she undertook in the last seven days. Individuals can report multiple activities, and report if they were involved in the activity with “full intensity” or “half intensity. The wages earned in the last week are reported for all activities separately (if that activity generated wages). The average wage earned by each individual is computed by dividing the total wage earned for each activity over the last seven days by the number of intensity-days worked (summing across days and treating full intensity as 1 day and half intensity as 0.5 days) in that activity.

The wage premium for skilled workers is computed by running a regression of log of wages on a dummy which takes the value 1 if the worker is skilled (ten or more years of education) controlling for potential experience (age minus years of education minus four) and its square, and dummies for each 4-digit industry, 2-digit occupation, state, sector (urban or rural), and sex. We restrict the sample to workers reporting their industry as manufacturing (2-digit NIC between 15 and 36) and individuals between the age of 15 and 65 only.

Column 1 of Table 20 reports the results for the regression. The coefficient on the dummy which takes value 1 if a person is classified as skilled is 0.45, implying a wage premium of 56.8 percent which is rounded up to 60 percent when calibrating the model.

Calibrating θ_{q_n} also requires $N-1$ ratios for equation 12, the unskilled to skilled ratio for different qualities relative to the lowest quality. As mentioned in the paper, size categories reported in the Employment-Unemployment survey are very coarse, and therefore cannot be used to compute eleven ratios for equation 12 for eleven different quality (size) levels. Instead the relation between the size of an establishment and the share of unskilled to skilled workers is extrapolated based on the values reported in Table 11. The table reports that plants of size five or less have a unskilled to skilled ratio of 5.05 while plants of size 5 to 20 have a unskilled to skilled ratio of 2.92. These two points are used to extrapolate the unskilled to skilled ratio for larger sized plants with the ratio taking a minimum value of 0.5 (hire twice as many skilled as compared to unskilled workers). These extrapolated values are used to compute equation 12 for different quality levels given the average size of each quality level.

H. Additional figures and tables

Table 16: Plants with more capital produce use more expensive input goods

	(1) log(input price)	(2) log(input price)	(3) log(input price)	(4) log(input price)
Machinery Capital/ Labor ratio	0.019*** (0.0056)			
Machinery Investment/ Labor ratio		0.018 (0.012)		
Total Capital/ Labor ratio			0.013*** (0.0040)	
Total Capital Investment/ Labor ratio				0.014* (0.0079)
Adjusted R^2	0.892	0.892	0.893	0.893
Winsor	Yes	Yes	Yes	Yes
Observations	105197	105197	106749	106749
State x Rural x Product FE	Yes	Yes	Yes	Yes
Number of Products	2190	2190	2190	2190
SE clusters:	Product	Product	Product	Product
Number of Clusters	1485	1485	1487	1487
Sample	ASI	ASI	ASI	ASI

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table examines whether plants with a higher capital to labor ratio use more expensive inputs. Column 1 (2) regresses the stock of machinery capital (total capital) to labor on input prices. Column 2 (4) regresses machinery investment (total investment) to labor on input prices. To prevent our results being distorted by outliers, the input prices are winsorized at the 1 percent level. Each regression includes the triple interaction of state, rural, and input product fixed effects. We cluster our standard errors for each input product.

Table 17: Summary Statistics: ASI and SUM

	Annual Survey of Industries			**Survey of Unorg. Manufacturing**		
	Observations	Plants	Employment	Observations	Plants	Employment
1989-90	45	88	6,999	94	13,279	26,968
1994-95	52	105	7,853	156	12,114	29,924
2000-05	30	119	7,762	220	16,994	37,016
2005-06	42	125	8,811	82	17,037	36,376
2009-10	41	144	11,506	98	17,211	34,910

Notes: All numbers are in thousands ('000). The data is from the Annual Survey of Industries and the Survey of Unorganized manufacturing for five different years. The row corresponding to the year 2009 reports results the ASI of 2009-10 but the SUM of 2010-11 . The column "Observations" reports the total number of observations surveyed in the year. The "Plants" and "Employment" columns report the total plants and the total employment in these plants after taking into account the survey weights provided with the surveys. Four states were excluded due to lack of coverage in some years of the ASI. Only plants which reported industries which corresponded to the 2-digit NIC 2004 classification between 15 and 36 were included.

Table 18: Summary Statistics: Consumer Expenditure Survey

	Items	Share of Expenses	Items with Prices	Share with Prices
Food	161	0.499	156	0.980
Fuel and Light	13	0.094	10	0.932
Clothing and Footwear	27	0.075	24	0.967
Other goods and services	86	0.288	0	0.000
Durables	52	0.044	19	0.449

Notes: The data is from the Consumer Expenditure Survey conducted by the NSS in 2004-05. The rows represent broad expenditure categories. The column "Item" gives the number of distinct goods in the category for which consumption was reported. "Share of Expenses" gives the share of total expenditure that was devoted to the particular expenditure category when summing over all households. "Items with Prices" reports the number of items in the category for which values and quantities were reported, allowing calculation of prices. "Share of Prices" reports the share of expenditure within the category which was devoted to items for which the price could be computed.

Table 19: Main Activity of Individual - 2003 Consumer Expenditure Survey

Description	Code
Worked in hh enterprise (self-employed): own account worker	11
Worked in hh enterprise (self-employed): employer	12
Worked as helper in hh enterprise	21
Worked as regular salaried/wage employee	31
Worked as casual wage labor: in public works	41
Worked as casual wage labor: in other types of work	51
Did not work but was seeking and/or available for work	81
Attended educational institution	91
Attended domestic duties only	92
Domestic duties and engaged in free collection of goods, sewing, tailoring, etc. for household use	93
Rentiers, pensioner, remittance recipients etc	94
Not able to work due to disability	95
Beggars, prostitutes	96
Others	97

Notes: The 2003 consumer expenditure survey asks each individual in the household to report their main activity during the year. The table lists the different activities which the individuals could report. People who reported codes 92, 93, 94, or 97 were classified as non-workers and households which had at least one person between the age of 15 and 70 who was classified as a non-worker were considered to have low opportunity cost of time.

Table 20: Wage Premium from Employment-Unemployment Survey

Dependent variable: log(wage)		
	(1)	(2)
skilled	0.450*** (0.0150)	0.445*** (0.0147)
Wage Premium	1.568	1.560
Winsorize 1%		Y
Observations	11,003	11,003

Notes: The data is from the Employment Unemployment Survey of 2004-05. Column 1 reports results for the regression of log of wages earned by an individual on a dummy which takes value 1 if the individual has 10 or more years of education. Column 2 winsorizes 1 percent tails of wages. All regressions include controls for potential experience (age minus years of schooling minus 4) and its square, and dummies for each 4-digit NIC industry, 2-digit occupation, state, sector (urban or rural), and sex. The wage premium implied by the coefficient estimate for skilled is given in the row labeled "Wage Premium". Robust standard errors are reported. ***p<0.01.

Table 21: Ranking of Industries Based on Tradability Index

	Herfindahl Index	Export-Import Index
Industries Below Median (Non-tradable)	151,152,153,154,155,171,201,202,210,221,222,241,242,251,252,261,269,272,273,281,289,291,292,311 312,343,361,369	152,153,154,155,160,171,182,201,202,210,222,231,251,252,269,271,281,293,311,313,314,315,341,342,343,352,359,361
Industries Above Median (Tradable)	160,172,173,181,191,192,223,232,243,271,293,300,313,314,315,319,321,322,323,331,332,333,341,342,351,352,353,359	151,172,173,181,191,192,221,232,241,242,243,261,272,289,291,292,300,312,319,321,322,323,331,332,333,351,353,369

Notes: The table lists the 3-digit industries (NIC04) which fall above and below the median of for the two tradability indexes.

Table 22: Size Income Relation Across States for Tradables vs. Non-tradables: No Weighting

Dependent Variable: share of employment in <=5 in industry ‘i’, state ‘s’, time ‘t’				
	(1)	(2)	(3)	(4)
log(per-capita SNDP)*tradability	0.015	-0.026	-0.043	-0.006
	(0.0376)	(0.0705)	(0.0278)	(0.0415)
Index	H-index	H-index	Exp-Imp	Exp-Imp
Cutoff	Median	Quartile	Median	Quartile
Observations	3,885	1,826	3,899	1,959

Notes: The data is from five rounds of the ASI and SUM. The table reports regression results for the share of employment in plants of size 5 or less in industry ‘i’ in state ‘s’ at time ‘t’ on log per-capita state NDP interacted with a dummy which takes value 1 if industry ‘i’ is classified as a tradable industry and 0 if it is classified as non-tradable. Column 1 classifies an industry as tradable if the Herfindahl Index across US counties for the industry was above the median of Herfindahl Indexes, and non-tradable if it was below the median. Column 2 uses top and bottom quartiles of the Herfindahl Index as cutoffs. Column 3 and 4 use the tradability index based on Indian exports and imports and uses the median and the top and bottom quartiles as cutoffs respectively. All regressions include fixed effects for industry interacted with time and state interacted with time. No weights are applied to the observations in the regressions. Standard errors are clustered at the state level.

Table 23: Size Income Relation Across States for Tradables vs. Non-tradables: Exclude “NEC” and “Others”

Dependent Variable: share of employment in <=5 in industry ‘i’, state ‘s’, time ‘t’				
	(1)	(2)	(3)	(4)
log(per-capita SNDP)*tradability	0.050	0.036	0.002	-0.006
	(0.0399)	(0.0525)	(0.0500)	(0.0549)
Index	H-index	H-index	Exp-Imp	Exp-Imp
Cutoff	Median	Quartile	Median	Quartile
Observations	3,219	1,531	3,233	1,593

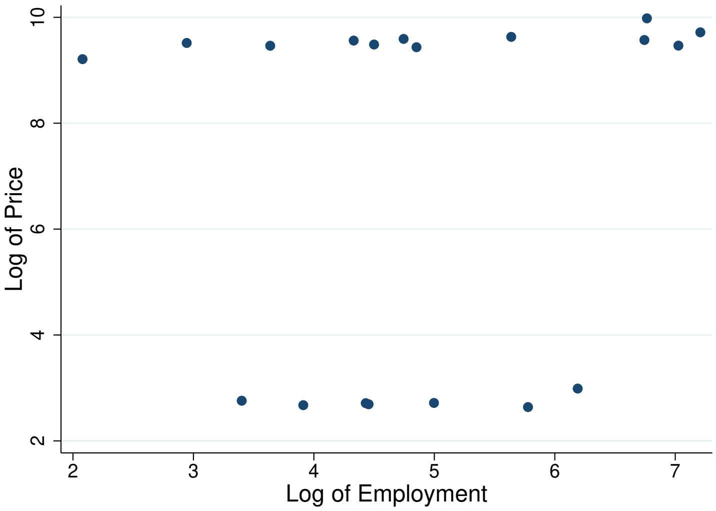
Notes: The data is from five rounds of the ASI and SUM. Residual industries (with words "NEC" or "other") are removed. The table reports regression results for the share of employment in plants of size 5 or less in industry ‘i’ in state ‘s’ at time ‘t’ on log per-capita state NDP interacted with a dummy which takes value 1 if industry ‘i’ is classified as a tradable industry and 0 if it is classified as non-tradable. Column 1 classifies an industry as tradable if the Herfindahl Index across US counties for the industry was above the median of Herfindahl Indexes, and non-tradable if it was below the median. Column 2 uses top and bottom quartiles of the Herfindahl Index as cutoffs. Column 3 and 4 use the tradability index based on Indian exports and imports and uses the median and the top and bottom quartiles as cutoffs respectively. All regressions include fixed effects for industry interacted with time and state interacted with time. Each observation is weighted by the share of observations in the state-industry cell out of the total observations in the ASI and SUM combined for the given year. Standard errors are clustered at the state level.

Table 24: Units Misreporting Problem in the ASI

Dependent Variable: log(output price)				
	(1)	(2)	(3)	(4)
log(labor)	0.096***	0.155***	0.106***	0.125***
	(0.0087)	(0.0125)	(0.0133)	(0.0152)
Units Problem Accounted For	Y	N	Y	N
Sample	ASI	ASI	Both	Both
Observations	46,704	46,704	75,161	75,161
Number of products	1,217	1,077	3,181	3,041
Number of clusters	1,078	1,078	3,042	3,042

Notes: The data is from the ASI and SUM of 2005-06. All columns report results for regressions of log of price charged by plants for their products on log of number of employees hired by the plant. Columns 1 and 2 restrict the sample to the ASI alone while columns 3 and 4 combine the ASI and the SUM. Columns 1 and 3 implement the manual units correction (same as reported in main text) while columns 2 and 4 do not correct for misreporting of units. 1 percent tails of prices (within a product) and plant size are winsorized. All regressions include product fixed effects and state times urban-rural fixed effects. Standard errors are clustered at the product level. ***p<0.01.

Figure 17: Price Charged for Milk by Different Plants Against Size


Log of Price
Log of Employment